Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra informatiky a výpočetní techniky

Diplomová práce

Segmentace historických obrazových dokumentů

Plzeň 2019 Bc. Miroslav Liška

Místo této strany bude zadání práce.

Prohlášení

Prohlašuji, že jsem diplomovou práci vypracoval samostatně a výhradně s použitím citovaných pramenů.

V Plzni dne 16. května 2019

Bc. Miroslav Liška

Abstract

Correct segmentation of image documents is one of the most important tasks in OCR. During segmentation, areas of interests such as text blocks, text lines and separators are automatically labeled. Historical image documents are often malformed, contain noise and have irregular structure. Because of these issues, successful segmentation of such documents presents a difficult challenge. This Master's thesis explores possible approaches to segmentation of historical documents, after subsequent analysis, fully convolutional neural networks ARU-Net and U-net and its variations were used for this task. This thesis also deals with manual creation of a data set, particularly with creation of expected results of segmentation and suitable tools for achieving this task. Based on results obtained from models trained on the created data set, variation of the U-net network was selected for text labeling and ARU-net for finding of separators and labeling of lines of text. The result of this thesis a program combining results of individual neural networks. This program is able to detect and cut lines of text from input images while retaining reading order. Text lines obtained from this program are suitable for further OCR processing.

Abstrakt

Správná segmentace obrazových dokumentů je jednou z nejdůležitějších součástí OCR systémů. Během ní jsou v obrázcích automaticky označovány oblasti zájmu, jako jsou například textové bloky, řádky textů, oddělovací čáry a jiné. Historické obrazové dokumenty jsou často různě deformované, obsahují šum a mají nepravidelnou strukturu a tak je úspěšná segmentace těchto dokumentů velkou výzvou. V rámci diplomové práce byly prozkoumány možné přístupy k segmentaci historických dokumentů a po následné analýze byly použity plně konvoluční neuronové sítě ARU-Net, U-Net a její úprava. Dále se práce věnuje tvorbě datové sady, zejména pak vytvoření očekávaných výsledků segmentace a nástrojům pro jejich vytvoření. Na základě dosažených výsledků modelů na této datové sadě je pro označování textu vybrána síť upravený U-Net, pro nalezení oddělovacích čar a pro označení řádků textu síť ARU-Net. Spojením výsledků jednotlivých sítí a jejich následným zpracováním byl vytvořen program, který ve vstupním obrázku detekuje a vyřezává řádky textu a to tak, aby bylo zachováno pořadí čtení. Takto získané řádky jsou vhodné pro další zpracování OCR.

Obsah

1 Úvod

Rozvoj počítačových technologií umožňuje stále snadněji převádět papírové dokumenty na digitální. Díky tomu vznikají projekty, které si kladou za cíl s využitím digitalizace zachovat kulturní dědictví lidstva. Jedním z nich je i projekt „Bavorsko-česká síť digitálních historických pramenů“ [20], jehož výsledkem bude opětovné spojení historicky k sobě patřících, česko-německých archivních fondů, které byly rozděleny vlivem 2. světové války. Vytvořené digitální reprodukce jednotlivých archiválií budou v budoucnu nahrány do jednoho úložiště, čímž vznikne virutální digitální archiv. Zdigitalizované dokumenty budou prostřednictvím webové stránky zveřejněny a zpřístupněny široké veřejnosti.

Pouhým naskenováním však práce s dokumentem nemusí končit. Dalším návazným krokem je analýza dokumentu. Ta zahrnuje nalezení bloků obsahujících informace. Může se jednat například o bloky obsahující text, obrázky, tabulky a jiné. Tento proces se nazývá segmentace a je předmětem této diplomové práce. Z textových bloků je dále extrahován text metodou optického rozpoznávání znaků (OCR). S tímto textem je možné dále pracovat. Může se jednat například o indexování, které umožní efektivní vyhledávání. Dále může být text přeložen do jiného jazyka. Vzhledem k tomu, že proces segmentace stojí na počátku celého procesu analýzy dokumentu, jsou výstupy tohoto procesu klíčové pro navazující kroky.

V úvodu práce bude nejprve provedena rešerše přístupů používaných pro segmentaci historických dokumentů. Následně bude provedena analýza dodané datové sady `Portafontium` a dalších podobných, volně dostupných sad, které kromě obrázků obsahují i metadata definující očekávaný výstup segmentace. Vzhledem k tomu, že datová sada `Portafontium` neobsahuje tato metadata, bude potřeba tato data vytvořit, ať kvůli trénování klasifikačního modelu nebo pro vyhodnocení jeho výstupů. Část práce je tak věnována možnostem tvorby a způsobu uložení těchto informací. Dále bude provedena rešerše metrik používaných pro vyhodnocení výsledků segmentace. Na základě analýzy budou vybrány metody segmentace, které budou součástí výsledného řešení. V závěru práce pak bude provedeno vyhodnocení dosažených výsledků.

2 Segmentace obrazu

Pojmem segmentace obrazu se rozumí soubor postupů, jehož výstupem je automatické označení oblastí zájmů. V případě naskenovaných dokumentů se jedná o označení jeho obsahu. Ten je následně dělen na text, obrázky, oddělující čáry, tabulky a jiné. Používané metody segmentace lze rozdělit na tři kategorie a budou popsány v následujícím textu:

- Extrakce příznaků a klasifikátoru.
- Konvoluční neuronové sítě.
- Plně konvoluční neuronové sítě.

2.1 Extrakce příznaků a jejich použití

Každý pixel v obrázku je reprezentován vektorem příznaků. Na základě těchto příznaků je během klasifikace každému pixelu v obrázku přiřazena odpovídající třída.

U mnoha metod segmentace začíná celý proces binarizací vstupních obrázků, kdy je vícekanálový obrázek (barevný) převeden na obrázek jednokanálový (šedotónový). Následně jsou jednotlivé pixely s využitím práhu rozděleny na dvě kategorie, čímž vznikne obrázek, kde každý pixel nabývá hodnoty 0 (černá barva) nebo 255 (bílá barva). Jedním z důvodů použití binarizace je snížení dimenzionality problému a tím tak klesá výpočetní náročnost [8]. Binarizací však dochází ke ztrátě barevné informace. Chen a kolektiv ukázali, že barevná informace pixelu a jeho okolí může sloužit k vytvoření důležitých příznaků [8]. Označme pixel v obrázku $I_{x,y}$, kde x a y představují jeho souřadnice na osách. Každý pixel $I_{x,y}$ má svoje okolí, definované jako $N(I_{x,y}, n)$, kde n je liché číslo a představuje délku stran čtverce rámujícího pixel $I_{x,y}$.

2.1.1 Příznaky pro reprezentaci obrázku

Metody pro jejich získání jsou popsány v [8] a [25] a lze je rozdělit na 3 kategorie:

- Souřadnice - Všechny obrázky v datové sadě jsou naškálovány na stejnou velikost. Jako příznaky jsou pak použity souřadnice pixelu x a y v novém obrázku.

- Barva - Jedná se o příznaky získané z barvy pixelu.

  - Hodnoty RGB - intezity jednotlivých kanálů barvy.

  - Suma okolí - součet intezity jednotlivých kanálů barvy v okolí N.

  - Maximální a minimální hodnota okolí - jako příznak je brána minimální/maximální hodnota intenzity kanálu barvy v okolí pixelu. V tomto případě není okolí myšleno jako pixely spadající do čtvercového okolí pixelu, ale pouze pixely sousedící horizontálně a vertikálně do vzdálenosti $(n-1)/2$

  - Suma ve sloupci - jako okolí se v tomto případě rozumí pixely z celé stránky, které jsou ve stejném sloupci. Příznakem je pak suma intezity barevného kanálu všech takto sousedících pixelů.

  - Střední hodnota v okolí - ve čtvercovém okolí pixelu je spočítána střední hodnota intenzita kanálu barvy. Označme střední hodnotu jako $M(p_{x,z}, n)^r$, kde r je kanál barvy a n je velikost oblasti kolem pixelu.

  - Variabilita v okolí - malý stupeň variability signalizuje, že jsou si pixely podobné. Variabilita se počítá z okolí pixelu jako:

$$V(p_{x,y}, n)^r = \frac{\sum_{i=-d}^{d} \sum_{j=-d}^{d} (I^r_{x+i,y+j} - M(p_{x,z}, n)^r)^2}{n \times n} \tag{2.1}$$

$$d = (n-1)/2 \tag{2.2}$$

  - Hladkost barvy (Color smoothness) - využívá informace o variabilitě. Čím vyšší hladkost, tím jsou si pixely v okolí podobnější:

$$SMO(p_{x,y}, n) = \frac{1}{1 + V(p_{x,y}, n)} \tag{2.3}$$

- Textury - Nejprve se obrázek převede na šedotónový:

  - Local Binary Pattern (LBP) [3] - Hodnota pixelu je spočítána z jeho okolí podle 2.4, kde P definuje počet okolních pixelů; R definuje poloměr kružnice kolem $p_{x,y}$ určující sousední pixely; g_c je intenzita pixelu uprostřed kružnice; g_p představuje intenzitu p-tého sousedního pixelu.

$$LBP_{P,R}(x,y) = \sum_{p=0}^{P-1} s(g_p - g_c) 2^p$$

$$s(x) = \begin{cases} 1, & \text{pokud } x < 0. \\ 0, & \text{jinak.} \end{cases} \tag{2.4}$$

- Gáborovy filtry - na každý pixel v okolí $N(p_{x,y}, n)$ se aplikují Gáborovy filtry $g(x, y; \lambda, \theta, \psi, \sigma, \gamma)$ s různou orientací θ. Filtr je reprezentován maskou (konvolučním jádrem), pod kterým si můžeme představit 2D pole s různými váhami. Číslo vycházející z Gáborova filtru dosahuje vyšších hodnot, pokud v nastaveném směru filtru existuje textura. Ohodnocením okolních pixelů je pak směr θ, ve kterém dosáhl filtr nejvyšších hodnot. Z takto ohodnocených sousedů se vytvoří histogram, kde osa x reprezentuje směr θ a osa y představuje počet sousedů, kteří ve směru θ_i dosáhli nejvyšších hodnot.

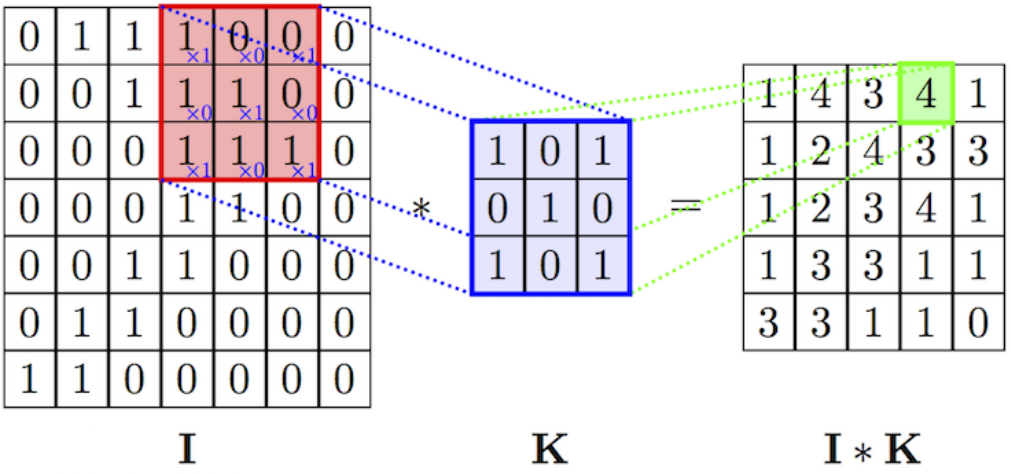
2.1.2 Výběr nejlepších příznaků

Vysoká dimenzionalita způsobená použitím všech výše popsaných příznaků vede k vyšší výpočetní náročnosti klasifikačního problému [8]. Cílem výběru příznaků je snížení dimenzionality tak, aby nebyla snížena přesnost klasifikace a to takovým způsobem, že jsou odebrány příznaky irelevantní či redundantní. K tomu je použit algoritmus Fast Correlation-Based Filter (FCBF) [27], jehož výstupem je podmnožina S', kde každý příznak velmi koreluje s klasifikační třídou a zároveň výrazně nekoreluje z žádným jiným příznakem. Je použit parametr práhu θ, který, pokud je vysoký, může odstranit některé důležité příznaky. Na druhou stranu, pokud je jeho hodnota příliš malá, může množina obsahovat redundatní příznaky. Ideální hodnota práhu závisí na použité datové sadě [8] a je potřeba s ní experimentovat.

Jakmile jsou jednotlivé pixely reprezentovány touto redukovanou množinou příznaků, je možné použít libovolný klasifikační model. Například v [8] byl zvolen klasifikační model SVM (Support Vector Machine) [12].

2.2 Konvoluční neoronové sítě

Oproti předchozímu přístupu, kdy si musíme příznaky vytvářet sami, je možné využít konvoluční neuronovou síť (CNN - Convolutional Neural Network)), která se trénováním sama naučí si z obrázku příznaky

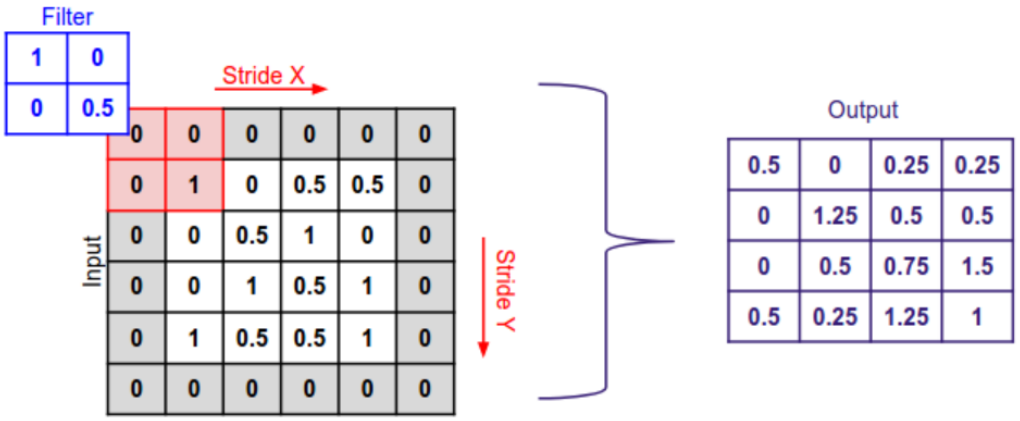


Obrázek 2.1: Ukázka konvoluce pro extrakci příznaku. Na vstupu je obraz I o velikosti 7x7x1. Konvoluce je prováděna s filtrem K o velikosti 3x3x1. Výsledkem konvoluce je matice o rozměrech 5x5x1. Krok posunu je nastaven na 1. Zdroj https://jenslaufer.com.

extrahovat [18]. Skryté vrstvy sítě jsou tvořeny alespoň jednou vrstvou konvoluční. Pro obrázky je operace konvoluce $*$ definovaná jako:

$$f[x,y] * g[x,y] = \sum_{n_1} \sum_{n_2} f[n_1, n_2] \cdot g[x - n_1, y - n_2] = I * K \tag{2.5}$$

kde $f[x,y]$ je vstupní obraz, na který bude použit tzv. konvoluční filtr $g[x,y]$ (někdy se používá výraz maska). Pod konvoučním filtrem $g[x,y]$ si můžeme představit dvourozměrnou matici vah, která právě díky násobení s obrazem detekuje příznaky (například vertikální hrany). Konvoluční filtr bývá obvykle menší než vstupní obraz. Jedna konvoluční vrstva může být tvořena i několika filtry, kdy každý filtr bude detekovat jiný typ příznaku. Konvoluce nad vstupním obrazem je realizována posuvným oknem, jak můžeme vidět na obrázku 2.1. O kolik pozic se okno posune ve vertikálním nebo horizontálním směru je jedním z parametrů konvoluční vrstvy (anglicky stride). Výstupem konvoluční vrstvy jsou tzv. mapy vlastností (feature maps). Na obrázku 2.1 si můžeme všimnout, že se rozměry vstupního obrazu nerovnají rozměrům výstupního obrazu. To znamená, že zde dojde ke ztrátě informace tzv. okrajovým efektem. Ten lze vysvětlit na obrázku 2.1. Všimněme si, že je možné horizontálně posunout okénko ještě o jednu pozici doprava. Pro pixely v 6. a 7. sloupci se ale konvoluce již nebude počítat. Ke stejnému jevu dojde i v případě vertikálního posunu, kdy konvoluce nebude provedena na řádky 6 a 7. Tomuto efektu je možné v případě potřeby předejít přidáním ohraničení kolem celého obrázku



Obrázek 2.2: Přidání nul kolem celého obrázku. Zdroj https://medium.com/.

(anglicky *padding*), jako můžeme vidět na obrázku 2.2 Výstupní rozměr matice je definován předpisem:

$$O = \frac{I - K + 2P}{S} + 1 \tag{2.6}$$

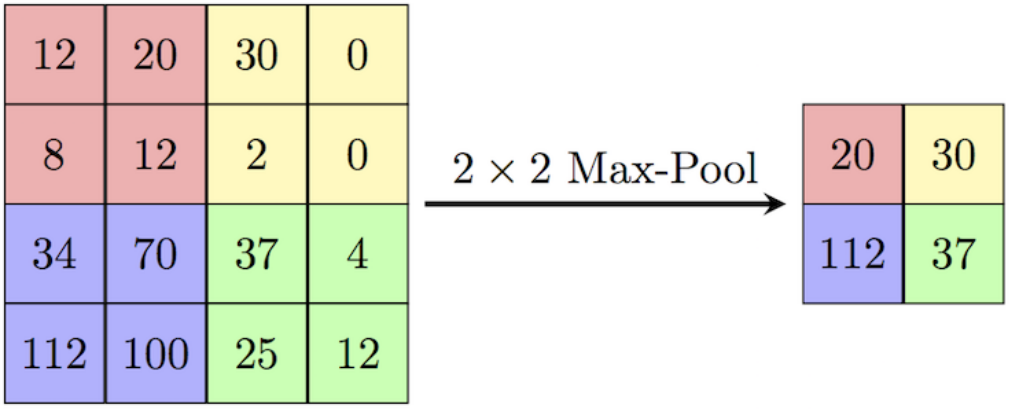
kde I je velikost vstupního obrazu, K velikost filtru, S je parametr posunu (stride) a P je šířka *paddingu*. Počet kanálů ve výstupním obrazu je roven počtu filtrů použitých během konvoluce.

Na výstupní obraz je následně aplikována nelineární aktivační funkce. Často používanou funkcí je například funkce ReLU (Rectified Linear Unit) dána předpisem:

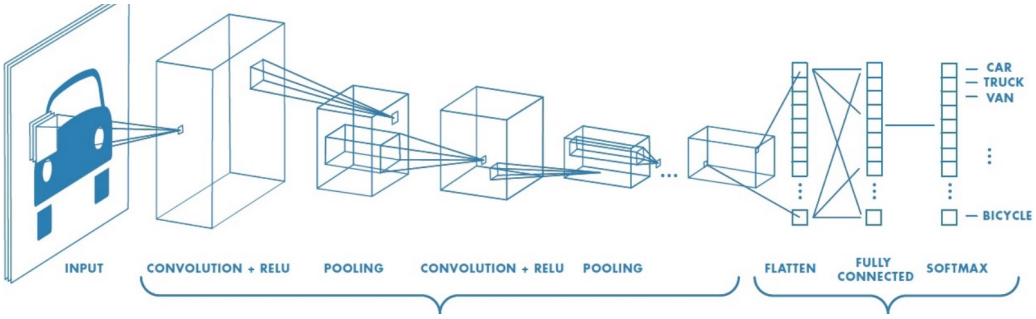
$$f(x) = max(x, 0) \tag{2.7}$$

To znamená, že každé záporné číslo ve vstupní matici je nahrazeno nulou a ostatní čísla jsou zachována.

Další vrstvou používanou v konvolučních sítích je vrstva podvzorkovací (anglicky pooling layer). Tato vrstva je definovaná velikostí filtru, jeho krokem posunu (stride) a funkcí definující chování filtru (používá se výběr maximální hodnoty v obrazu - Max-Pool; nebo je spočítána průměrná hodnota - Avg-Pool). Vizualizaci podvzorkování můžeme vidět na obrázku 2.3. Napříč vstupním obrazem je posouváno okénko, na které je použit filtr. Výsledek posouvání a použití filtru je výstupem podvzorkovací vrstvy. Podvzorkování se například používá pro zajištění invariance vůči lokálnímu posunu. Efektem podvzorkování je mimo jiné i snížení dimenzionality příznaků. Zároveň však dochází ke ztrátě informace, kde se v obrazu zkoumaný znak nachází, což může být v některých případech nežádoucí. Výstupem z konvoluční sítě je reprezentace vstupního obrazu formou



Obrázek 2.3: Vizualizace podvzorkování. Použita je funkce Max-Pool s velikostí filtru 2×2 a krokem posunu 2. Zdroj https://computersciencewiki.org.



Obrázek 2.4: Ukázka konvoluční sítě. Zdroj https://medium.com/.

příznaků (mapa příznaků). Tyto příznaky jsou pak typicky přivedeny na vstup plně propojené neuronové sítě, která na základě příznaků provede klasifikaci do jedné ze tříd. Architekturu takové sítě je možné vidět na obrázku 2.4

2.2.1 Konvoluční sítě pro segmentaci historických dokumentů

Chen a kolektiv přichází s návrhem konvoluční sítě, který se osvědčil pro segmentaci dokumentů [9].

Vstupní obrázek se nejprve předzpracuje a to tak, že jsou v něm nalezeny tzv. superpixely. Pod superpixelem si můžeme představit množinu pixelů, které patří ke stejnému objektu. Příklad může být například písmeno v textu, kdy všechny pixely tvořící toto písmeno jsou součástí superpixelu. Namísto klasifikace všech pixelů zvlášť se klasifikuje pouze

střed superpixelů (jeden pixel). Na základě výsledku klasifikace středového pixelu jsou stejnou třídou označeny i zbylé pixely. Pro nalezení superpixelů v obrázku je použit algoritmus Simple Linear Iterative Clustering (SLIC) [2].

Architektura navržené konvoluční sítě je pak následující: Velikost vstupního obrázku je 28 × 28 × 1 (obrázek je šedotónový). Za vstupní vrstvou následuje vrstva konvoluční o rozměrech 26 × 26 se 4 filtry o velikosti 3 × 3.

Za touto konvoluční vrstvou následuje plně propojená vrstva se 100 neurony. ReLU je používáno jako aktivační funkce. V poslední vrstvě je použita softmax aktivační funkce, díky které získáme pravděpodobnost jednotlivých tříd. Ta je definovaná jako:

$$P(y = i|x, W_1, \cdots, W_M, b1, \cdots, b_M) = \frac{e^{W_i x + b_i}}{\sum_{j=1}^{M} e^{W_j x + b_j}} \tag{2.8}$$

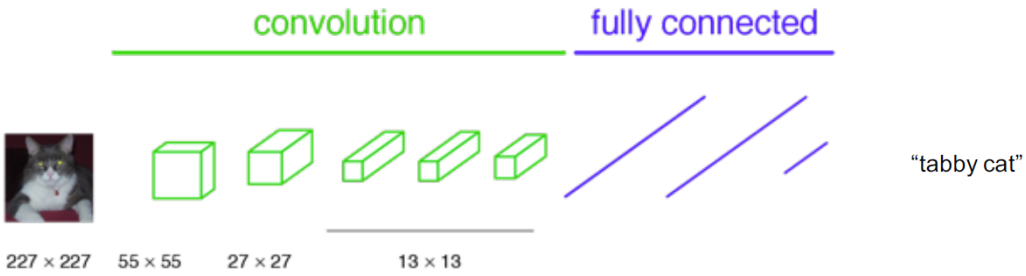
kde x je výstupem z plně propojené vrstvy, W_i a b_i jsou váhy a prahy (biases) i. neuronu v této vrstvě a M je počet klasifikačních tříd. Jako výsledná třída $\hat{y}$ je vybrána třída s nejvyšší pravděpodobností.

Pro trénování je potřeba ke každému obrázku mít i jeho `ground-truth`. Trénování probíhá tak, že jsou v obrázku nejprve nalezeny superpixely. Z každého superpixlu se následně vybere prostřední pixel. Kolem tohoto pixelu je vybraná oblast o rozměrech 28 x 28. Tímto způsobem je z původního obrázku vyříznuta pouze jeho část, tzv. *patch*. Tento *patch* je následně přiveden na vstup CNN. Očekáváné označení tohoto obrázku odpovídá označení středového pixelu *patche*. V CNN je cenová funkce definovaná jako:

$$L(X, Y) = -\frac{1}{n} \sum_{i=1}^{n} (\ln a(x^{(i)}) + (1 - y^{(i)}) ln(1 - a(x^{(i)}))) \tag{2.9}$$

kde $X = \{x^{(1)}, \cdots, x^n\}$ jsou trénovací *patche* a $Y = \{y^{(1)}, \cdots, y^n\}$ je odpovídající množina s očekávanými třídami. Počet trénovacích *patche* je označen jako n. CNN je trénovaná s využitím stochastického gradientního sestupu a použitím *dropout* techniky [24]. Smyslem použití *dropout* techniky je zabránění přeučení modelu umělým zanesením šumu do trénovacích vzorků.

Při klasifikaci neviděných dat se obrázek opět rozdělí na superpixely, kolem jejichž středů se vytvoří *patche*. Tyto *patche* jsou přivedeny na vstup CNN. Z výstupu sítě je pak určena nejpravděpodobnější třída. Tato třída je následně nastavena všem pixelům superpixelu.



Obrázek 2.5: Kovnoluční síť spojená s plně propojenou vrstvou. Zdroj https://medium.com/.

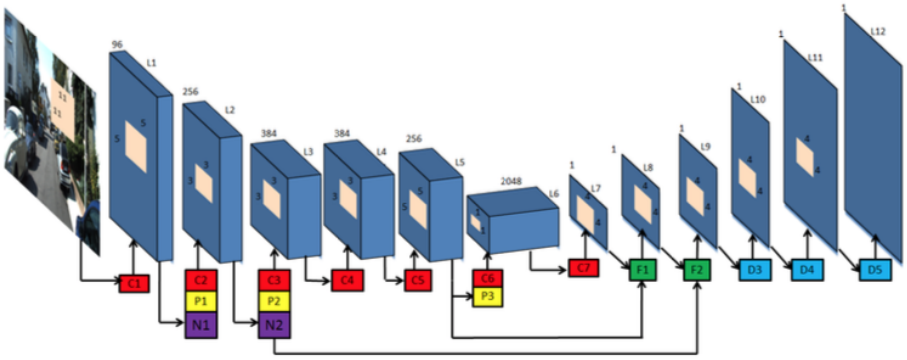
2.3 Plně konvoluční neuronové sítě

V případě použití konvolučních neuronových sítí je vstupní obrázek postupně zpracováván konvolucemi a podvzorkováním. Výsledek je následně spojen s plně propojenou vrstvou, jejíž výstupem je nejpravděpodobnější třída (*label*) pro celý vstupní obrázek. Jak bylo zmíněno v předchozí sekci 2.2.1, vstupní obrázek je potřeba nejprve rozdělit na podobrázky (*patche*). Podobrázky jsou následně vstupem konvoluční sítě.

V případě plně konvolučních neuronových sítí (FCNN - Fully Convolutional Neural Network) vstupuje do sítě celý obrázek. Na celý obrázek jsou postupně aplikovány konvoluce a podvzorkování. Namísto následného spojení s plně propojenou vrstvou, jako tomu bylo u CNN (viz. obrázek 2.5)), je postup otočen. Konvoluční operátory jsou nahrazeny operátory dekonvolučními.

V neuronových sítích se dále používají tzv. skip-spojení. Toto spojení slouží k tomu, aby výstup z jedné z vrstev byl kromě následující vrstvy spojen i s jinou vrstvou. Ve FCNN jsou takto spojovány např. konvoluční a dekonvoluční vrstvy se stejnou dimenzí. Toto spojení můžeme vidět na obrázku 2.6. Spojení je vytvářeno z důvodu ztráty prostorové informace způsobené především podvzorkováním a následným nadvzorkováním. Díky tomuto propojení je možné spojit informace získané na vyšší úrovni s informacemi získanými na nižší úrovni.

Obraz získaný z výstupní vrstvy má stejné rozměry jako vstupní obrázek a navíc bude mít tolik kanálů, kolik je počet předdefinovaných tříd. Pokud například má vstupní obrázek šířku 1024, výšku 768 a každý pixel je klasifikován do jedné z patnácti tříd, bude mít výstupní vrstva rozměry $1024 \times 768 \times 15$.



Obrázek 2.6: Použité skip-vrstvy. Převzato z [14].

2.3.1 Plně konvoluční sítě pro segmentaci historických dokumentů

Plně konvoluční síť navržená v [26] dosahuje v úloze segmentace historických dokumentů lepších výsledků, než dosahovaly doposud známé metody [26]. Autoři tuto síť nijak nepojmenovali, zadefinujme si ji pro tuto práci jako `upravený U-Net`. Během trénování síť požaduje vstupní obrázek a k němu odpovídající `ground-truth` obrázek. Vstupní obrázek je nejprve převeden na šedotónový a následně je provedena binarizace s využitím nástroje *Ocropy* [1]. Reprezentace `ground-truth` obrázku je volena tak, že černá barva reprezentuje pozadí a ostatní barvy jsou rozděleny pro jednotlivé třídy. Ukázku můžeme vidět na obrázku 2.7.

— 96 —

Ankündigungen aller Art.

Peruanischen-Guano,

direkt aus dem Depot der peruanischen Regierung bezogen, verkauft unter der Garantie der Aechtheit sowohl in einzelnen Säcken, als auch in ganzen Wagenladungen billigst

Hermann Roth,
in Oelsnitz. (Voigtland.)

Bezugnehmend auf vorstehende Bekanntmachung bin ich bereit Aufträge entgegen zu nehmen, und halte ich stets etwas Lager in meinem Hause.

Asch, den 10. März 1866.

Christof Jäger.
Stein Nr. 128.

Verkauf.

75 Centner gutes Grummet sowie einige hundert Mäßel

Kartoffeln

sind zu verkaufen und das Nähere zu erfragen Nr. 42 auf der Wieden.

Verkauf von drei Kirchenstühlen u. z. einer für Männer, und zwei für Frauen. Zu erfragen in der Expedition dss. Bl.

Bei meinem Abgange vom Hrn. **E. Beier** im Schießhause, erinnre ich hiermit die Hrn. Gäste die bei mir noch im Rückstande stehen, solchen an Obengenannten direkt zu zahlen.

Asch, am 23. März 1866.

Carl Schäffler.
Kellner.

Erklärung.

Im Höfer Anzeiger Nr. 30 erkühnt sich ein gewisser Herr Wirth als Zahnarzt sich zu empfehlen.

Obgleich ich Unterzeichneter keine Concurenz fürchte, so halte ich mich doch dem verehrl. Publikum gegenüber für verpflichtet, folgende Erklärung zu geben:

„Erwähnter Herr Wirth ist einfacher Goldarbeitergehilfe, und durch meine freundschaftliche längere Unterstützung hat sich derselbe auch einige Kenntnisse in der Zahntechnik erworben. Da Herr Wirth nicht die Zeugnisse, wie ich genau weiß, beibringen kann, daß derselbe als Zahnarzt aufzutreten berechtigt wäre, so verweise ich solchen hiemit in die Grenzen seines rechtlichen Standpunktes, da derselbe sich höchstens als Zahntechniker annonciren könnte, und ich fordre Herrn Wirth zugleich auf, sich als Zahnarzt öffentlich zu legitimiren.

Christoph Hofmann,
Goldarbeiter und Zahntechniker in Hof.

Reelles Heirathsgesuch.

Ein lediger, bemittelter Mann in den besten Jahren, — Kaufmann und selbstständig — sucht auf diesem in der Neuzeit viel und mit Glück betretenen Wege eine Lebensgefährtin, sei es Jungfrau oder Wittwe, welche natürlich nicht über 40 und nicht unter 18 Jahren alt sein darf. Sonstige Bedingungen wären: angenehmes Aeußere, Bildung und gesellschaftliche Tournüre, sanfter Charakter sowie ein disponibles Vermögen von 6—10,000 fl. Adressen womöglich mit Photographie unter der Chiffre: **H. v. B.** befördert die Exped. d. Bl. Strengste Discretion wird zugesichert und beansprucht.

Masthammel-Auction.

Auf dem Rittergute Ottengrün bei Oelsnitz sollen Freitag den 23. März 1866 von Mittag 12 Uhr an 104 Stück ganz fette Hammel, bayrische Raçe, gegen gleich baare Bezahlung meistbietend verkauft werden.

Kirchliche Nachrichten der letzten Woche.

Geboren: C. Wunderlich, Schuhmacher, ein Sohn. — W. März, Farbgehilfe, eine Tochter. — H. Jäger, Büttnermeister, ein Sohn. — Joh. Dittrich, Weber, ein Sohn. — K. Biedermann, Weber in Neuenbrand, ein Sohn. — Joh. Richter, Oek. in Nassengrub, ein Sohn. — C. Ludwig, Gastwirth in Nassengrub, eine Tochter. — Adam Benker, Oek. in Wildenau, eine Tochter. — C. Friedrich in Schönlind, eine Tochter. — **St. Niklas**: B. Hat, Handarb., ein Sohn. — Johann Müller, Schneider, eine Tochter. — **Neuberg**: W. Böttiger, Müller, ein Sohn. — C. Zeitler, eine Tochter. — Joh. Künzel, Weber, ein Sohn. — J. Steinel, Oek. in Grün, eine Tochter.

Gestorben: C. Wunderlich, Schuhmacher, ein Sohn. — Lorenz Richter, Strumpfw. — M. Geier, Strumpfwirkerswittwe. — Josef Wächter, Strumpfw., ein Sohn. — J. Schindler, Schuhmacher, ein Sohn. — Adam Geier, Müllermstr. von Mühlbach. — **St. Niklas**: R. Lindermeier, Handarb., in Oberreuth. — **Neuberg**: J. Gläsel. — J. Zöfel in Thonbrunn. — J. Wittmann, Webermstr. in Krugsreuth.

Tageskalender.

Telegraph. Course, Wien 23. März 1866, k. k. öffentl. Börse.

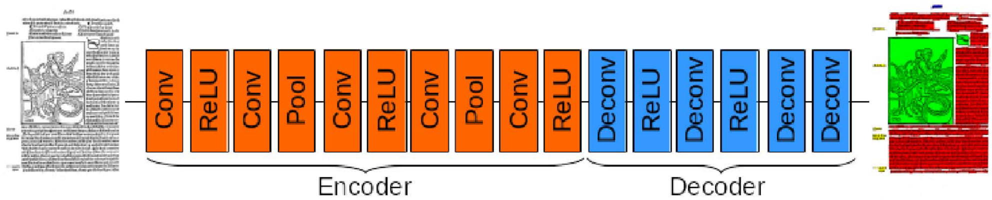
5% Metalliques	59. 60	Bankaktien	702.
5% National-Anlehen	62.	Creditaktien	140. 30
1860r. Staats-Anlehen	79. 10	London	106. 90
k. k. Münzdukaten	5. 14	Silber	106. 75

Prag, den 14. März 1866.
4. 65. 66. 75. 22.
Nächste Ziehung 24. März.

Brünn, den 17. März 1866.
46. 87. 31. 37. 43.
Nächste Ziehung 28. März.

Verantwortliche Redaction, Druck und Verlag von Gustav Becker in Asch.

Obrázek 2.7: Ukázka `ground-truth` obrázku. Pozadí je reprezentované černou barvou a hledané textové bloky barvou bílou.



Obrázek 2.8: Architektura FCNN pro segmentaci historických dokumentů podle [26].

Všechny vstupní obrázky jsou také naškálovány na velikost 260 × 390, což vede ke snížení výpočetního času a zároveň velikosti sítě [26].

Architektura sítě se skládá z konvolučních a dekonvolučních vrstev a je inspirována architekturou U-Net [22]. Architekturu je možné vidět na obrázku 2.8.

Lze si všimnout, že po dekonvolucích není prováděno nadvzorkování. Zvýšení rozlišení obrázku je možné totiž zajistit během dekonvoluce nastavením rozměru filtru a jeho posunu (`stride`) na stejnou velikost. Rozlišení je tak zvětšeno právě o tuto velikost. V každé konvoluční vrstvě

je filtr nastaven na velikost 5 × 5 a je použit *padding* pro zachování rozměrů. Počty filtrů v konvolučních vrstvách jsou postupně 40, 60, 120, 160, 240 a v dekonvolučních 240, 120, 60 a 6. Poslední dekonvoluční vrstva je použita pro získání predikce do jedné ze 6 tříd. Během podvzorkování je používán filtr o rozměrech 2 × 2. Stejnou velikost filtru také používají první dvě dekonvoluční vrstvy. Díky tomu je rozlišení obrazu vstupujícího do sítě rovno rozlišení výstupního obrazu. Pro vyhodnocení úspěšnosti je výstupní obraz naškálován zpět na velikost původního obrazu.

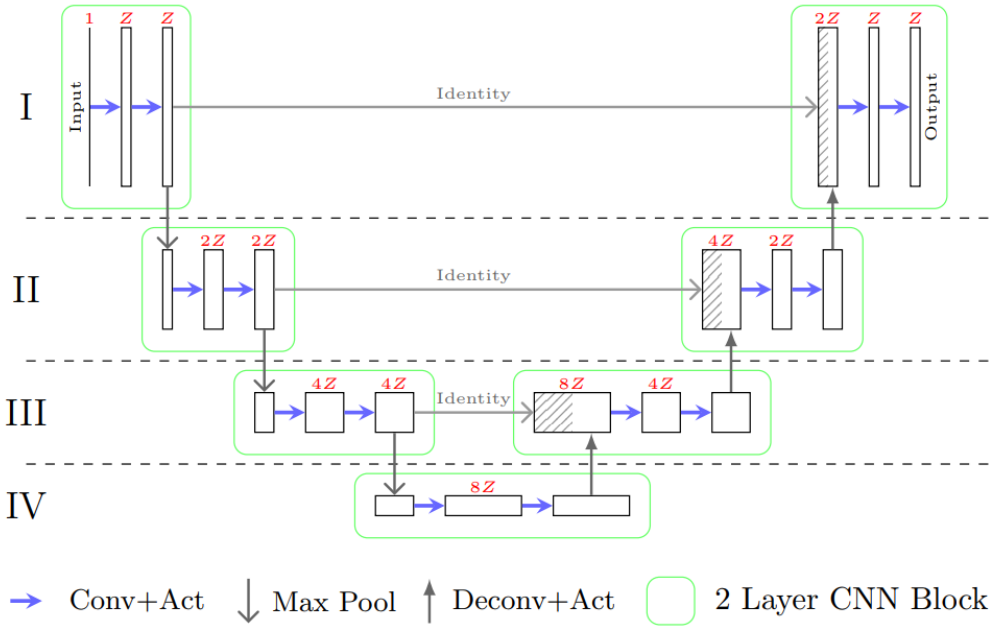
2.3.2 ARU-Net

V [16] byla představena hluboká neuronová síť ARU-Net, která byla navržena především pro detekci řádků textu v historických dokumentech. Architektura sítě byla inspirovaná architekturou plně konvoluční sítě U-Net [22], kterou rozšiřuje o dva další koncepty: reziduální bloky a pozornost (bude používán anglický výraz `attention`). Architekturu sítě U-Net můžeme vidět na obrázku 2.9. Značení `Conv+Act` znamená provedení konvoluce, jejíž výstup je přiveden do aktivační funkce. Max Pool představuje podvzorkování s hledáním maxima. `Deconv+Act` značí dekonvoluci a aktivační funkci. Zeleným obdélníkem jsou pak značeny dvouvrstvé bloky konvoluční neuronové sítě. Vstupem je obrázek určité velikosti. Každý obdélník představuje trojrozměrné pole (*výška* × *šířka* × *početmappříznaků*). V každém škálovacím prostoru (označeném římskými číslicemi) jsou prostorové rozměry map příznaků stejné. Na obrázku je tato informace vizualizována výškou obdélníku. Počet map příznaků je vizualizován šířkou obdélníku. Dále si můžeme všimnout skip-spojení mezi bloky ve stejném škálovacím prostoru.

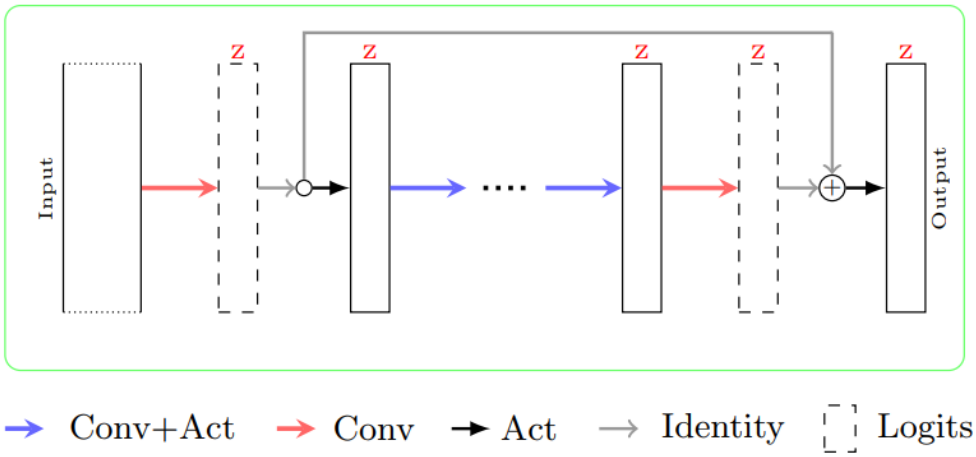
Oproti U-Net architektuře jsou dvouvrstvé bloky (v obrázku 2.9 zeleně označené bloky) konvoluční neuronové sítě nahrazeny bloky reziduálními. Reziduální bloky mohou mít různou architekturu. Použitá architektura reziduálního bloku v ARU-Net je zobrazena na obrázku 2.10. Reziduální bloky využívají skip-spojení, které umožňuje zpětnou propagaci chyby a propagaci identity i pro velmi hluboké sítě [16]. Takto upravená U-Net síť se nazývá RU-Net.

Vzhledem k tomu, že se v dokumentech často vyskytuje různě velký text, je dále zaveden mechanismus `attention`, kterého využívá síť A-Net. A-Net je vícevrstvou konvoluční neuronovou sítí, jejímž výstupem je jedna mapa příznaků. Jednotlivé hodnoty této mapy dosahují vysokých hodnot pro oblasti, které obsahují vysokou míru informace.

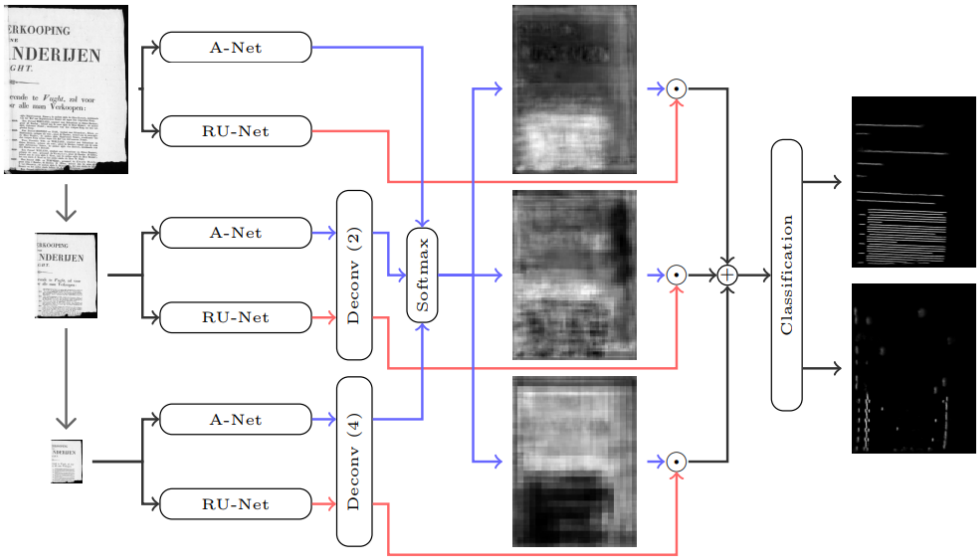
Spojením sítě A-Net a RU-Net vzniká síť ARU-Net. Jednotlivé operace



Obrázek 2.9: U-Net. Převzato z [16].



Obrázek 2.10: Reziduální blok. Na vstup je aplikována konvoluce. Výsledek je použit dvakrát. V prvním případě je přiveden na vstup aktivační funkce a dále zpracován dalšími konvoučními vrstvami. V druhém případě je výsledek s využitím skip-spojení přímo přiveden na vstup součtového uzlu (`Summation node`) [16].



Obrázek 2.11: ARU-Net. Převzato z [16].

můžeme vidět na obrázku 2.11 Síť A-Net společně s RU-Net bude aplikována pro různé škály. Napříč všemi škálami jsou používány stejné váhy sítě (sdílení vah). Označme jednotlivé škálované verze obrázku jako $I_1, I_2, I_4, I_8, \cdots, I_s$, kde s označuje faktor škály. Tyto obrázky jsou přivedeny na vstup A-Net a RU-Net. Výstupy z A-Net a RU-Net jsou vstupem dekonvolučních vrstev (odpovídající škály) pro získání mapy příznaků prostorové dimenze odpovídající dimenzi vstupu. $A_1, \cdots, A_s$ označme nadvzorkované mapy příznaků získaných z A-Net (označuje se jako `attention` mapa) a $RU_1, \cdots, RU_s$ z RU-Net. `Attention` mapa je následně normalizována softmax funkcí podle:

$$\hat{A}_i(y,x) = \frac{exp(A_i(y,x))}{\sum_{j\in\{1,2,\cdots,s\}} exp(A_j(y,x))}$$

Spojení `Attention` map příznaků a map příznaků z RU-Net je realizováno jako:

$$ARU = \sum_{i\in\{1,2,\cdots,s\}} RU_i \odot \hat{A}_i$$

kde $\odot$ představuje Hadamardův součin. Ukázka Hadamardova součinu pro matice o rozměrech 2×2:

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \odot \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} & a_{12}b_{12} \\ a_{21}b_{21} & a_{22}b_{22} \end{bmatrix}$$

Na obrázku 2.11 můžeme vidět, že pro různé škály stejného obrázku generuje A-Net různé výstupy. Jeví se, že se RU-Net zaměřuje na určitou velikost písma a A-Net odděluje oblasti s rozdílnou velikostí písma [16].

V případech, že je síť použita pro přiřazení jednotlivých tříd pixelům, je za celou ARU-Net zařazen klasifikátor, jehož výstupem je pravděpodobnost příslušnosti každého pixelu k jedné z definovaných tříd.

Přestože je síť navržena pro hledání řádků textu, je možné ji využít i pro další úlohy pro klasifikování jednotlivých pixelů, jako je například segmentace [16].

3 Data

3.1 Europeana

Projekt Europeana si klade za cíle umožnit na jednom centrálním místě snadný přístup k digitalizovanému, kulturnímu bohatství Evropy. Europeana Collection obsahuje přes 50 milonů záznamů, zahrnujících noviny, knihy, fotografie, umělecká díla, audionahrávky a další. Mnoho z těchto dat je volně dostupných pro jakékoliv využití. Pro přístup k těmto datům byla vytvořena webová stránka `europeana.eu`, kde je možné si data zobrazit, stáhnout, filtrovat podle kategorií atp.

Datová sada europena obsahuje celkem 528 stránek zdigitalizovaných historických novin ve formátu TIF. K datům jsou přiložena i metadata ve formátu PAGE (viz. sekce 4.1.1), obsahující informace o textových regionech, oddělovačích, pořadí čtení atp. Z této sady byla vybrána podmnožina 95 stránek, které jsou původem z různých druhů novin – například General-Anzeiger für Hamburg-Altona, Hamburger Anzeiger, Bozner Nachrichten a dalších. Jednotlivé stránky byly vybírány tak, aby se jejich struktura podobala dodané datové sadě Portafontium (viz. sekce 3.2). Datová sada Portafontium navíc obsahuje text psaný v německém jazyce. Z tohoto důvodů tvoří vybranou podmnožinu pouze stránky s německým textem. Titulní stránky typicky obsahují nadpis psaný velkým ozdnobným písmem. Některé nadpisy jsou navíc různě zahnuté nebo se nachází uvnitř nějakého rámování. Počet sloupců na stránce se pouhybuje v rozmezí od 2 do 12 sloupců. Součástí některých stránek jsou i obrázky rozprostřené i přes několik sloupců. Datová sada byla rozdělena na sadu trénovací a validační. Trénovací sada obsahuje 76 obrázků a sada validační 19.

3.2 Portafontium

V rámci projektu Porta fontium byly jako jedny z několika zdigitalizovány noviny z Aše – Ascher Zeitung pro roky 1866 až 1867 a 1869 až 1900. Jedná se o týdeník. Každé vydání obsahuje titulní stranu. Tato stránka začíná nadpisem psaným ozdobným písmem. Pro roky 1881 až 1887 je nadpis dokonce zahnutý do oblouku. Rozložení titulní stránky je pak různorodé. V některých případech je text jednosloupcový, jindy je text ve

dvou sloupcích. Stejně je tomu tak i u ostatních stránek. Často jsou zde stránky, které obsahují inzeráty. Tyto inzeráty jsou typicky skládány do dvou sloupců. V datech se také vyskytují dvousloupcové stránky obsahující pouze text, případně malé nadpisy.

Pro všechny typy stránek platí, že jsou jednotlivé sloupce a logicky související bloky vizuálně odděleny čárami (oddělovači).

Poskytnutá data obsahují spoustu nekvalitních digitalizací. Stránky například nejsou vycentrovány, na některých stránkách je text na krajích deformovaný (zahnutý), což indikuje ohnutý papír během procesu digitalizace. Na obrázcích je již také vidět časová degradace novin. Na stránkách je například otisknutý text z předchozí stránky, text je v některých místech značně vybledlý, či stránky byly ohnuty nebo jiným způsobem znehodnoceny.

K jednotlivým obrázkům nejsou v současné době poskytována žádná metadata, ze kterých by bylo možné stroji řici, kde se vyskytují textové bloky.

3.3 Datová sada PRImA Layout Analysis

Jedná se o soubor obrázků sloužících primárně pro vyhodnocení úspěšnosti metod z oblasti analýzy dokumentu [21]. Obsahuje realistické dokumenty s širokou škálou struktur. Zvláštní důraz je kladen na časopisy a technické/vědecké publikace, které se budou v budoucnu pravděpodobně digitalizovat.

Ke každému obrázku v datové sadě jsou přiložena metadata ve formátu PAGE (4.1.1), kde jsou definované ground-truth bloky. Během skenování dokumentů došlo k minimálnímu zanesení šumu. Stránky jsou vycentrované a nejsou deformované (ohnuté, potrhané). Na některých stránkách je vidět prosvícení obsahu z předchozí stránky. Datová sada obsahuje celkem 478 obrázků.

3.4 RDCL 2017 dataset

Jedná se o podmnožinu PRImA Layout Analysis Dataset (viz. 3.3). Data jsou rozdělena na trénovací a testovací. Trénovací množina obsahuje 7 obrázků s přiloženými metadaty ve formátu PAGE. Testovací množina pak obsahuje 75 obrázků. Tato datová sada byla vytvořena pro potřeby soutěže ICDAR 2017 [11].

4 Ground-truth

Pojem ground-truth je termín používaný v oblasti strojového učení. Tento termín představuje výsledek, kterému se snažíme přiblížit. Ground-truth je tedy klíčovým měřítkem při vyhodnocování výsledků získaných z procesu strojového učení a zároveň tvoří nezbytnou část trénovacích a testovacích dat.

Ground-truth v případě segmentace definuje, kde se na obrázku vyskytují hledané textové bloky, oddělovače atd. Textové bloky mají navíc přiřazeny další informace, např. zda se jedná o blok, ve kterém je nadpis nebo běžný text, jaké je pořadí čtení bloků, jaký konkrétní text blok obsahuj atp. Textové bloky mohou být rozděleny i na konkrétní řádky. Součastí ground-truth bývají i obrázky, či jiné grafické prvky.

4.1 Formáty ground-truth

Ground-truth informace je potřeba nějakým způsobem ukládat. Dříve se tyto informace ukládaly do obyčejného textového souboru. V současné době je trendem uchovávat informace v XML souboru, kde jsou tyto informace ukládány do stromové struktury.

4.1.1 PAGE

PAGE (Page Analysis and Ground-truth Elements) představuje reprezentaci obrázku, založenou na XML, která slouží pro uchování informací o struktuře stránky a jejím obsahu [19]. Tento formát umožňuje přesný popis jakéhokoliv elementu, který se může vyskytovat v naskenovaném dokumentu. Uvnitř stránky jsou všechny elementy definovány jako regiony (bloky) určitého typu. Mezi nejčastější typy patří:

- Image - obrázek.
- Line - čára.
- Graphic/Drawing - obrázek/kresba.
- Table - tabulka.
- Chart - graf.

- Separator - oddělovač, typicky čára mezi textovými elementy.
- Maths - matematické vzorce
- Noise - šum

Textové regiony je možné dále dělit na části:

- Lines - řádky.
- Words - jednotlivá slova.
- Glyphs - grafická vyjádřená písmena, čísla, znaky, interpunkční znamínka a jiné.

Každý element reprezentující region je popsán obrysem ve formě polygonu definovaný souřadnicemi (`coordinates`). Dále má každý region identifikátor `id` a podle typu regionu další specifická data. Textový region má například jazyk `language`, `script`, písmo `font`, směr čtení `reading direction`, barvu textu `text colour`, barvu pozadí `background colour`, typ `type` (například zda se jedná o hlavičku, odstavec, nadpis, patičko a tak dále) a mnoho dalších. Regiony mohou mít také nastavené pořadí čtení.

Ukázku PAGE formátu můžeme vidět v ukázce 4.1. Kořenovým elementem je element `PcGts`, který v sobě má element `Page`. V elementu `Page` je nadefinovaná cesta k obrázku, jeho šířka a výška. Dále si můžeme všimnout, že element `Page` má v sobě zanořené dva elementy reprezentující regiony.

Prvním elementem je `TextRegion` s id r0, typu `paragraph` (odstavec). Tento element má definované souřadnice s využitím `Coords`, kde jsou jednotlivé souřadnice oddělovány mezerou. Dále můžeme vidět element `TextEquiv`, kde je napsané, jaký text je v regionu obsažen.

Druhým elementem je `SeparatorRegion` s id r6, který v sobě také obsahuje informace o souřadnicích. Tento element reprezentuje stránkové oddělovače, jako je například vertikální čára mezi dvěma sloupci.

```
<PcGts
  xmlns="..."
  xmlns:xsi="..."
  xsi:schemaLocation="...">
  <Page
    imageFilename="i.png"
    imageWidth="1811"
```

```
    imageHeight="2431">
    <TextRegion id="r0" type="paragraph">
      <Coords points="220,466 220,499 128,499 128,466"/>
      <TextEquiv>
          <Unicode>Text regionu</Unicode>
      </TextEquiv>
    </TextRegion>
    <SeparatorRegion id="r6">
      <Coords points="107,521 107,523 193,523 193"/>
    </SeparatorRegion>
  </Page>
</PcGts>
```

Ukázka 4.1: Ukázka PAGE formátu. Pro zjednodušení byly hodnoty atributů `xmlns`, `xmlns:xsi` a `xsi:schemaLocation` přepsány na tři tečky.

4.1.2 GEDI

Stejně jako PAGE, formát GEDI (Groundtruthing Environment for Document Images) představuje XML strukturu, ve které jsou uloženy informace o struktuře stránky a blocích. Bloky jsou v tomto případě nazývány zonámi (`DL_ZONE`). Rozdílem oproti PAGE formátu je, že typy regionů nejsou nijak standardizované. Uživatel si typy regionů definuje sám a následně je přiřazuje k elementu `DL_ZONE` s použitím atributu `gedi_type`. Definice regionu bude vysvětlena na ukázce 4.2.

```
<?xml version="1.0" encoding="UTF-8"?>

<GEDI GEDI_date="07/29/2013" GEDI_version="2.4" xmlns="...">
<DL_DOCUMENT docTag="xml" NrOfPages="1" src="a.jpg">
    <DL_PAGE src="a.jpg"
    GEDI_orientation="0" pageID="1"
    gedi_type="DL_PAGE"
    height="2431" width="1811">
        <DL_ZONE gedi_type="TextRegion" id="4"
            height="105" width="1402" row="320" col="190">
        </DL_ZONE>
        <DL_ZONE gedi_type="TextRegion" id="2"
            polygon="(500,202);(500,51);(1289,51);(1289,202)"
            Text="Wochenblatt">
```

```
        </DL_ZONE>
        <DL_ZONE gedi_type="Separator" id="5"
            polygon="(104,427);(243,434);(609,436);">
        </DL_ZONE>
    </DL_PAGE>
</DL_DOCUMENT>
</GEDI>
```

Ukázka 4.2: Ukázka GEDI formátu. Pro zjednodušení byla hodnota atributu `xmlns` přepsána na tři tečky.

Kořenovým elementem je element `GEDI`, kde je mimo jiné napsáno, v jaké verzi nástroje GEDI (viz. 4.2.1) byl soubor vytvářen. Element `GEDI` v sobě obsahuje element `DL_DOCUMENT`, který údává cestu ke zdrojovému souboru a počet stránek. Například obrázek ve formátu TIFF může obsahovat i více než jednu stránku. Uvnitř následuje element `DL_PAGE`, který reprezentuje již samotnou jednu stránku obrázku. Jako atributy obsahuje například výšku a šířku obrázku a jak bylo s obrázkem při tvorbě GEDI souboru otáčeno.

Samotný element `DL_PAGE` pak obsahuje již elementy reprezentující samotné regiony `DL_ZONE`. V případě, že je region obdélníkového tvaru, je zde zadaná počáteční pozice atributy `row` a `col`. Následně je udávána výška a šířka regionu v atributech `height` a `width`. Pokud má však region tvar polygonu, tyto atributy se zde nevyskytují. Namísto toho je zde atribut `polygon`, kde jsou v závorkách udávány jednotlivé souřadnice bodů polygonu.

V ukázce 4.2 můžeme také vidět, že elementy `DL_ZONE` obsahují atribut `gedi_type`. Hodnoty v tomto atributu nejsou nikde definované formátem GEDI, ale uživatel si je definoval sám. Stejně tak u elementu `DL_ZONE` s id 2 si můžeme všimnout atributu `Text`. Tento atribut je také uživatelsky definovaný. Jak tedy můžeme vidět, tento formát nechává uživateli volnou ruku na atributy jednotlivých elementů. V některých případech to můžeme považovat za výhodu, nícméně pokud je potřeba mít striktnější formát, je formát PAGE lepším řešením.

4.2 Tvorba ground-truth

Poskytnutá datová sada Portafontium (viz. sekce 3.2), která je pro tuto diplomovou práci klíčová, nemá přiložená `ground-truth` data a tak je potřeba je připravit. V současné době jsou pro tvorbu `ground-truth` dat volně k dispozici nástroje GEDI a Aletheia. V těchto nástrojích je možné v grafickém prostředí označovat hledané bloky přímo na obrázku, zatímco se

na pozadí vytváří `ground-truth` soubor. Nyní budou jednotlivé nástroje popsány.

4.2.1 GEDI

Nástroj GEDI je vysoce konfigurovatelný nástroj pro anotaci naskenovaných dokumentů. Nástroj umožňuje vytvářet regiony různých typů. Vzhledem k charakteru formátu GEDI (viz. 4.1.2) je možné vytvářet vlastní typy a to v nastavení programu. K těmto typům je možné také přiřazovat další atributy, které mohou být užitečné například pro přepis textu obsaženém v daném regionu. Jednotlivé regiony mohou mít tvar obdélníku, nebo je možné je vytvořit jako polygon. Tyto regiony je možné také slučovat dohromady, nebo je naopak rozdělit. Ukázku programu můžeme vidět na obrázku A.1, kde si mimo jiné můžeme všimnout vizualizace pořadí čtení jednotlivých regionů. Výstupem z anotace je xml soubor ve formátu GEDI.

4.2.2 Aletheia

Aletheia je nástroj pro přesnou analýzu a anotaci naskenovaných dokumentů. Uživatelům poskytuje řadu automatizovaných a poloautomatizovaných nástrojů. Program například umožňuje automaticky vyhledat a označit regiony. Nicméně výsledné označený bylo na mnou zkoušených datech nepřesné. Nástroj kromě regionů umožňuje označovat i jednotlivé řádky, slova nebo dokonce i písmena. Regiony mají pak předdefinované různé atributy, které je možné během označování nastavit. Označení může mít tvar obdélníku či polygonu. Regiony je možné stejně jako v GEDI slučovat, či je rozdělovat. Program můžeme vidět na obrázku A.2. Nástroj umožňuje, a to dokonce automaticky, určit pořadí čtení jednotlivých bloků. Tato funkce je však nefunguje korektně na stránkách, kde jsou dva textové sloupce vedle sebe a to i přesto, že je mezi nimi oddělovač a je tak nutné toto pořadí nastavovat ručně. Výstupem programu je xml soubor ve formátu PAGE.

4.2.3 Volba nástroje

Během tvorby `ground-truth` byly vyzkoušeny oba dva nástroje. Jako přehlednější a intuitivnější se jeví program Aletheia. Na druhou stranu se v programu GEDI snadněji označují polygonové bloky. Polygon se stejně jako v Aletheia definuje body, které se zadávají klikáním do obrázku. GEDI však již během tvorby vizualizuje možné dokončení vytvářeného polygonu

a označuje plochu uvnitř. Aletheia pouze zobrazuje čary mezi za sebou jdoucími body. Oba dva programy poskytují pro označování stejné možnosti. Jako výhodu považuji u nástroje Aletheia, že jeho výstupem je soubor ve formátu PAGE, ve kterém jsou ukládány `ground-truth` data i v jiných datových sadách. (viz 3.1). Z tohoto důvodu byl tento nástroj zvolen.

4.2.4 Předzpracování obrázku před označováním

Před začátkem tvorby `ground-truth` je obrázek nejprve binarizován a transormován tak, aby řádky textu byly vodorovně. Během skenování obrázku mohlo totiž například dojít k tomu, že byl papír ve skenneru pootočen a řádky tak mohou být zahnuté. K tomu je použit nástroj Ocropy. [1] Jedná se o nástroj napsaný v jazyce Python a je navržený pro analýzu dokumentu a OCR. Takto předzpracované obrázky jsou následně anotovány.

4.3 Vytvořená datová sada

Tvorba `ground-truth` pro celou datovou sadu je časově velmi náročná. Označení textových bloků a oddělovačů na jedné stránce mi zabralo přibližně 45 minut. Z tohoto důvodu bylo označeno celkem 17 stránek. Vytvořená datová sada byla rozdělena na 3 podmnožiny: trénovací, validační a testovací. Během analýzy dodané datové sady bylo zjištěno, že se typy stránek dělí na titulní, stránky s textem ve dvouch sloupcích a dvousloupcové stránky s inzeráty. V tabulce 4.1 je navržené rozdělení označené datové sady, které bude použito pro trénování a vyhodnocení modelů. Ukázka typů stránek je na obrázku 4.1.

Typ sady	Titulní	Dvousloupcové	Inzeráty
Trénovací	2	5	2
Validační	1	–	2
Testovací	2	2	1
Celkem	5	7	5

Tabulka 4.1: Typy stránek v jednotlivých podmnožinách.

[1] https://github.com/tmbdev/ocropy

Obrázek 4.1: Vlevo je titulní strana, uprostřed dvousloupcový text a vpravo stránka s inzeráty.

5 Přístupy k vyhodnocení segmentace

Jelikož segmentace stojí na začátku procesu analýzy dokumentu, je správnost výstupu klíčová pro všechny další navazující kroky analýzy dokumentu. K vyhodnocení je potřebné mít vždy dvojici obrázků: očekávaný výstup a výsledek segmentace. Podle [10] se v současné době pro vyhodnocení úspěšnosti používají dva přístupy: blokově založené a pixelově založené. Každý přístup se liší použitými metrikami, požadavky na vstup a výpočetní náročností. Nyní budou jednotlivé přístupy popsány.

5.1 Pixelově založené

V pixelově založeném přístupu pro vyhodnocení je každý pixel obrázku označen třídami, které reprezentuje. Vstupem metrik jsou tedy dva obrázky (ground-truth a výsledek segmentace) s metadaty obsahujícími informace, jakých jednotlivých klasifikačních tříd pixel nabývá. Ground-truth obrázek je při vyhodnocení reprezentován jako:

$$L = (\{c \in C\}_i \mid 1 \leq i < n) \tag{5.1}$$

C představuje množinu unikátních tříd v dokumentu a n je celkový počet pixelů v obrázku. Stejným způsobem je reprezentován i obrázek získaný segmentací, označme ho jako $L^{'}$. Během vyhodnocení je tedy porovnáváno přiřazení tříd k pixelu v ground-truth obrázku vůči přiřazení v segmentaci, pixel po pixelu. Při vyhodnocení přiřazení tříd pixelům mohou nastat následující případy:

- True Positive (TP): pixelu v L a $L^{'}$ byla přiřazena stejná třída c_i.
- True Negative (TN): pixelu v L a $L^{'}$ nebyla třída c_i přiřazena.
- False Positive (FP): pixelu v $L^{'}$ byla přiřazena třída c_i, ale v L nikoliv.
- False Negative (FN): pixelu v $L^{'}$ nebyla přiřazena třída c_i a zároveň pixelu v L byla třída c_i přiřazena.

V následující sekci budou popsány metriky, které podle [4] patří mezi hlavní metriky používané v oblasti segmentace, doplněné o metriku představenou v [26].

5.1.1 Metriky

Přesná shoda

Přesná shoda (Exact match) představuje nejvíc striktní metriku, která počítá, kolik je zde přesných shod mezi označením v ground-truth a obrázku získaném segmentací.

$$EM = \frac{\mid \{p \mid L_p^{'} = L_p\}}{N} \tag{5.2}$$

V případě, že je každému pixelu přiřazená pouze jedna třída, jedná se o vzoreček pro přesnost (accuracy).

Hammingovo skóre

Hammingovo skóre oproti metrice přesné shody poučítá i s částečnou shodou.

$$H = 1 - \frac{1}{n \cdot \mid C \mid} \cdot \sum \mid L_i^{'} \oplus L_i \mid \tag{5.3}$$

Přesnost

Přesnost (precision) klasifikátoru je definovaná jako pravděpodobnost, že náhodnému pixelu v obrázku získaném segmentací je správně přiřazena třída c.

$$P_c = \frac{TP_c}{TP_c + FP_c} \tag{5.4}$$

Úplnost

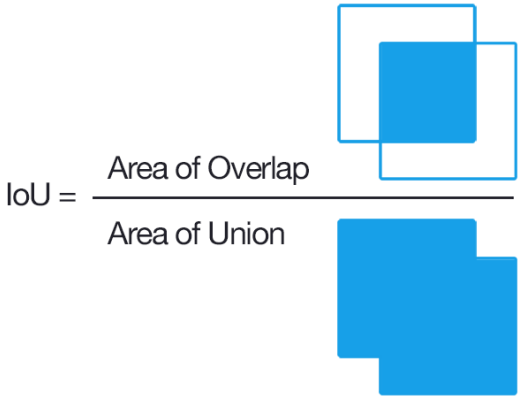
Úplnost (recall) klasifikátoru je definovaná jako pravděpodobnost, že se náhodný pixel z ground-truth obrázku s třídou c bude vyskytovat i v obrázku získaném segmentací.

$$R_c = \frac{TP_c}{TP_c + FN_c} \tag{5.5}$$

F1-skóre

F1-skóre odpovídá harmonickému průměru mezi přesností a úplností. Přesnosti a úplnosti je tak dána stejná váha.

$$F_{1_c} = \frac{2 \cdot TP_c}{2 \cdot TP_c + FP_c + FN_c} = 2 \cdot \frac{P_c \cdot R_c}{P_c + R_c} \tag{5.6}$$



Obrázek 5.1: Vizualizace Jaccardova indexu podobnosti. Převzato z [23].

Jaccardův koeficient

Jaccardův koeficient podobnosti [17] představuje metriku používanou pro porovnání podobnosti a odlišnosti množin.

$$J = IU_c = \frac{TP_c}{TP_c + FP_c + FN_c} \tag{5.7}$$

Vizualizaci rovnice můžeme vidět na obrázku 5.1. V [4] zmiňují, že přestože odrazují čtenáře od porovnávání výkonnosti algoritmů pouze jednou metrikou, jedná se o metriku, kterou by pro tento případ použili.

Průměrování výsledků

Výše zmíněné metriky určují výsledky pro každou třídu c zvlášť. Označme výsledek metriky pro třídu c jako ψ_c. Z těchto dílčích výsledků je možné vypočítat průměrnou hodnotu pro všechny třídy c jako:

$$\psi^M = \frac{1}{|C|} \cdot \sum_c \psi_c$$

Přesnost pixelů v popředí

Pokud se podíváme na obrázek 5.2, lze říci, že všechny výstupy jednotlivých segmentací označily text stejně dobře. Liší se pouze předpovězenou oblastí kolem textu. V případě použití metrik jako jsou přesnost, úplnost či Jaccardův koeficient (tyto metriky jsou rozebrány v sekci 5.1) se však nejedná o totožné výsledky. Pro určení úspěšnosti modelu tak byla zavedena metrika, která určuje přesnost určení popředí (*Foreground Pixel Accuracy - FgPA*). Během vyhodnocení se v této

Obrázek 5.2: Různé výsledky segmentace stejného textu. Převzato z [26].

metrice berou v úvahu pouze pixely, které jsou označeny jako popředí. V tomto případě se jedná o pixely jednotlivých písmen.

Tato metrika byla zavedena pro vyhodnocení a porovnání výsledků plně konvoluční neuronové sítě v [26]. Nechť $\delta_x = 1$ v případě, že se označení pixelu v `ground-truth` shoduje s označením v obrázku získaném segmentací. Dále je definováno $b_x = 1$, pokud je pixel x v `ground-truth` obrázku označení jako popředí. Přesnost určení pixelů v popředí je pak definováno jako:

$$FgPA = \frac{\sum_x \delta_x \cdot b_x}{\sum_x b_x}$$

Nástroj pro vyhodnocení

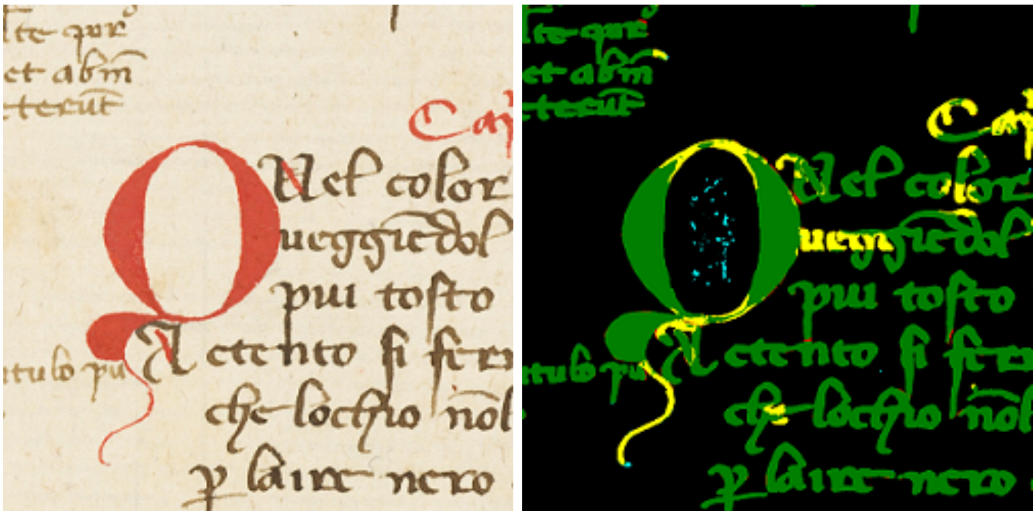
Pro potřeby vyhodnocení dosažených výsledků algoritmů, přihlášených do soutěže RDCL 2017, byl vyvinut program DIVA layout analysis evaluator [5]. Tento program je uvolněn široké veřejnosti a slouží pro vyhodnocení výsledků segmentace s využitím pixelově založených metod. Zároveň obsahuje všechny výše zmíněné metriky kromě přesnosti pixelů v popředí ($FgPA$).

Vstupem programu jsou tzv. pixel-label obrázky, kde třídy přiřazené pixelu jsou uloženy v modrém kanálu, zatímco červená a zelená je nastavená na 0. Hodnota modrého pixelu může například nabývat hodnot:

- 0x000001: pokud se jedná o pozadí.
- 0x000002: pokud se jedná o oddělovač.
- 0x000004: pokud se jedná o text.
- 0x000006: pokud se jedná o oddělovač a zároveň text.

Program tedy na vstupu přijímá takto upravený ground-truth obrázek a stejně tak i výsledek segmentace. Výstupem jsou výsledky výše zmíněných metrik. Nástroj zároveň umožňuje výsledky vyhodnocení vizualizovat, jako je tomu na obrázku 5.3. Barevné hodnoty pixelů ve vizualizaci představují:

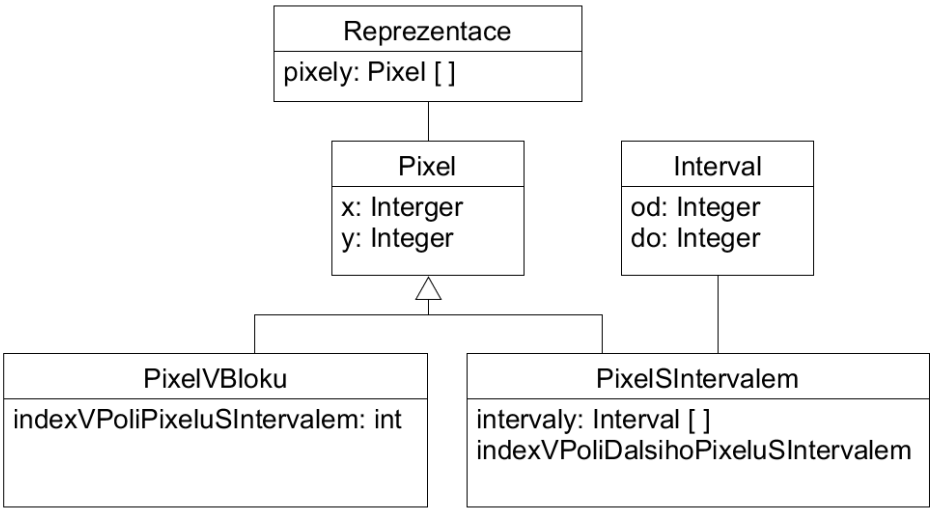
- Zelená: pixel je správně označen jako popředí se správnou třídou.
- Žlutá: pixel je správně označen jako popředí, ale je mu přiřazena špatná třída (např. Text místo komentáře)
- Černá: pixel je správně označen jako pozadí.
- Červená: pixel, který měl být označen jako pozadí byl označen jako popředí
- Modrá: pixel, který byl označen jako popředí byl označen jako pozadí



Obrázek 5.3: Vlevo je originální obrázek, vpravo je vizualizace vyhodnocení. Zdroj [4].

5.2 Blokově založené

V [10] je popsáno, jakým způsobem je možné provádět blokově založené vyhodnocování. Analýza výkonnosti segmentačních metod je rozdělena do tří částí. Nejprve je potřeba všechny regiony reprezentovat s pomocí speciální intervalové reprezentace, která umožňuje další efektivní zpracovávání. Na základě těchto reprezentací jsou následně nalezeny vztahy mezi ground-truth bloky a bloky v obrázku získaném segmentací. Na závěr jsou identifikovány, kvantifikovány a klasifikovány chyby. Nyní budou jednotlivé kroky popsány.



Obrázek 5.4: Možná struktura pro uchování intervalové reprezentace.

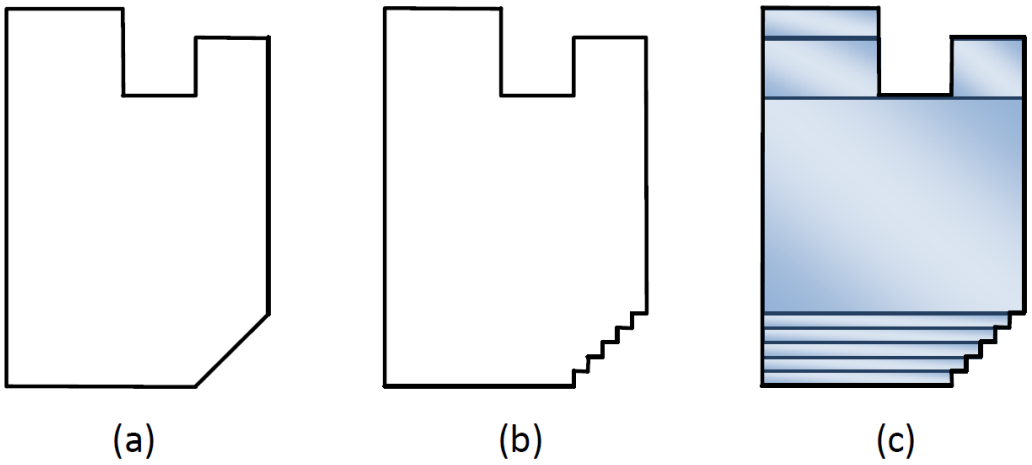
5.2.1 Reprezentace regionu

Podle [10], každý blok stránky je reprezentován svým obrysem ve formě polygonu. Díky tomu je možné reprezentovat jakýkoliv blok velmi přesně. Pro efektivní porovnání označených polygonů mezi sebou, je potřeba zvolit vhodnou datovou strukturu.

Nejprve se polygon převede na pravoúhelník (anglicky isothetic polygon), tedy polygon, který obsahuje pouze horizonatální a vertikální hrany. Vzhledem k tomu, že digitální obrázky dokumentů jsou rastrové, tato transformace nevede ke ztrátě informace. Z takto upraveného polygon se následně určí jeho intervalová reprezentace [6]. Pro uchování této reprezentace je použito jednorozměrné pole o velikosti počtu pixelů obrázku. Navržená datová struktura pro reprezentaci je k vidění na obrázku 5.4. Můžeme zde vidět, že datové struktury obsažené v tomto poli mohou být dvojího typu.

První struktura (označme si jí jako S_1), která v sobě kromě informace o souřadnici daného pixelu, nese i informace o intervalech, které se na řádku pixelu vyskytují. Tyto intervaly jsou definovány souřadnicemi od do, tedy představují, kde daný interval začíná a kde končí. Dalším atributem této struktury je ukazatel na další strukturu tohoto typu. Z rozdílů y-ových souřadnic těchto struktur je možné určit velikost bloku.

Další struktura (S_2) v sobě kromě souřadnic daného pixelu obsahuje pouze ukazatel na strukturu S_1. Díky tomuto ukazateli je možné bez nutnosti výpočtu určit, do kterého bloku daný pixel spadá. Jak můžeme vidět na obrázku 5.5 c), po procesu intervalové dekompozice, je originální



Obrázek 5.5: Proces intervalové dekompozice: (a) Originální polygon; (b) Pravoúhelník; (c) Intervalová reprezentace. Zdroj [10].

polygon rozdělen na několik obdélníků. Počet těchto obdélníků závisí na složitosti tvaru původního polygonu. Například pokud je původní polygon pouze obdélník, reprezentace bude tvořena pouze jedním intervalem. Jak ale můžeme vidět na obrázku 5.5 c), celkem jednoduchý tvar je ve výsledku reprezentován devíti intervaly.

5.2.2 Určení vztahů mezi regiony

V tomto kroku se sleduje zda a jak moc se bloky v ground-truth obrázku překrývají s bloky ze segmentace. Informace o překrytí jsou uloženy do dvou vyhledávacích tabulek. Klíčem první tabulky jsou bloky z ground-truth obrázku a hodnotami jsou bloky z obrázku získaným segmentací, které ground-truth blok překrývají. Druhá tabulka je přesně obráceně, tedy klíčem jsou bloky ze segmentace a hodnotami jsou bloky z ground-truth obrázku, které bloky ze segmentace překrývají. Tyto tabulky jsou použity k identifikaci následujících situací [7]:

- Sloučení: Blok ze segmentace překrývá více než jeden ground-truth blok.
- Rozdělení: Ground-truth blok je překryt více než jedním bloku ze segmentace.
- Minutí / částečné minutí: Ground-truth blok není, nebo je pouze částečně pokryt bloky segmentace.
- Chybná detekce: Blok ze segmentace nepřekrývá ground-truth blok.

Dále jsou používány metriky přesnost, úplnost a F-míra.

5.2.3 Kvantifikace a kvalifikace chyb

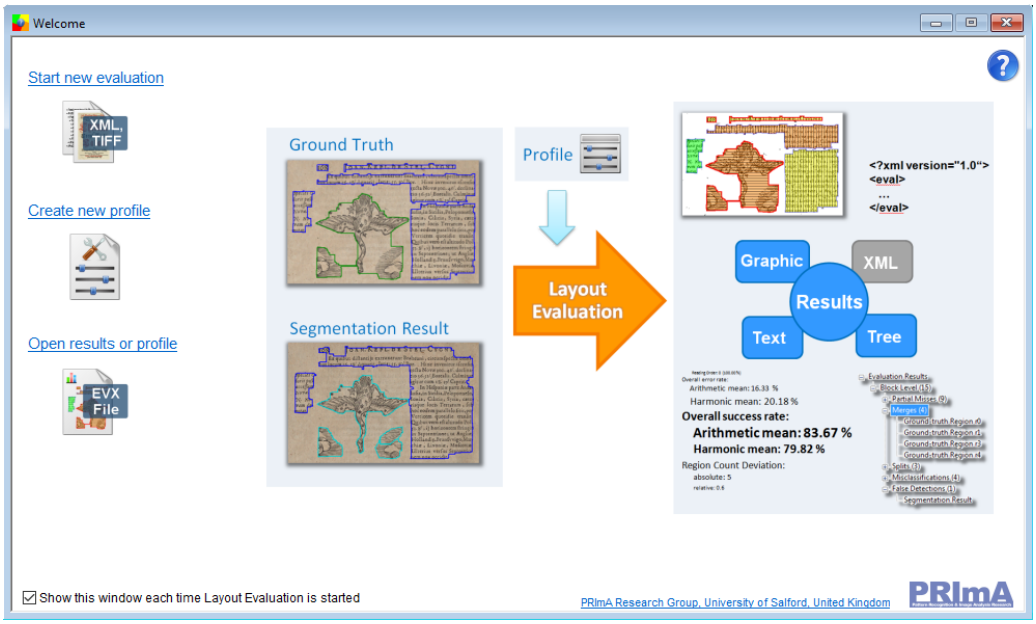
Na základě měřítek zmíněných v předchozí části můžeme jednotlivé chyby kvantifikovat. Chyby jsou pak rozděleny podle jejich významu. Jsou zde dva typy významu chyb [10]. Prvním je implicitní kontextově závislý význam, který představuje logické a geometrické vztahy mezi bloky. Příkladem mohou být přípustná a nepřípustná sloučení. Sloučení dvou vertikálně přilehlých odstavců v jednom sloupci může být označeno jako přípustné, jelikož toto sloučení sotva ovlivní výsledek OCR. Sloučení odstavců mezi dvěma různými sloupci už ale přípustné není, jelikož tímto sloučením bude výsledkem OCR nesmyslný tok slov. Pro určení, zda se jedná o přípustnou nebo nepřípustnou chybu, jsou využívány i další infomrace jako je pořadí čtení.

Druhý význam si uživatel definuje sám podle toho, jaké jsou jeho aktuální potřeby vyhodnocení (vyhodnocovací scénáře). Například, pokud jsou vyhodnocovány výsledky programu, který je zaměřen na vytvoření obsahu, má správnost určení bloku s číslem stránky a nadpisů největší váhu. Naopak správné určení grafického bloku je v tomto případě nepodstatné.

5.2.4 Nástroj PRImA Performance Evaluation

Tento volně dostupný nástroj vychází z výše popsaných kroků. Regiony jsou reprezentovány s využitím intervalů, následně je určeno překrývání bloků a chybám jsou přiřazené patřičné váhy. Vstupem vyhodnocení je ground-truth soubor ve formátu PAGE (viz. 4.1.1) a výsledek segmentace, také ve formátu PAGE. Nástroj umožňuje si vytvářet vlastní vyhodnocovací scénáře. Zároveň běžně používané scénáře jsou již od autorů předdefinované:

- Všeobecné rozpoznávání: Scénář pro běžné rozpoznávací úlohy, obsahuje všechny chyby a typy regionů s nastavenými váhami.
- Fulltextové rozpoznávání: Scénář zaměřen pouze na textové bloky.
- Indexování textu: Scénář zaměřený na výstupy pro indexování a hledání klíčových slov. Chybám jako je slučování a rozdělování jsou nastaveny nižší váhy.
- Obrázky a grafy: Scénář zaměřený na obrázky. Ostatní typy bloků jsou ignorovány.



Obrázek 5.6: Nástroj PRImA Performance Evaluation.

6 Volba metod pro implementaci

Výstupem metod pro segmentaci je klasifikace na úrovni pixelů – v našem případě se jedná o rozdělení pixelů do dvou tříd: text a ostatní. Následně je možné vzít původní obrázek a vymaskovat ho s označeným textem, čímž docílíme toho, že v původním obrázku zůstane pouze text. Označme výsledný obrázek segmentace jako I_{seg}. Jednotlivé bloky je možné nalézt v obrázku I_{seg} použitím algoritmů pro analýzu spojených komponent. Může však dojít k jevu, kdy textové bloky, které se nachází vertikálně těsně vedle sebe, spojí klasifikátor do jednoho bloku. Z tohoto vyplývá, že je nutné v obrázku nalézt i oddělovače. Může se jednat o klasickou černou čáru oddělující jednotlivé bloky, avšak bloky mohou být také děleny pouze vynecháním místa, kdy má oddělovač barvu pozadí (bílou). Analýza dodané datové sady Portafontium ukázala, že oddělovače jsou zde reprezentovány pouze černými čarami. S využitím oddělovačů je již možné spojené bloky rozdělit. Algoritmus rozdělení bloků podle oddělovačů bude popsán v sekci 6.2.6. Po správném nalezení textových bloků je již snadné aplikací algoritmů (viz. sekce 6.3) rozdělit bloky na jednotlivé řádky textu. Nalezené řádky pak mohou být vstupem OCR.

Z výše popsaných požadavků vyplývá, že bude potřeba vytvořit klasifikátory pro tři typy úloh: hledání textových bloků, hledání oddělovačů a pro hledání řádků v textu. Z prostudovaných metod používaných pro segmentaci byly vybrány plně konvoluční sítě a to konkrétně ARU-Net 2.3.2, U-Net a upravený U-Net 2.3.1. Síť ARU-Net se osvědčila při detekci řádek textu v historických dokumentech [16]. Textové bloky v těchto dokumentech úspěšně detekuje upravený U-Net popsaný v sekci 2.3.1. Na základě experimentů budou vybrány nejlepší modely pro zmíněné úlohy.

V rámci diplomové práce byla z dodaných dat vytvořena datová sada s anotovanými obrázky (viz sekce 4.3). Pro trénování bylo vyhrazeno celkem 10 obrázků. Obecně je však 10 stránek pro trénování neuronové sítě málo. Klasifikační modely tak budou nejprve natrénovány na podmnožině datové sady Europena (viz sekce 3.1), která obsahuje 95 obrázků, z toho 76 z nich tvoří trénovací sadu a 19 sadu testovací. Vzápětí pak proběhne dotrénování modelu na vytvořené datové sadě Portafontium.

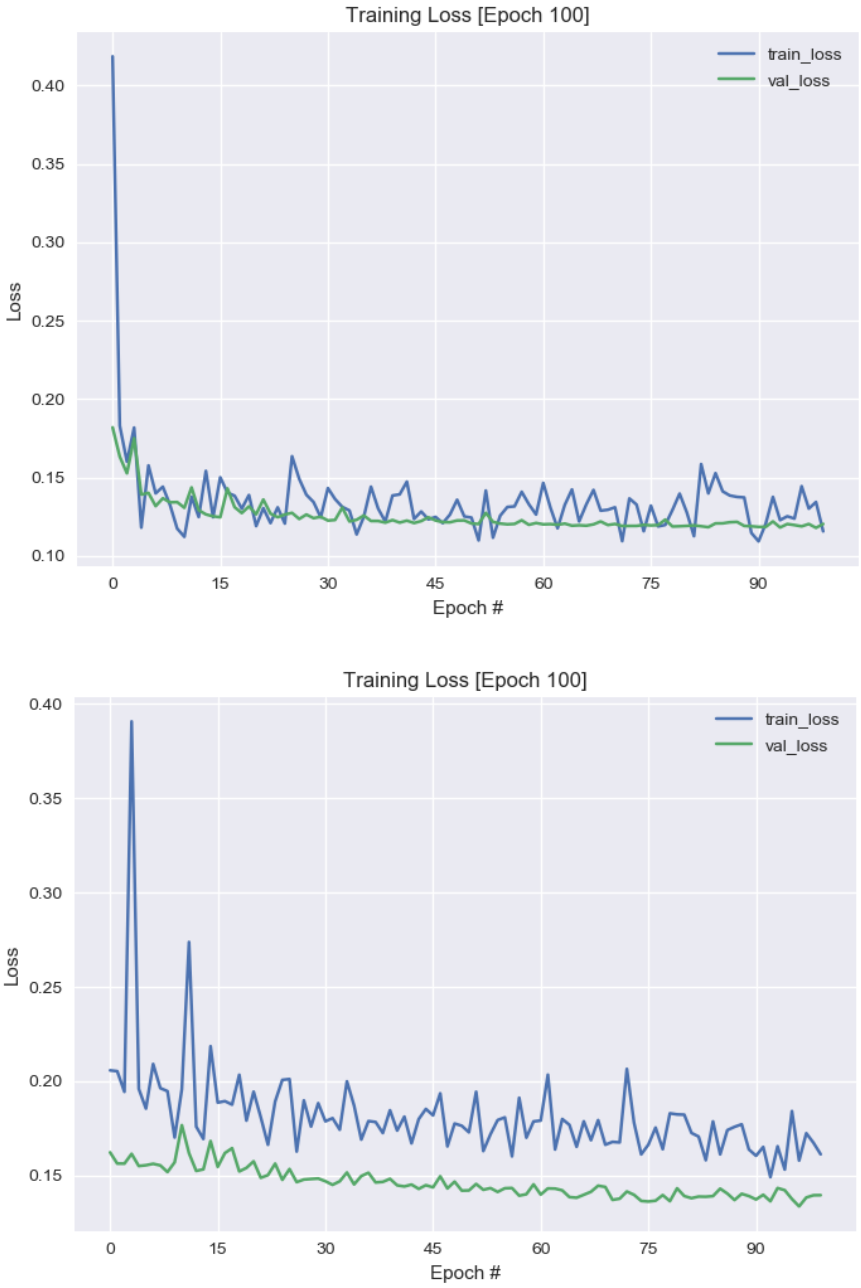
6.1 Segmentace textu

Datová sada pro trénování klasifikátoru pro segmentaci textu je tvořena dvojcemi obrázků: obrázkem, který má být klasifikován a obrázkem s označenými, hledanými, textovými bloky (`ground-truth`). Pixely tvořící textové bloky jsou reprezentovány bílou barvou, zbylé jsou černé. Odpovídajícím způsobem je tvořena i testovací sada.

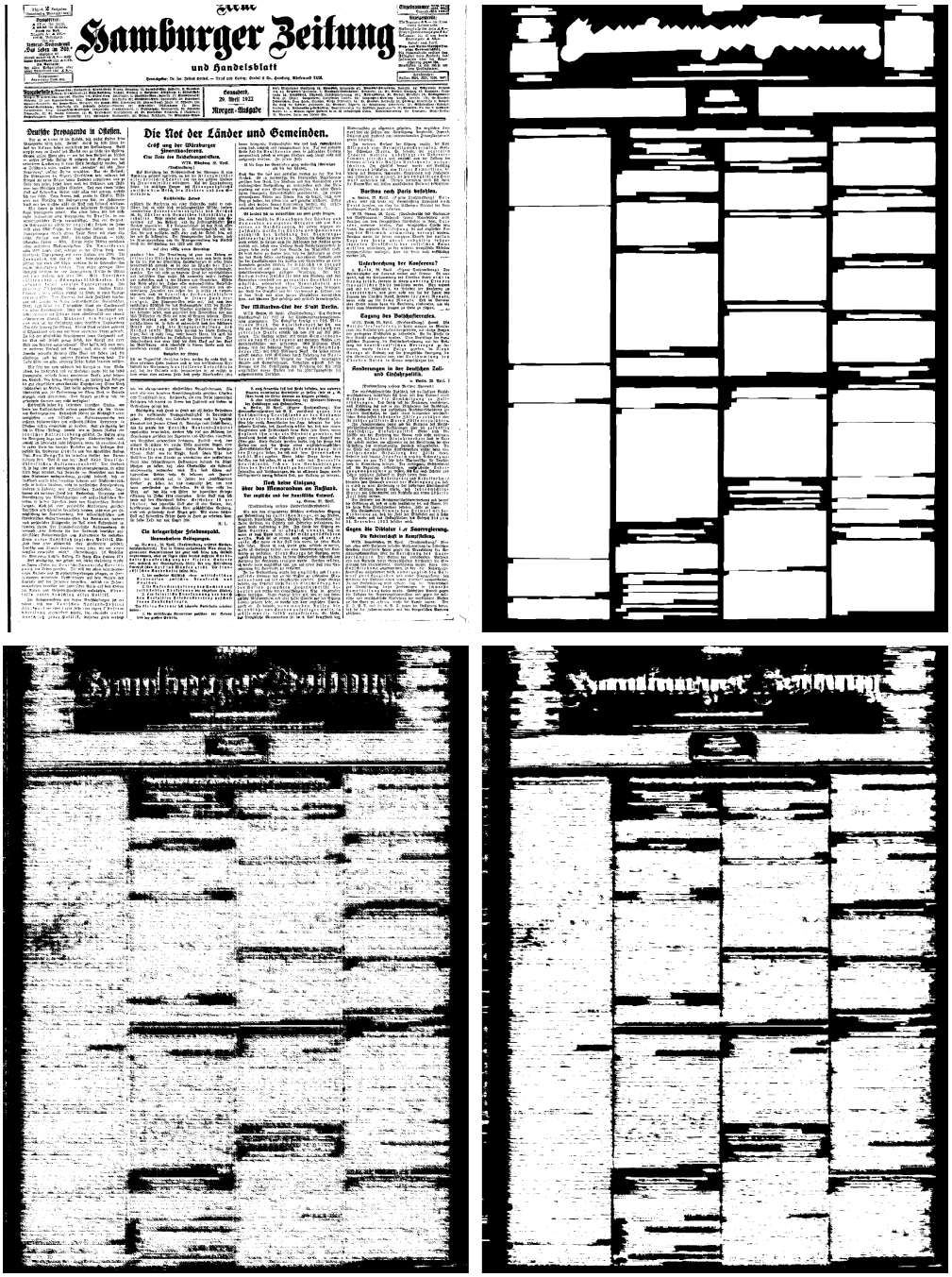
6.1.1 ARU-Net

Jak bylo zmíněno v sekci 2.3.2, síť ARU-NET využívá tzv. *attention*, což by mělo podle [16] vést k lepšímu naučení detekce různě velkého textu. Průběh trénování můžeme vidět na obrázku 6.1. Můžeme si všimnout, že se ztráta na validačních datech postupně snižuje, nicméně s narůstajícím počtem epoch dochází pouze k minimálnímu zlepšení. Z tohoto důvodu bylo trénování ukončeno vždy po 100. epoše. Pro model natrénovaný na datové sadě Europeana, označme ho jako ARU_{eutext}, bylo dosaženo nejmenší ztráty na validačních datech v 99. epoše a to konkrétně $0,1179$. Tento model byl dále dotrénován na datově sadě `Portafontium`, kdy bylo dosaženo nejmenší ztráty $0,1336$ po dokončení 97. epochy. Označme ho jako ARU_{pftext}.

Na obrázcích 6.3 a 6.2 můžeme vidět ukázky výsledků segmentace. Z ukázek je vidět, že dotrénování modelu na datech `Portafontium` vede k ucelenějšímu označování textových bloků. Zároveň jsou v některých místech odstraněny osamocené pixely nebo jsou naopak zvýrazněny. Můžeme si všimnout, že síť také v některých případech označuje oddělovací čáry, přestože to nebylo předmětem učení. Co se týče nadpisů a textů psaných větším písmem, síť tyto bloky označuje daleko menší oblastí, než je očekáváno. V obrázku s textem 6.2 se v levém sloupci vyskytuje lehce vybledlý text. Síť i přesto tento blok označuje a to především v dotrénované verzi. Dotrénovaná verze také lépe odděluje jednotlivé textové bloky a to i v případě, že jsou bloky těsně vedle sebe nebo blízko nad sebou.



Obrázek 6.1: Průběh trénování síte ARU-NET pro segmentaci textových bloků na datové sadě Europeana můžeme vidět na horním obrázku. Následné dotrénování na datové sadě Portafontium je na obrázku spodním.



Obrázek 6.2: ARU-Net segmentace textu. Ukázka výsledků segmentace na datové sadě Europeana. Obrázky: Originální obrázek, očekávaný výstup, výsledek modelu ARU_{eutext}, výsledek modelu ARU_{pftext}.

Wochenblatt
für
Asch und Umgegend.

№ 1. Sonnabend, den 6. Januar 1866. III. Jahrgang.

Amtliche Mittheilungen.

Executive Feilbietung.

Der Mann in Eisen.

Erzählung von Hans Wachenhusen.

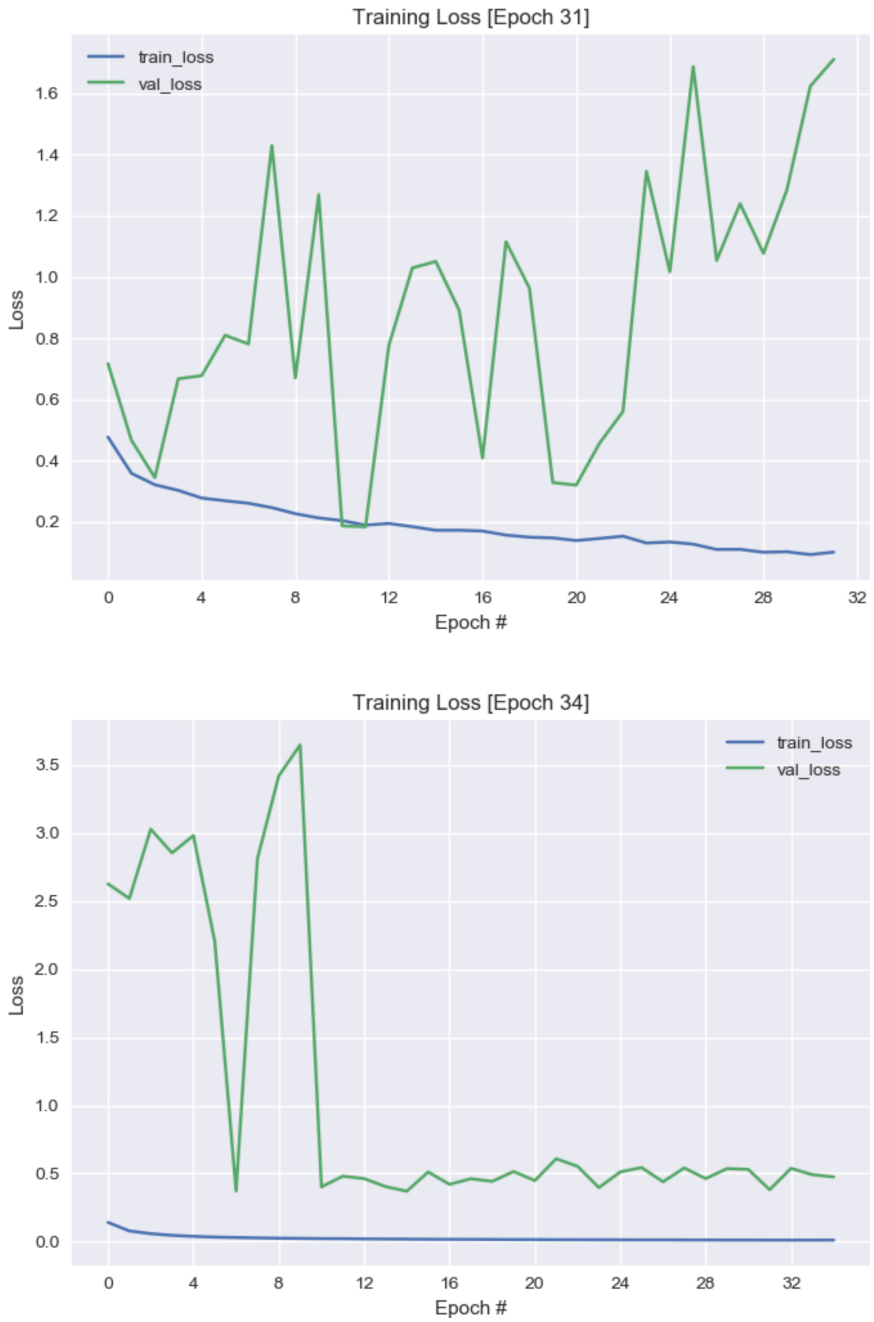
(Fortsetzung)

Obrázek 6.3: ARU-Net segmentace textu. Ukázka výsledků segmentace na datové sadě Portafontium. Obrázky: Originální obrázek, očekávaný výstup, výsledek modelu ARU_{eutext}, výsledek modelu ARU_{pftext}.

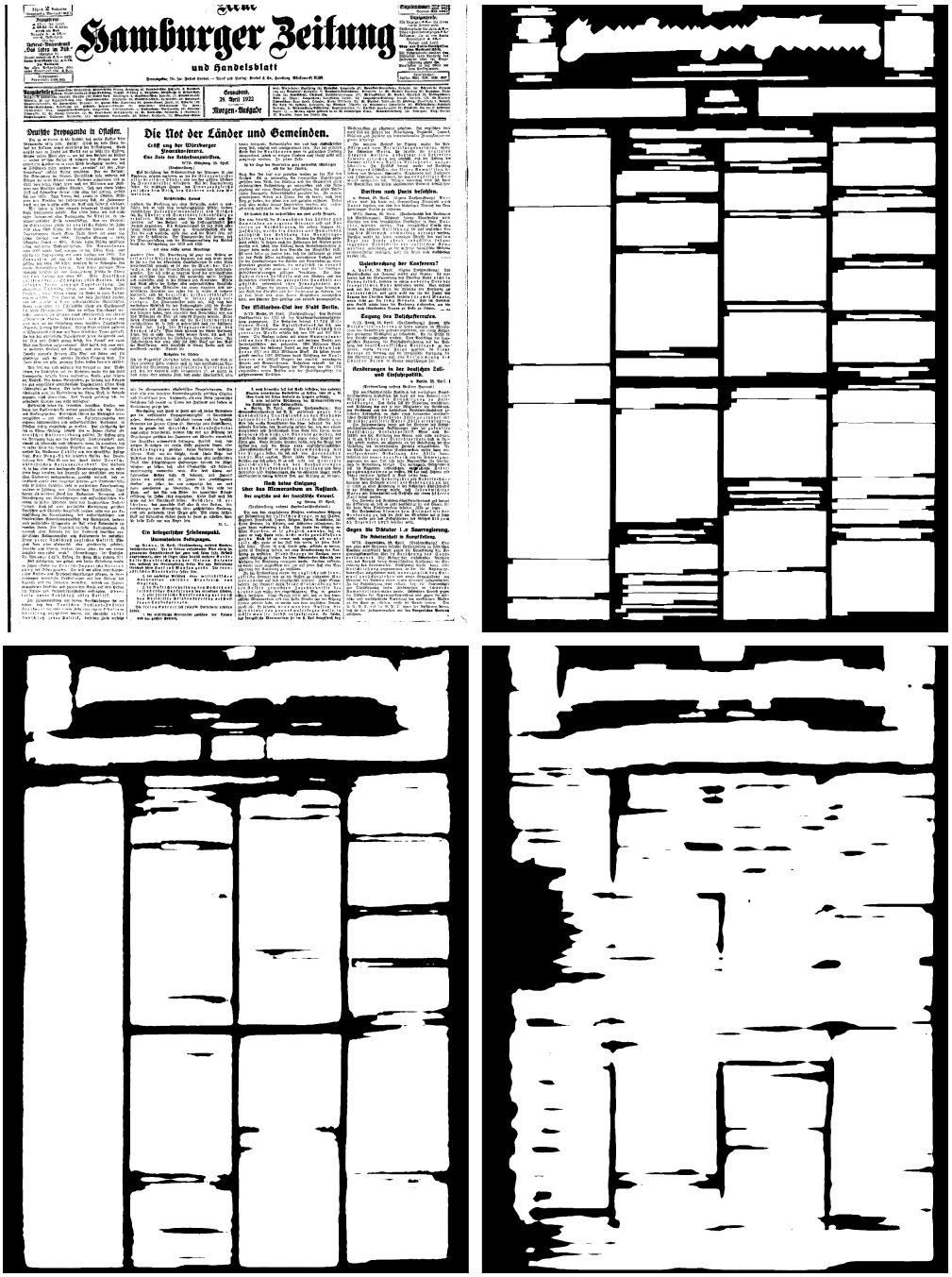
6.1.2 U-Net

Průběh trénování sítě U-Net můžeme vidět na obrázku 6.4. Nejlepší ztráty na validačních datech bylo dosaženo pro datový set Europeana po dokončení 11. epochy a to konkrétně $0,1858$. Přestože ztráta na trénovacích datech stále klesá, ztráta na validačních začíná narůstat, což může značit přeučení. Tento model označme jako U_{eutext}. Následné dotrénování na datové sadě Portafontium dosáhlo nejlepší validační ztráty po dokončení 13. epochy a to konkrétně $0,4557$. Model bude označen jako U_{pftext}.

Ukázky dosažených výsledků můžeme vidět na obrázcích 6.6 a 6.5. Na obrázku 6.5 si všimněme, že dotrénování vede k znepřesnění výsledků pro obrázek z datové sady Europeana. Jednotlivé bloky jsou spojovány do jednoho. Model U_{pftext} si navíc ani neporadí s lehce vyblednutým textem a navíc označuje oddělovače. Podle předpokladů model U_{pftext} lépe označuje obrázky z datové sady Portafontium. Na ukázce 6.6 můžeme vidět, že jsou přesněji označovány jednotlivé odstavce včetně odsazení první věty. Lépe jsou také označovány nadpisy a to i v případech, kdy jsou psány drobnějším písmem.



Obrázek 6.4: Průběh trénování síte U-Net pro segmentaci textových bloků na datové sadě Europeana můžeme vidět na horním obrázku. Následné dotrénování na datové sadě Portafontium je na obrázku spodním.



Obrázek 6.5: U-Net segmentace textu. Ukázka výsledků segmentace na datové sadě Europeana. Obrázky: Originální obrázek, očekávaný výstup, výsledek modelu U_{eutext}, výsledek modelu U_{pftext}.

Wochenblatt
für
Asch und Umgegend.

№ 1. Sonnabend, den 6. Januar 1866. III. Jahrgang.

Amtliche Mittheilungen.

OKRESNÍ ARC...
V CHEBU
1339/52

Executive Feilbietung.

Der Mann in Eisen.
Erzählung von Hans Wachenhusen.
(Fortsetzung.)

Executive Feilbietung.

Executive Feilbietung.

Executive Feilbietung.

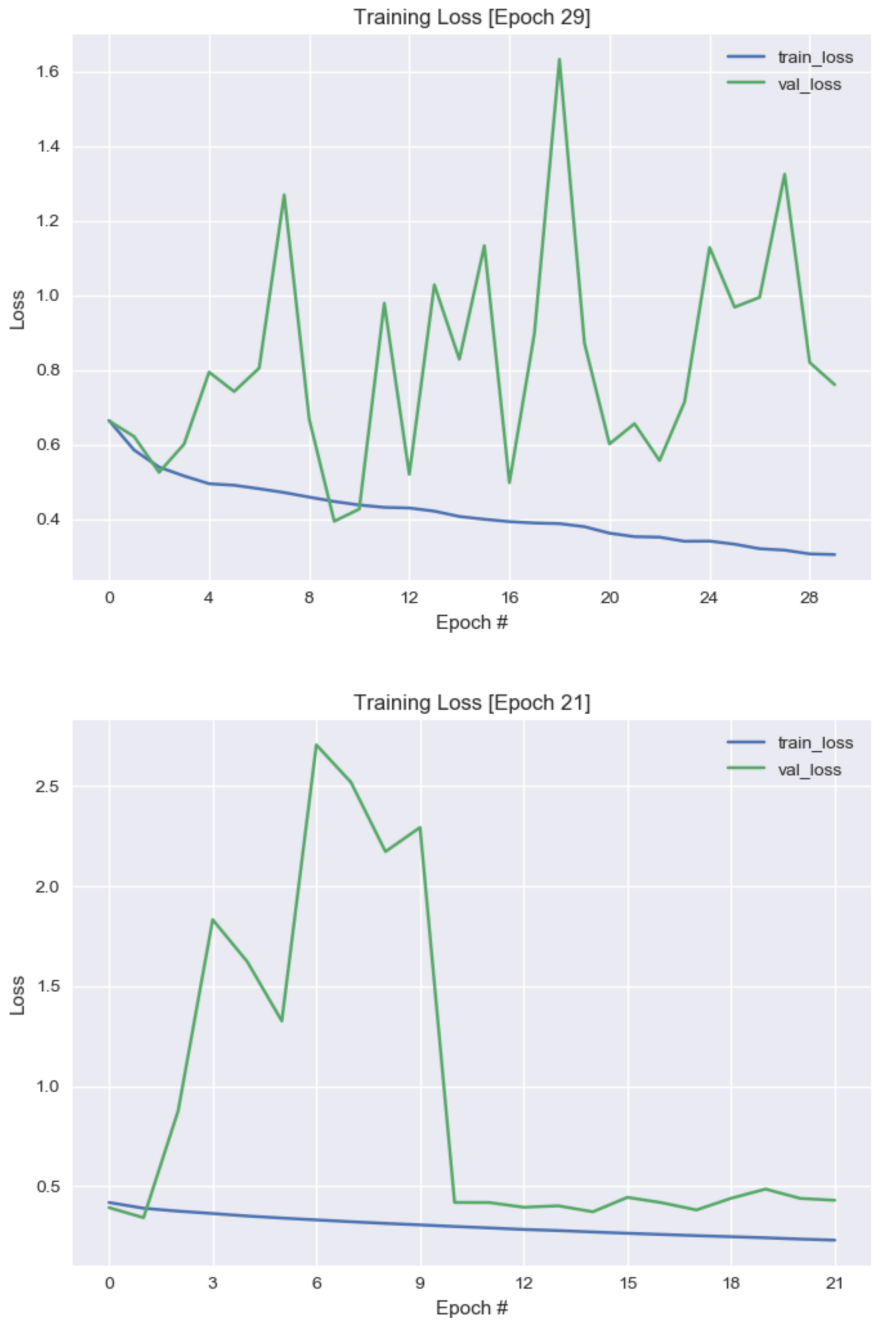
Executive Feilbietung.

Obrázek 6.6: U-Net segmentace textu. Ukázka výsledků segmentace na datové sadě Portafontium. Obrázky: Originální obrázek, očekávaný výstup, výsledek modelu U_{eutext}, výsledek modelu U_{pftext}.

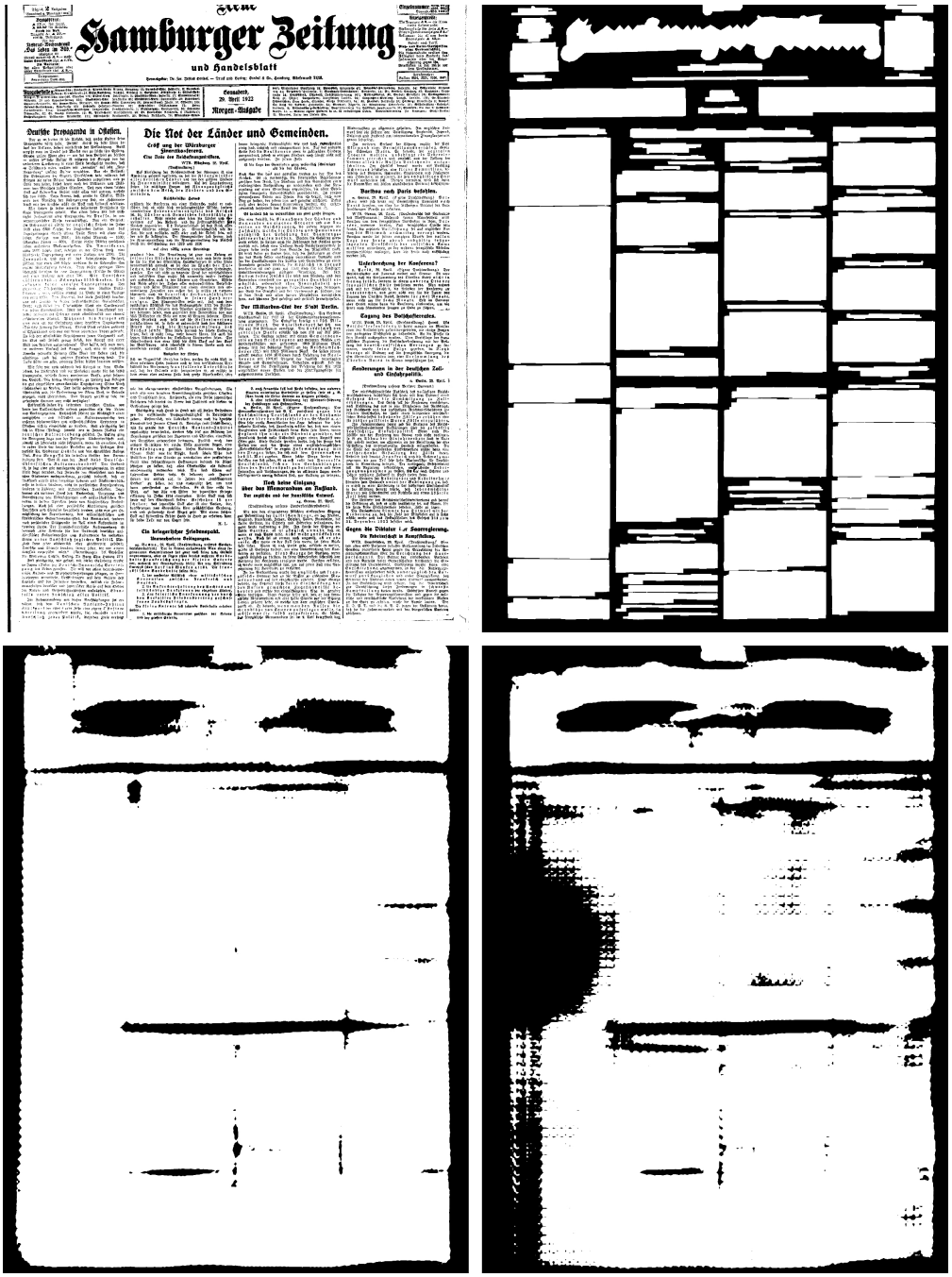
6.1.3 Upravený U-Net

Další experimenty byly prováděny se sítí upravený U-Net. Průběh trénování je na obrázku 6.7. Nejlepší ztráty na validačních datech pro datovou sadu Europeana bylo dosaženo v 9. epoše – $0,3978$, označme model jako uU_{eutext}. Od této epochy se validační ztráta horšila. Během dotrénování pak bylo dosaženo nejmenší ztráty hned ve 2. epoše – $0,3103$. Model označme uU_{pftext}. Po několika epochách se ztráta ustálila na hodnotách podobných z druhé epochy, nicméně nejlepší výsledek již nebyl překonán.

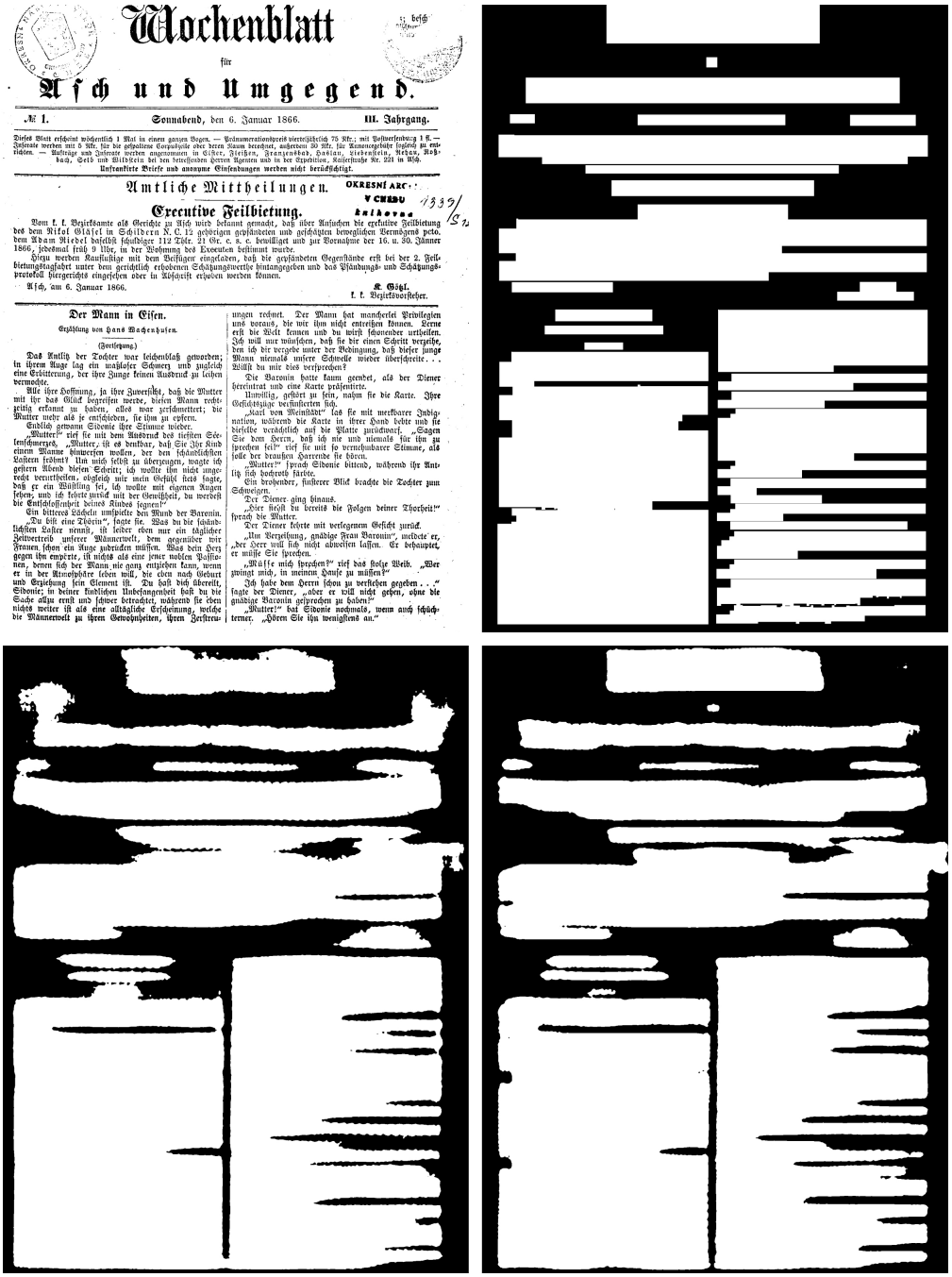
Na obrázcích 6.8 a 6.9 pak můžeme vidět ukázku výsledků segmentace. Na obrázku 6.8 vidíme, že síť pravděpodobně označila většinu textu. Zároveň je však označováno i okolí textu včetně oddělovačů. Dotrénovaný model navíc vede ještě ke zhoršení výsledku v místech, kde je text vybledlý. Dotrénování naopak vede ke zlepšení výsledků na obrázku 6.9. Označení textových bloků je srovnatelné s označením z U-Net, avšak síť U-Net lépe označuje odsazení první věty odstavce.



Obrázek 6.7: Průběh trénování síte upravený U-Net pro segmentaci textových bloků datové sadě Europeana můžeme vidět na horním obrázku. Následné dotrénování na datové sadě Portafontium je na obrázku spodním.



Obrázek 6.8: Segmentace textu sítí upravený U-Net. Ukázka výsledků segmentace na datové sadě Europeana. Obrázky: Originální obrázek, očekávaný výstup, výsledek modelu uU_{eutext}, výsledek modelu uU_{pftext}.



Wochenblatt

für

Asch und Umgegend.

№ 1. Sonnabend, den 6. Januar 1866. III. Jahrgang.

Dieses Blatt erscheint wöchentlich 1 Mal in einem ganzen Bogen. — Pränumerationspreis vierteljährlich 75 Kr.; mit Postversendung 1 fl. — Inserate werden mit 5 Kr. für die gespaltene Corpuszeile oder deren Raum berechnet, außerdem 30 Kr. für Remunerogebühr entrichtet. — Anfträge und Inserate werden angenommen in Eger, Franzensbad, Haslau, Liebenstein, Selb und Wildstein bei den betreffenden Herrn Agenten und in der Expedition, Kaiserstraße Nr. 221 in Asch.

Unfrankirte Briefe und anonyme Einsendungen werden nicht berücksichtigt.

okresní arc...
V CHEBU
Kulhová
1339/12

Amtliche Mittheilungen.

Executive Feilbietung.

Vom k. k. Bezirksamte als Gerichte zu Asch wird bekannt gemacht, daß über Ansuchen die executive Feilbietung des dem Nikol Gläsel in Schildern N. C. 12 gehörigen gepfändeten und gefchätzten beweglichen Vermögens peto. dem Adam Riedel hofelbst schuldiger 112 Thlr. 21 Sgr. c. s. c. bewilliget und zur Vornahme der 16. u. 30. Jänner 1866, jedesmal früh 9 Uhr, in der Wohnung des Executen bestimmt wurde.

Hiezu werden Kauflustige mit dem Beifügen eingeladen, daß die gepfändeten Gegenstände erst bei der 2. Feilbietungstagfahrt unter dem gerichtlich erhobenen Schätzungswerthe hintangegeben und das Pfändungs- und Schätzungsprotokoll hiergerichts eingesehen oder in Abschrift erhoben werden können.

Asch, am 6. Januar 1866.

K. Götzl.
k. k. Bezirksvorsteher.

Der Mann in Eisen.

Erzählung von Hans Wachenhusen.

(Fortsetzung.)

Das Antlitz der Tochter war leichenblaß geworden; in ihrem Auge lag ein wahnloser Schmerz und zugleich eine Erbitterung, der ihre Zunge keinen Ausdruck zu leihen vermochte.

Alle ihre Hoffnung, ja ihre Zuversicht, daß die Mutter mit ihr das Glück begreifen werde, diesen Mann recht-zeitig erkannt zu haben, alles war zerschmettert; die Mutter mehr als je entschieden, sie ihm zu opfern.

Endlich gewann Sidonie ihre Stimme wieder.

„Mutter!“ rief sie mit dem Ausdruck des tiefsten Seelenschmerzes, „Mutter, ist es denkbar, daß Sie Ihr Kind einem Manne hinwerfen wollen, der den schändlichsten Lastern fröhnt? Um mich selbst zu überzeugen, wagte ich gestern Abend diesen Schritt; ich wollte ihn nicht ungerecht verurtheilen, obgleich mir mein Gefühl stets sagte, daß er ein Wüstling sei, ich wollte mit eigenen Augen sehen, und ich kehrte zurück mit der Gewißheit, du werdest die Entschlossenheit deines Kindes segnen!“

Ein bitteres Lächeln umspielte den Mund der Baronin.

„Du bist eine Thörin“, sagte sie. Was du die schändlichsten Laster nennst, ist leider eben nur ein täglicher Zeitvertreib unserer Männerwelt, dem gegenüber wir Frauen schon ein Auge zudrücken müssen. Was dein Herz gegen ihn empörte, ist nichts als eine jener rohen Passionen, denen sich der Mann, wie ganz entziehen kann, wenn er in der Atmosphäre leben will, die eben nach Geburt und Erziehung sein Element ist. Du hast dich übereilt, Sidonie; in deiner kindlichen Unbefangenheit hast du die Sache allzu ernst und schwer betrachtet, während sie eben nichts weiter ist als eine alltägliche Erscheinung, welche die Männerwelt zu ihren Gewohnheiten, ihren Zerstreuungen rechnet. Der Mann hat mancherlei Privilegien uns voraus, die wir ihm nicht entreißen können. Lerne erst die Welt kennen und du wirst schonender urtheilen. Ich will nur wünschen, daß sie die einen Schritt vergeße, den ich dir vorgeschlagen unter der Bedingung, daß dieser junge Mann niemals unsere Schwelle wieder überschreite . . .

Wüßt du mir dies versprechen?

Die Baronin hatte kaum geendet, als der Diener hereintrat und eine Karte präsentirte.

Unwillig, gestört zu sein, nahm sie die Karte. Ihre Gesichtszüge verfinsterten sich.

„Karl von Meinlieder“ las sie mit merkbarer Indignation, während die Karte in ihrer Hand bebte und sie dieselbe verächtlich auf die Platte zurückwarf. „Sagen Sie dem Herrn, daß ich nie und niemals für ihn zu sprechen sei!“ rief sie mit so vernehmbarer Stimme, als solle der draußen Harrende sie hören.

„Mutter!“ sprach Sidonie bittend, während ihr Antlitz sich hochroth färbte.

Ein drohender, finsterer Blick brachte die Tochter zum Schweigen.

Der Diener ging hinaus.

„Hier siehst du bereits die Folgen deiner Thorheit!“ sprach die Mutter.

Der Diener kehrte mit verlegenem Gesicht zurück.

„Um Verzeihung, gnädige Frau Baronin“, meldete er, „der Herr will sich nicht abweisen lassen. Er behauptet, er müsse Sie sprechen.

„Müsse mich sprechen?“ rief das stolze Weib. „Wer zwingt mich, in meinem Hause zu müssen?“

Ich habe dem Herrn schon zu verstehen gegeben . . .“ sagte der Diener, „aber er will nicht gehen, ohne die gnädige Baronin gesprochen zu haben!“

„Mutter!“ bat Sidonie nochmals, wenn auch schüchterner. „Hören Sie ihn wenigstens an.“

Obrázek 6.9: Segmentace textu sítí upravený U-Net. Ukázka výsledků segmentace na datové sadě Portafontium. Obrázky: Originální obrázek, očekávaný výstup, výsledek modelu uU_{eutext}, výsledek modelu uU_{pftext}.

6.1.4 Shrnutí a výsledky

Model	Jaccard	F1-skóre	FgPA
ARU_{pftext}	0,687	0,814	0,765
ARU_{eutext}	0,614	0,760	0,718
U_{pftext}	**0,847**	**0,917**	0,867
U_{eutext}	0,768	0,866	0,741
uU_{pftext}	0,843	0,914	0,915
uU_{eutext}	0,820	0,901	**0,926**

Tabulka 6.1: Naměřené hodnoty pro validační data z datové sady Portafontium. Metriky: průměrná hodnota Jaccardova koeficientu, průměrná hodnota F1-Skóre a přesnost pixelu v popředí $FgPA$.

Tabulka 6.1 ukazuje hodnoty naměřených výsledků pro validační data Portafontium. Největší přesnosti pixelů v popředí $FgPA$ dosáhly modely sítě upraveného U-Netu a to konkrétně model, který nebyl dotrénovaný na datové sadě Portafontium (uU_{eutext}). Dotrénovaný model uU_{pftext} však dosahuje lepšího F1-skóre a Jaccardova koeficientu a zároveň přesnost pixelů v popředí není výrazně horší. Nejlepších hodnot Jaccardova koeficientu a F1-skóre bylo dosaženo modelem U_{pftext}, avšak přesnost pixelů v popředí není tak dobrá, jako u sítě upravený U-Net. Navíc se hodnoty F1-skóre a Jaccardova koeficientu mezi modely U_{pftext} a uU_{pftext} liší pouze minimálně a to i přes to, že model U_{pftext} lépe označuje odsazení první věty v odstavcích. Modely ARU-Net zaostávají ve všech sledovaných statistikách. Pokud se podíváme na obrázek 6.3, lze horší výsledky vysvětlit dírami v označených blocích a také tím, že větší nadpisy jsou označovány menší oblastí, než je označeno v `ground-truth`. Jako nejlepší model pro označení textových bloků je vybrán model uU_{pftext}, který je kompromisem mezi F1-skóre, Jaccardovým koeficientem a $FgPA$.

6.2 Segmentace oddělovačů

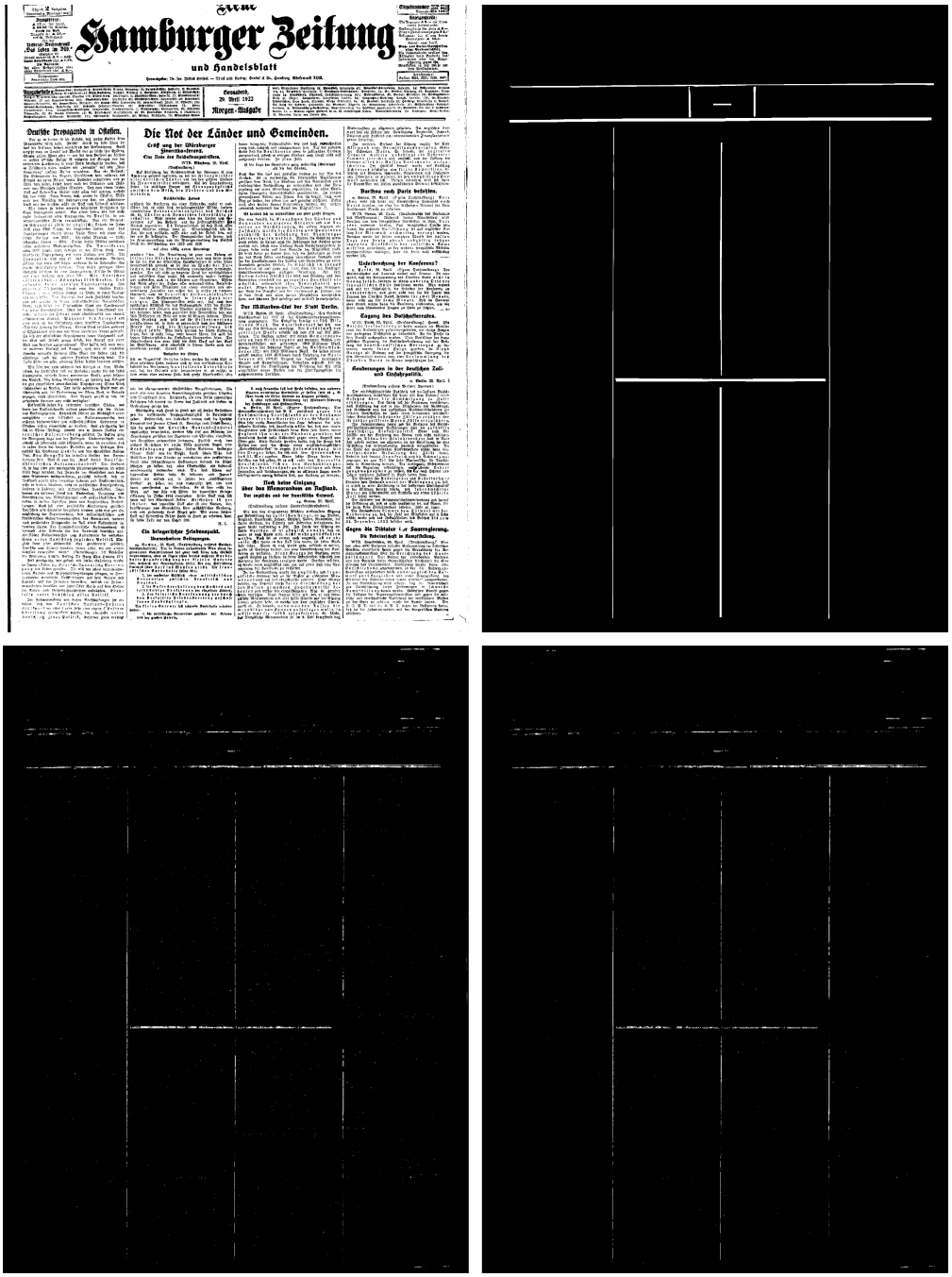
Datová sada pro trénování klasifikátoru segmentace textu je tvořena dvojcemi obrázků: obrázkem, který má být klasifikován a obrázkem s označenými, hledanými, oddělovači (`ground-truth`). Pixely tvořící oddělovače jsou reprezentovány bílou barvou, ostatní jsou černé. Odpovídajícím způsobem je tvořena i testovací sada.

6.2.1 ARU-Net

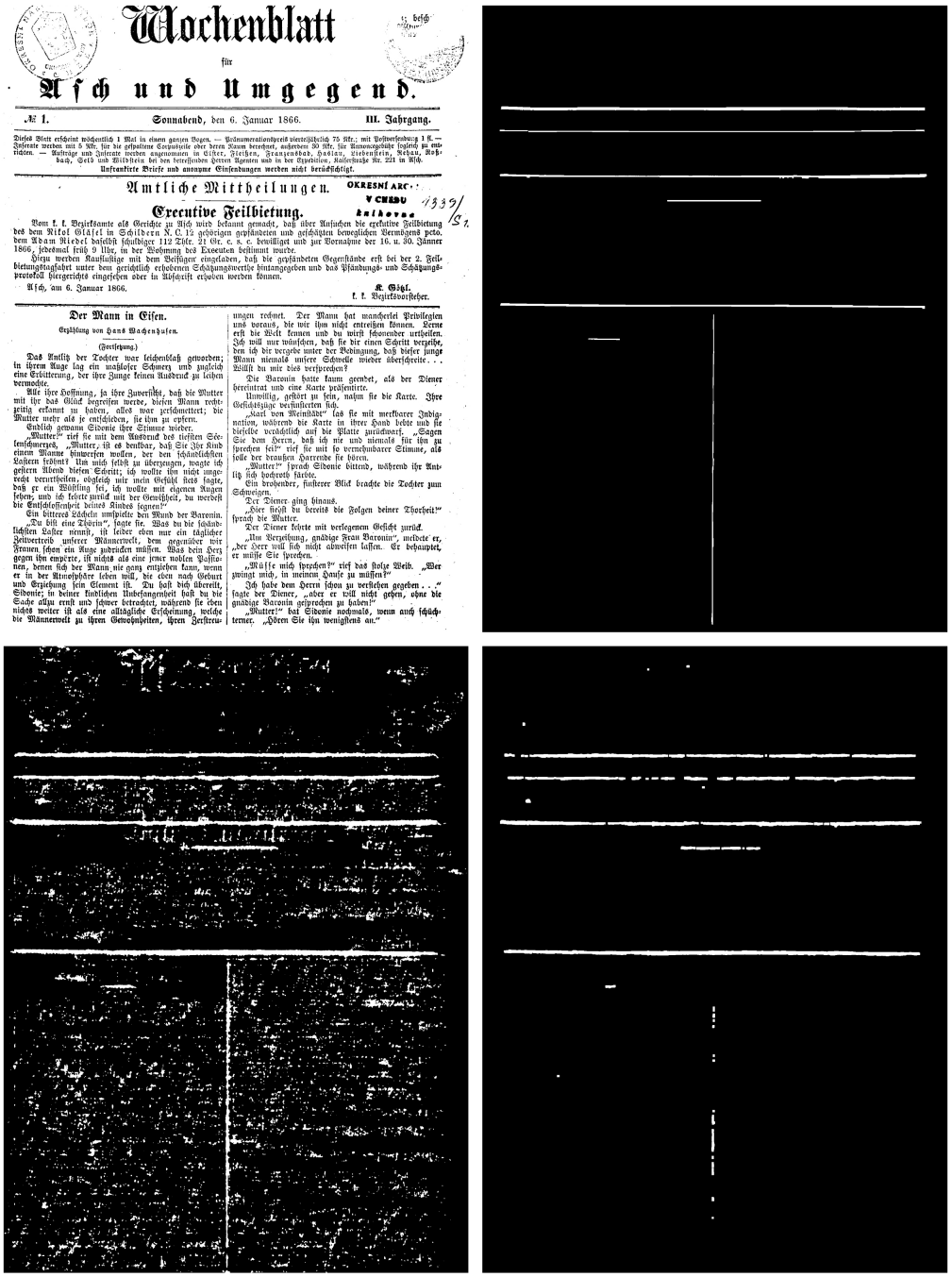
Na obrázku 6.10 můžeme vidět průběh trénování. Vzhledem k tomu, že se ztráta během trénování modelu na datové sadě Europena zmenšovala již minimálně, bylo trénování ukončeno po 100 epochách. Nejlepší ztráty bylo dosaženo v epoše 99. a to konkrétně s hodnotou $0,2857$. Označme tento model jako ARU_{eusep}. Následným dotrénováním tohoto modelu na datové sadě Portafontium bylo dosaženo nejlepší ztráty na validačních datech po dokončení 45. epochy s hodnotou ztráty $0,0099$. Od této epochy nedocházelo ke zlepšení validační ztráty a tak bylo trénování ukončeno po 100 epochách. Získaný model označme jako ARU_{pfsep}.

Na obrázcích 6.11 a 6.12 můžeme vidět ukázky výsledků segmentace oddělovačů. Na obrázku 6.11 si můžeme všimnout, že dotrénovaný model ARU_{pfsep} lehce zvýrazňuje hledané čáry a odstraňuje některé osamocené bílé pixely. Ani jeden z modelů nedokázal výrazněji označit nejsvrchnější horizontální čárů. Pod touto čarou modely také neoznačily dvě vertikální čáry. Dotrénování modelu naopak výrazně pomohlo v ukázce na datové sadě Portafontium, viz. obrázek 6.12. Model ARU_{eusep} sice lépe označuje hledané čáry, nicméně je také označeno velké množství pixelů, které oddělovače netvoří. Model ARU_{pfsep} je na tom v tomto ohledu mnohem lépe. Naopak však došlo k rozdělení některých oddělovačů, především vertikálního uprostřed obrázku.

Obrázek 6.10: Průběh trénování sítě ARU-Net pro segmentaci oddělovačů na datové sadě Europeana můžeme vidět na horním obrázku. Následné dotrénování na datové sadě Portafontium je na obrázku spodním.



Obrázek 6.11: ARU-Net segmentace oddělovačů. Ukázka výsledků segmentace na datové sadě Portafontium. Obrázky: Originální obrázek, očekávaný výstup, výsledek modelu ARU_{eusep}, výsledek modelu ARU_{pfsep}.



Wochenblatt
für
Asch und Umgegend.

№ 1. Sonnabend, den 6. Januar 1866. III. Jahrgang.

Amtliche Mittheilungen.

Executive Feilbietung.

Der Mann in Eisen.

Erzählung von Hans Wachenhusen.

(Fortsetzung.)

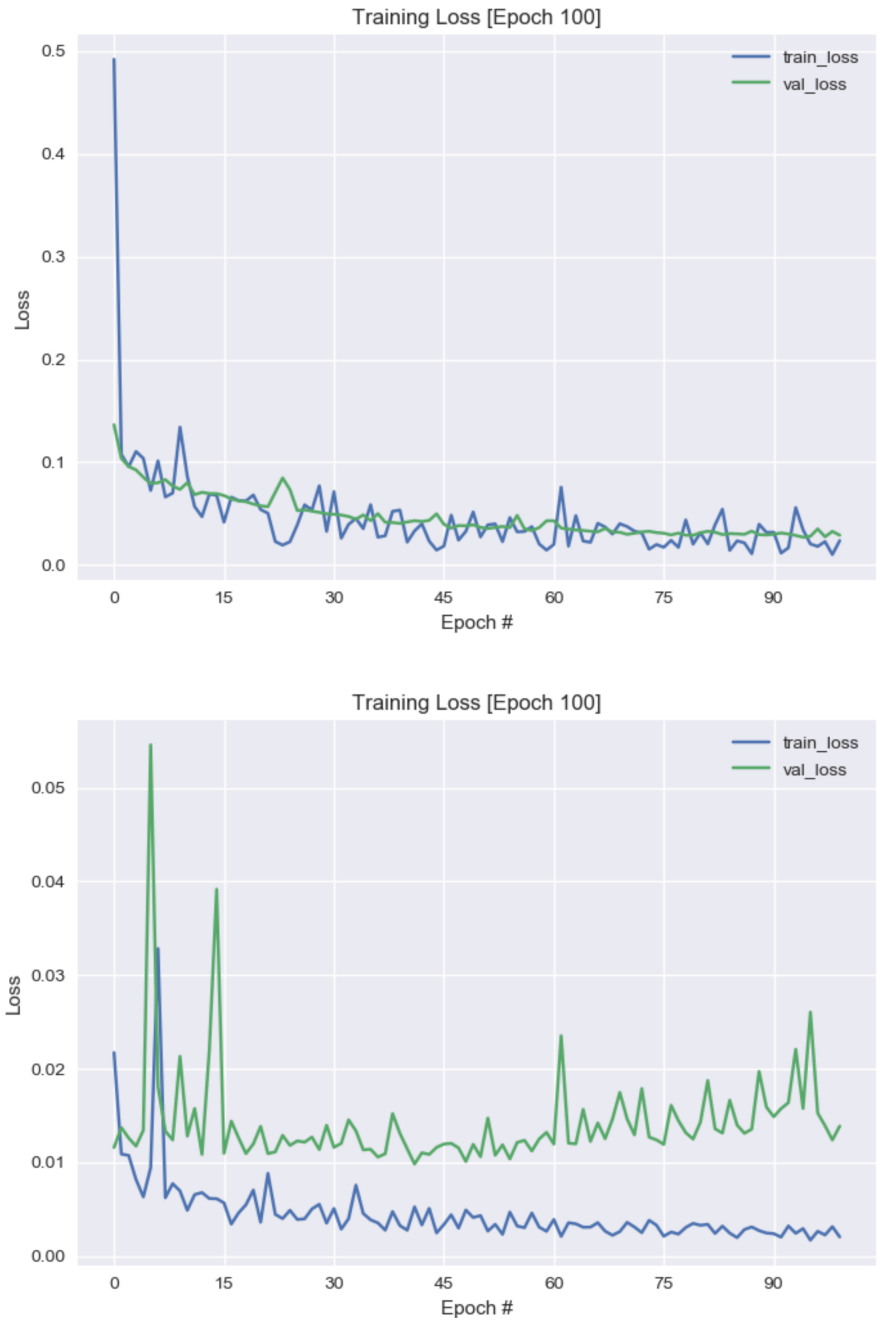
Obrázek 6.12: ARU-Net segmentace oddělovačů. Ukázka výsledků segmentace na datové sadě Europeana. Obrázky: Originální obrázek, očekávaný výstup, výsledek modelu ARU_{eusep}, výsledek modelu ARU_{pfsep}.

6.2.2 RU-Net

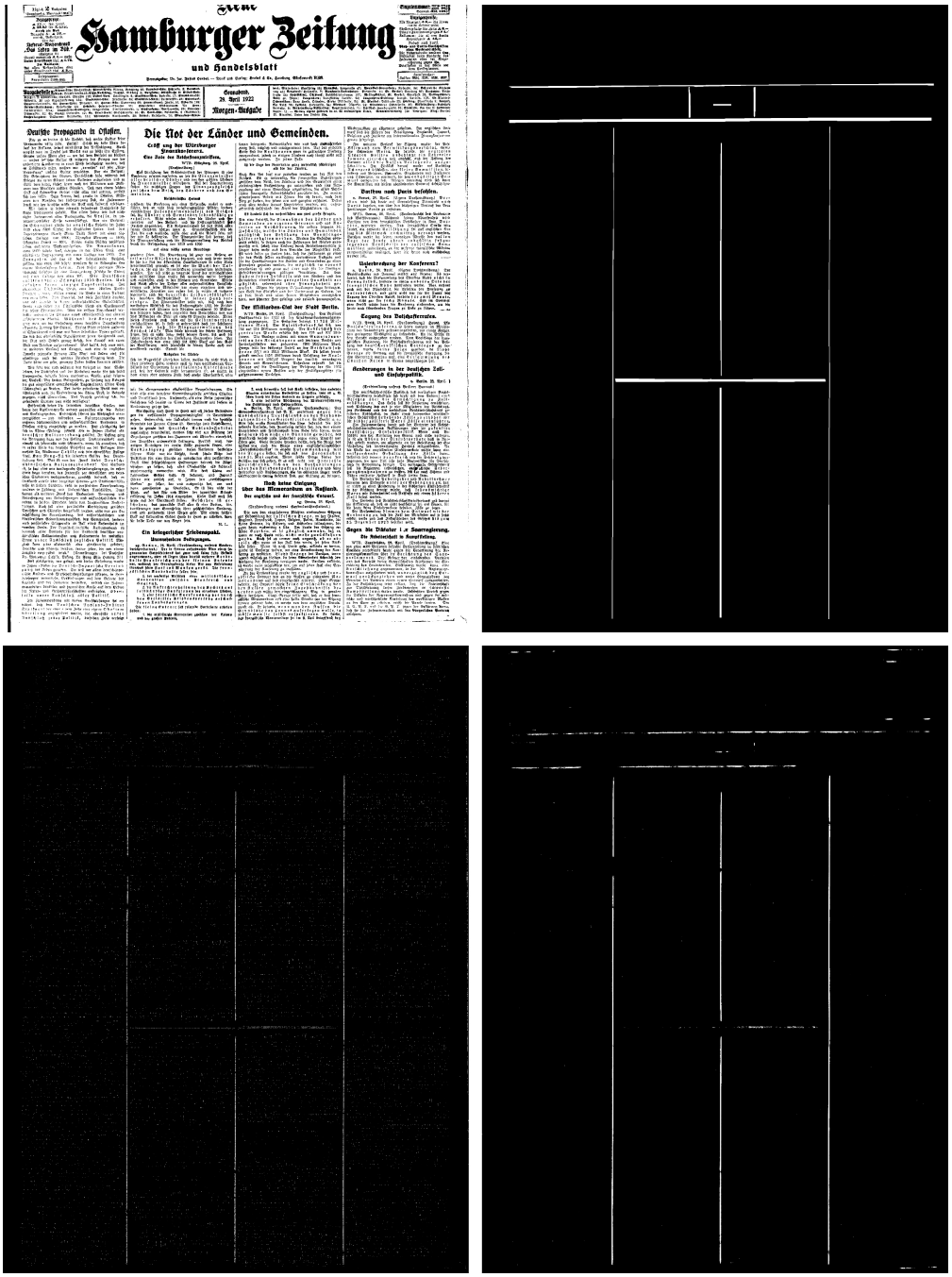
Síť RU-Net byla popsána v sekci 2.3.2. Jedná se o síť ARU-Net, která nepoužívá `attention`. Vzhledem k tomu, že nyní hledáme pouze oddělovače, které mají obvykle stejnou tloušťku, jeví se použití `attention` jako neopodstatněné.

Obrázek 6.13 ukazuje průběh trénování. Nejlepší ztráty na validačních datech pro datovou sadu Europeana bylo dosaženo v 95. epoše – $0,0270$. Vytvořený model označme jako RU_{eusep}. Validační chyba s příbývajícími epochami již výrazně neklesala a tak bylo trénování ukončeno po 100 epoše. Dotrénování na datové sade Portafontium dosáhlo nejlepší ztáty $0,009$ na validačních datech po dokončení 42. epochy. S dalšími epochami dochází ke zhoršení validační chyby, což může signalizovat přetrénování modelu.

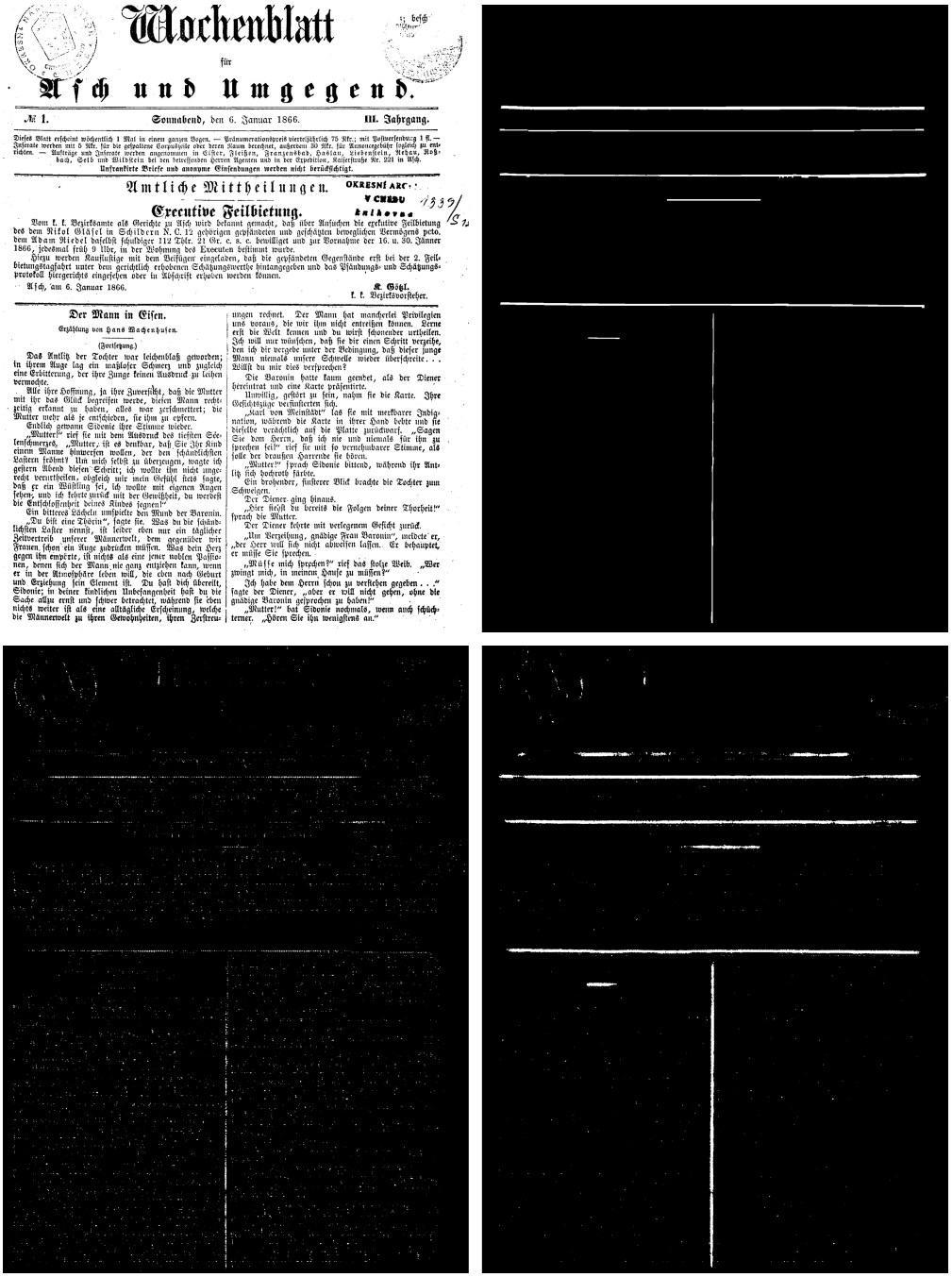
Na obrázcích 6.14 a 6.15 můžeme vidět ukázky výsledků segmentace oddělovačů. Všimněme si, že v obou případech vede dotrénovaní k zvýraznění hledaných čar a částečnému odstranění šumu. Oproti modelům ARU-Net jsou výrazněji a kompaktněji označovány vertikální oddělovače a naopak dochází ke zhoršení označování horizontálních oddělovačů.



Obrázek 6.13: Průběh trénování sítě RU-Net pro segmentaci oddělovačů na datové sadě Europeana můžeme vidět na horním obrázku. Následné dotrénování na datové sadě Portafontium je na obrázku spodním.



Obrázek 6.14: RU-Net segmentace oddělovačů. Ukázka výsledků segmentace na datové sadě Europeana. Obrázky: Originální obrázek, očekávaný výstup, výsledek modelu RU_{eusep}, výsledek modelu RU_{pfsep}.



Wochenblatt

für

Asch und Umgegend.

№ 1. Sonnabend, den 6. Januar 1866. III. Jahrgang.

Dieses Blatt erscheint wöchentlich 1 Mal in einem ganzen Bogen. — Pränumerationspreis vierteljährlich 75 Kr.; mit Postversendung 1 fl. — Inserate werden mit 5 Kr. für die gespaltene Corpuszeile oder deren Raum berechnet, äußerdem 30 Kr. für Remunerationsgebühr entrichtet. — Aufträge und Inserate werden angenommen in Eger, Tirschen, Franzensbad, Haslau, Liebenstein, Neuhaus, Asch, Graslitz, Selb und Wildstein bei den betreffenden Herrn Agenten und in der Expedition, Kaiserstraße Nr. 221 in Asch.

Unfrankirte Briefe und anonyme Einsendungen werden nicht berücksichtigt.

Amtliche Mittheilungen.

OKRESNÍ ARC··
V CHEBU
Kulhovan
1339/
52

Executive Feilbietung.

Vom k. k. Bezirksamte als Gerichte zu Asch wird bekannt gemacht, daß über Ansuchen die exekutive Feilbietung des dem Nikol Gläsel in Schildern N. C. 12 gehörigen gepfändeten und gesetzlichen beweglichen Vermögens peto. dem Adam Riedel hiesselbst schuldiger 112 Thlr. 21 Sgr. c. s. c. bewilliget und zur Vornahme der 16. u. 30. Jänner 1866, jedesmal früh 9 Uhr, in der Wohnung des Executen bestimmt wurde.

Hiezu werden Kauflustige mit dem Beifügen eingeladen, daß die gepfändeten Gegenstände erst bei der 2. Feilbietungstagfahrt unter dem gerichtlich erhobenen Schätzungswerthe hintangegeben und das Pfändungs- und Schätzungsprotokoll hiergerichts eingesehen oder in Abschrift erhoben werden können.

Asch, am 6. Januar 1866.

K. Götzl.
k. k. Bezirksvorsteher.

Der Mann in Eisen.

Erzählung von Hans Wachenhusen.

(Fortsetzung.)

Das Antlitz der Tochter war leichenblaß geworden; in ihrem Auge lag ein wahnsloser Schmerz und zugleich eine Erbitterung, der ihre Zunge keinen Ausdruck zu leihen vermochte.

Alle ihre Hoffnung, ja ihre Zuversicht, daß die Mutter mit ihr das Glück begreifen werde, diesen Mann rechtzeitig erkannt zu haben, alles war zerschmettert; die Mutter mehr als je entschieden, sie ihm zu opfern.

Endlich gewann Sidonie ihre Stimme wieder.

„Mutter!“ rief sie mit dem Ausdruck des tiefsten Seelenschmerzes, „Mutter, ist es denkbar, daß Sie Ihr Kind einem Manne hinwerfen wollen, der den schändlichsten Lastern fröhnt? Um mich selbst zu überzeugen, wagte ich gestern Abend diesen Schritt; ich wollte ihn nicht ungerecht verurtheilen, obgleich mir mein Gefühl stets sagte, daß er ein Wüstling sei, ich wollte mit eigenen Augen sehen, und ich kehrte zurück mit der Gewißheit, du werdest die Entschlossenheit deines Kindes segnen!“

Ein bitteres Lächeln umspielte den Mund der Baronin.

„Du bist eine Thörin“, sagte sie. Was du die schändlichsten Laster nennst, ist leider eben nur ein täglicher Zeitvertreib unserer Männerwelt, dem gegenüber wir Frauen schon ein Auge zudrücken müssen. Was dein Herz gegen ihn empörte, ist nichts als eine jener rohen Passionen, denen sich der Mann, wie ganz entziehen kann, wenn er in der Atmosphäre leben will, die eben nach Geburt und Erziehung sein Element ist. Du hast dich übereilt, Sidonie; in deiner kindlichen Unbefangenheit hast du die Sache allzu ernst und schwer betrachtet, während sie eben nichts weiter ist als eine alltägliche Erscheinung, welche die Männerwelt zu ihren Gewohnheiten, ihren Zerstreuungen rechnet. Der Mann hat mancherlei Privilegien uns voraus, die wir ihm nicht entreißen können. Lerne erst die Welt kennen und du wirst schonender urtheilen. Ich will nur wünschen, daß sie die einen Schritt vergeße, den ich dir vergebe unter der Bedingung, daß dieser junge Mann niemals unsere Schwelle wieder überschreite. . . Willst du mir dies versprechen?

Die Baronin hatte kaum geendet, als der Diener hereintrat und eine Karte präsentirte.

Unwillig, gestört zu sein, nahm sie die Karte. Ihre Gesichtszüge verfinsterten sich.

„Karl von Meinlieder“ las sie mit merkbarer Indignation, während die Karte in ihrer Hand bebte und sie dieselbe verächtlich auf die Platte zurückwarf. „Sagen Sie dem Herrn, daß ich nie und niemals für ihn zu sprechen sei!“ rief sie mit so vernehmbarer Stimme, als solle der draußen Harrende sie hören.

„Mutter!“ sprach Sidonie bittend, während ihr Antlitz sich hochroth färbte.

Ein drohender, finsterer Blick brachte die Tochter zum Schweigen.

Der Diener ging hinaus.

„Hier siehst du bereits die Folgen deiner Thorheit!“ sprach die Mutter.

Der Diener kehrte mit verlegenem Gesicht zurück.

„Um Verzeihung, gnädige Frau Baronin“, meldete er, „der Herr will sich nicht abweisen lassen. Er behauptet, er müsse Sie sprechen.

„Müsse mich sprechen?“ rief das stolze Weib. „Wer zwingt mich, in meinem Hause zu müssen?“

Ich habe dem Herrn schon zu verstehen gegeben . . .“ sagte der Diener, „aber er will nicht gehen, ohne die gnädige Baronin gesprochen zu haben!“

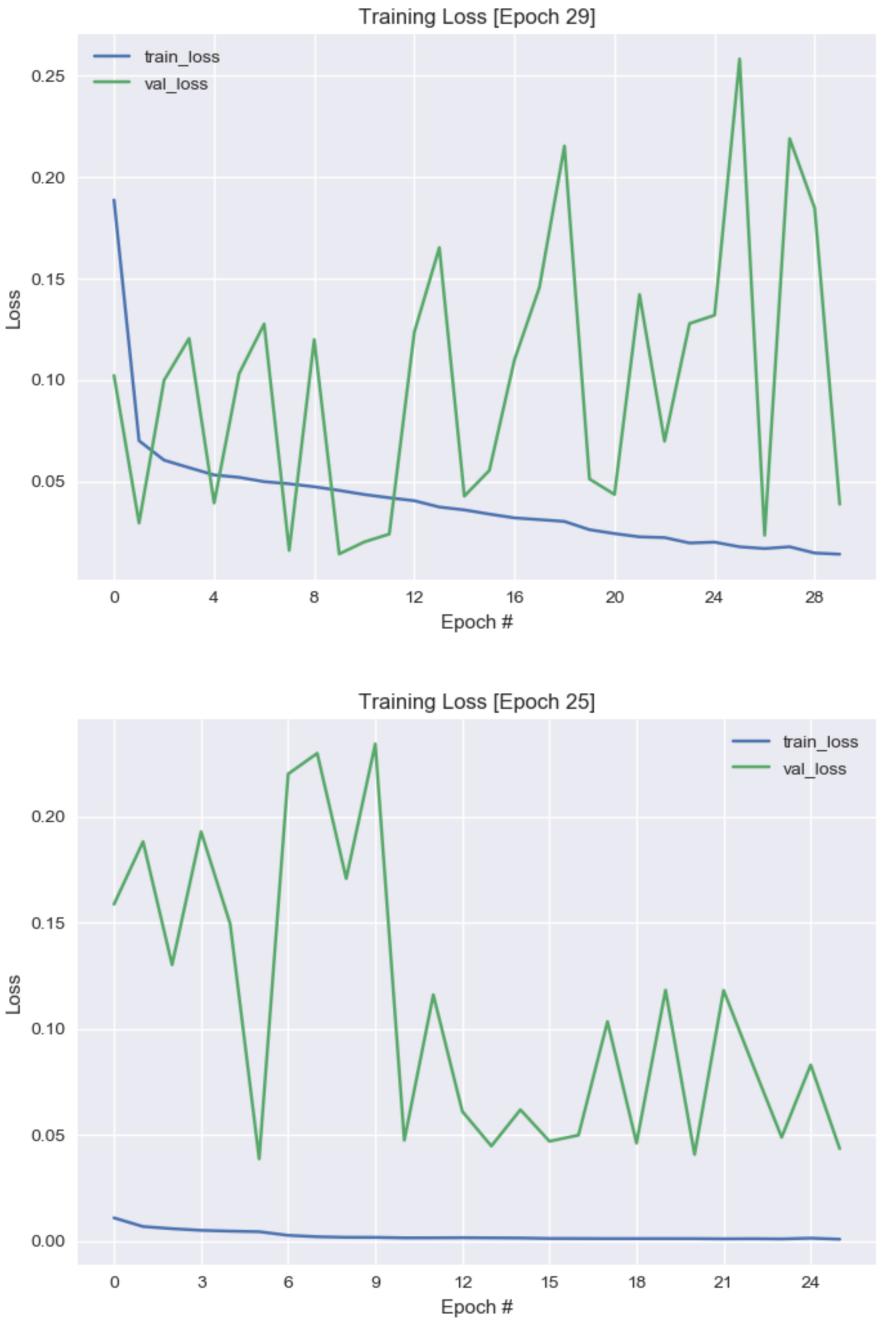
„Mutter!“ bat Sidonie nochmals, wenn auch schüchterner. „Hören Sie ihn wenigstens an.“

Obrázek 6.15: RU-Net segmentace oddělovačů. Ukázka výsledků segmentace na datové sadě Portafontium. Obrázky: Originální obrázek, očekávaný výstup, výsledek modelu RU_{eusep}, výsledek modelu RU_{pfsep}.

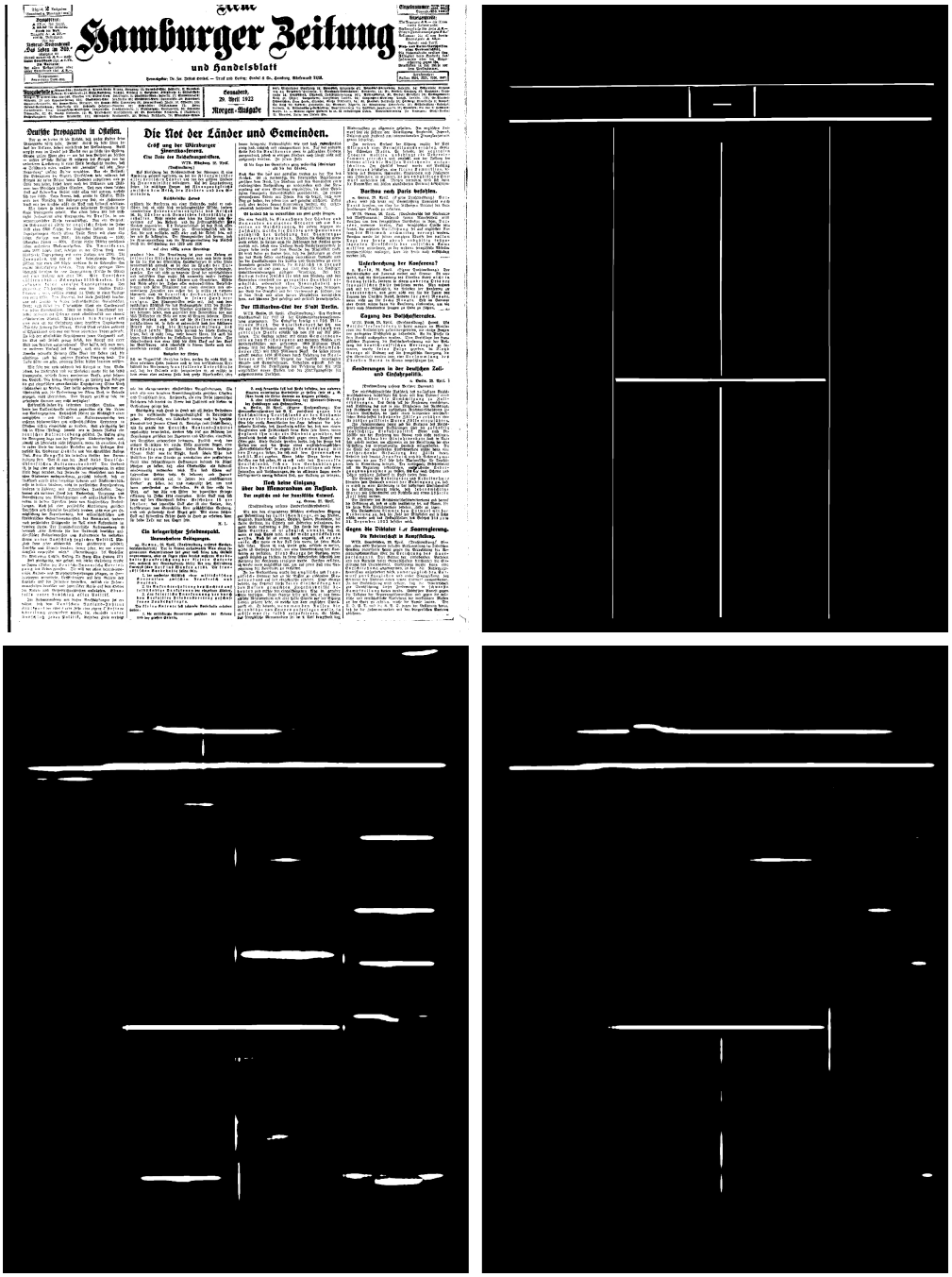
6.2.3 U-Net

Síť U-Net byla natrénována nejprve na datové sadě Europeana, kdy bylo dosaženo nejlepší ztráty $0,014$ na validačních datech po dokončení 9. epochy. Získaný model označme jako U_{eusep}. Dotrénovaný model U_{pfsep} na datové sadě Portafontium dosáhl nejlepší validační ztráty $0,039$ v 5. epoše. Průběh trénování můžeme vidět na obrázku 6.16.

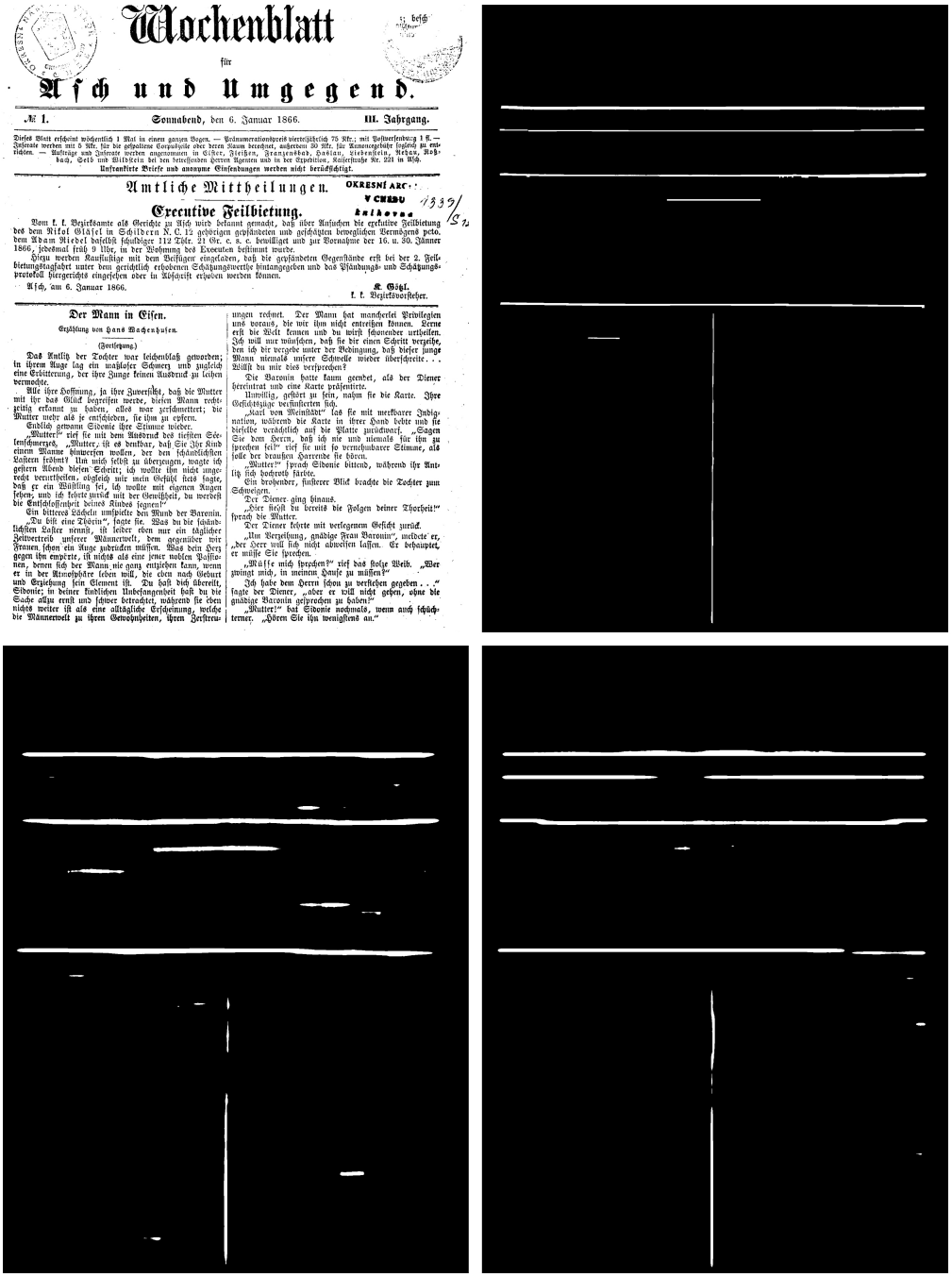
Ukázky výsledných segmentací oddělovačů můžeme vidět na obrázcích 6.17 a 6.18. Na obrázku 6.17 si můžeme všimnout, že mají modely problémy s označením vertikálních čar oddělující sloupce textů. Kromě toho jsou v textu označeny i některé nadpisy psanné tučným písmem. Dotrénovaný model U_{pfsep} se naučil některé z těchto nadpisů neoznačovat, nicméně můžeme vidět, že některé z nich zůstávají stále označené. Lepší označení můžeme vidět na ukázce výsledků z datové sady Portafontium, viz. obrázek 6.18. Můžeme si všimnout, že dotrénovaný model U_{pfsep} lépe označil vertikální čáru mezi sloupci textu. Výsledky modelů se také liší v označení horizontálních čar. Horizontální čára, která je druhá ze shora je označena modelem U_{pfsep}, kdežto model U_{eusep} tuto čaru neoznačuje. Naopak čtvrtá vertikální čára shora je lépe označena modelem U_{eusep}. Model U_{eusep} však kromě čar označuje i některý text. Můžeme si také všimnout, že horizontální oddělovač v levém sloupci s textem správně neoznačil ani jeden z modelů.



Obrázek 6.16: Průběh trénování sítě U-Net pro segmentaci oddělovačů na datové sadě Europeana můžeme vidět na horním obrázku. Následné dotrénování na datové sadě Portafontium je na obrázku spodním.



Obrázek 6.17: U-Net segmentace oddělovačů. Ukázka výsledků segmentace na datové sadě Europeana. Obrázky: Originální obrázek, očekávaný výstup, výsledek modelu U_{eusep}, výsledek modelu U_{pfsep}.



Wochenblatt
für
Asch und Umgegend.

№ 1. Sonnabend, den 6. Januar 1866. III. Jahrgang.

Amtliche Mittheilungen.

Executive Feilbietung.

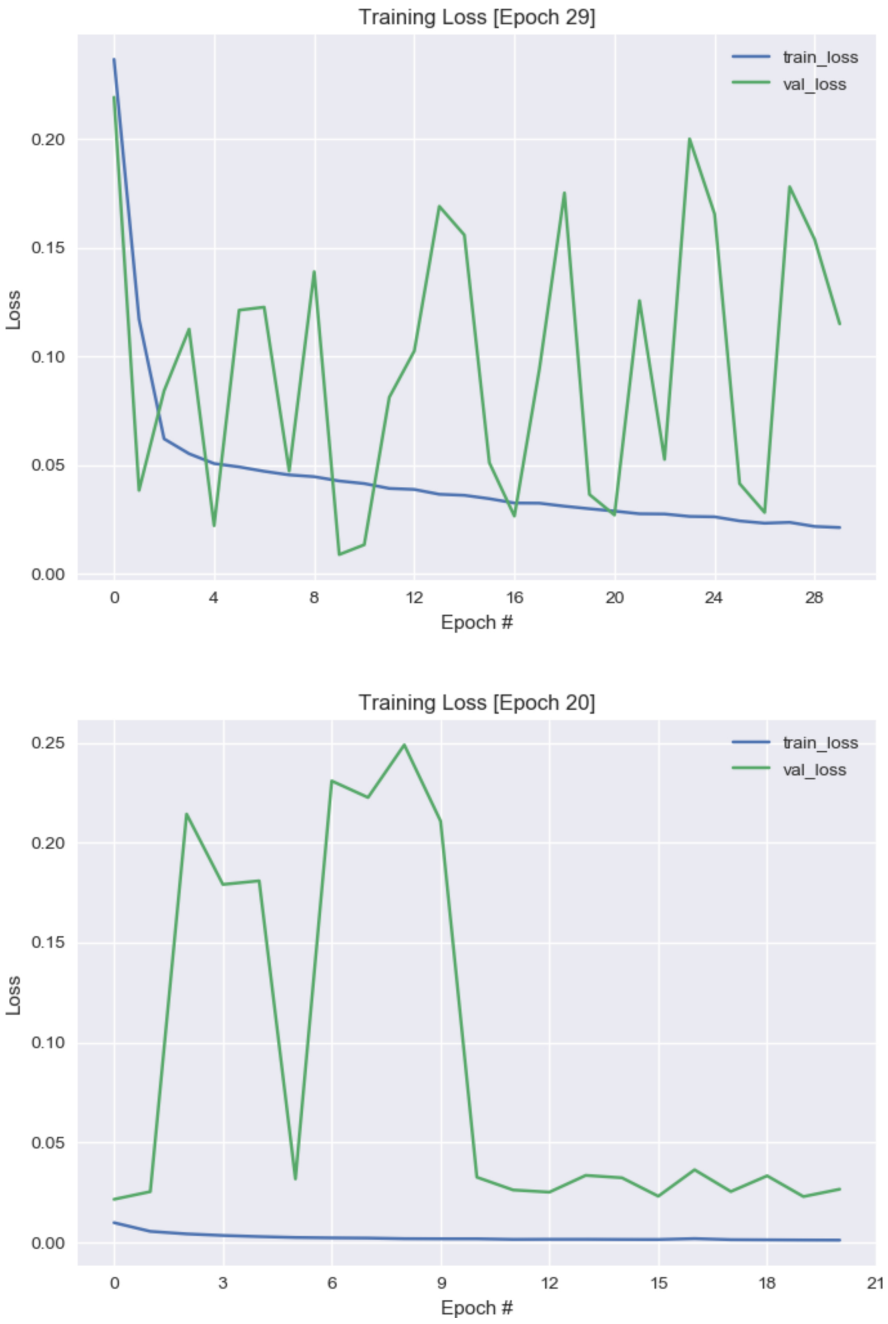
Der Mann in Eisen.
Erzählung von Hans Wachenhusen.
(Fortsetzung.)

Obrázek 6.18: U-Net segmentace oddělovačů. Ukázka výsledků segmentace na datové sadě Portafontium. Obrázky: Originální obrázek, očekávaný výstup, výsledek modelu U_{eusep}, výsledek modelu U_{pfsep}.

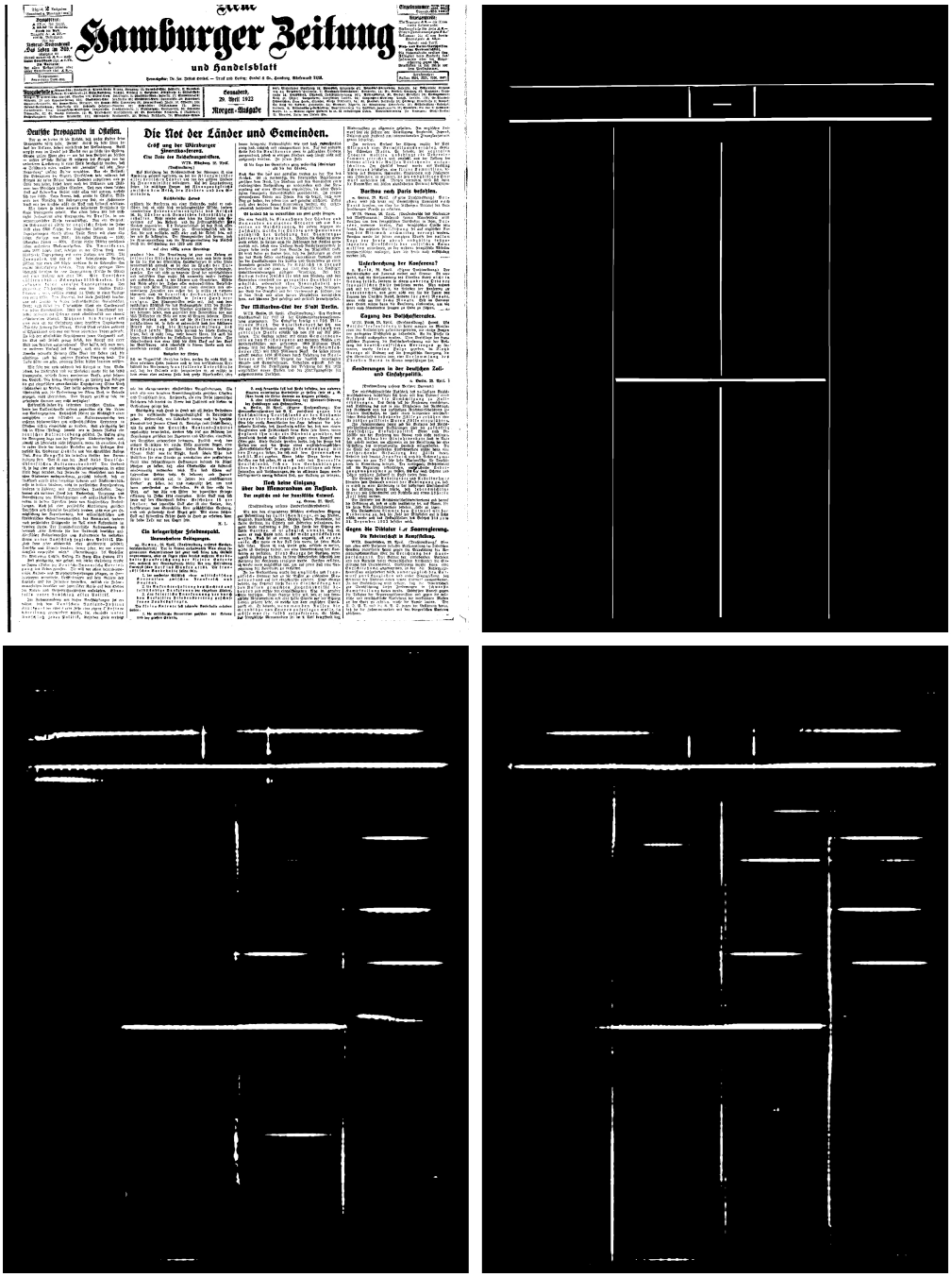
6.2.4 Upravený U-Net

Během trénování modelu uU_{eusep} na datové sadě Europeana bylo dosaženo nejlepší ztráty na validačních datech $0,0083$ v 9. epoše. Od této epochy se validační ztráta horšila a tak bylo trénování ukončeno po 27 epochách. Během dotrénování na datové sadě Portafontium dosáhla validační ztráta nejmenší hodnoty $0,0215$ hned po první epoše. Označme tento model jako uU_{pfsep}. V pozdějších epochách se ztráta tomutu výsledku přibližovala, nícméne již nebyla překonána a trénování bylo ukončení 21. epochou. Průběh trénování můžeme vidět na obrázku 6.19.

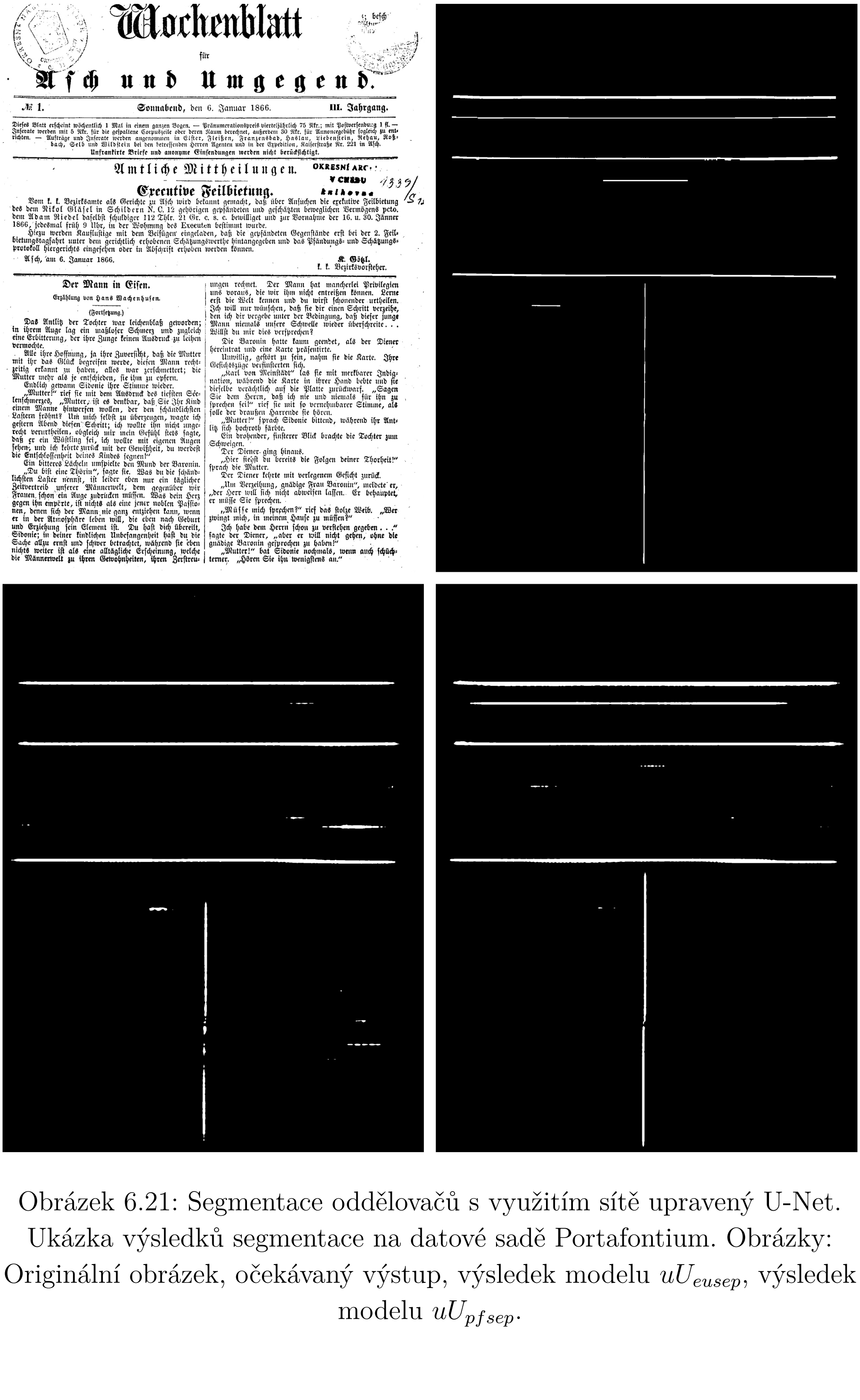
Ukázku dosažených výsledků můžeme vidět na obrázcích 6.20 a 6.21. Oproti modelům U-Net si můžeme všimnout, že jsou v obrázku 6.20 lépe označovány vertikální separátory mezi sloupci textu, nicméně stále nejsou označeny tak dobře jako modely ARU-Net a RU-Net. Upravený U-Net také jako U-Net v některých případech označuje některé tučně psané napdisy jako oddělovače. Modely uU_{eusep} a uU_{pfsep} jako jediné ze všech trénovaných modelů označují oddělovače pod první horizontální čarou v obrázku 6.20. Co se týče ukázky na datové sadě Portafontium, můžeme si všimnout, že dotrénovaný model uU_{pfsep} oproti modelu uU_{eusep} označil v pořadí druhou horizontální čáru. Zároveň došlo i k odstranění označení některých částí textu. Horizontální oddělovač čtvrtý od shora byl označen pouze minimálně a to pouze modelem uU_{pfsep}. Můžeme si také všimnout, že horizontální oddělovač v levém sloupci s textem správně neoznačil ani jeden z modelů, stejně jako tomu bylo u modelů U-Net.



Obrázek 6.19: Průběh trénování sítě upravený U-Net pro segmentaci oddělovačů na datové sadě Europeana můžeme vidět na horním obrázku. Následné dotrénování na datové sadě Portafontium je na obrázku spodním.



Obrázek 6.20: Segmentace oddělovačů s využitím sítě upravený U-Net. Ukázka výsledků segmentace na datové sadě Europeana. Obrázky: Originální obrázek, očekávaný výstup, výsledek modelu uU_{eusep}, výsledek modelu uU_{pfsep}.



Obrázek 6.21: Segmentace oddělovačů s využitím sítě upravený U-Net. Ukázka výsledků segmentace na datové sadě Portafontium. Obrázky: Originální obrázek, očekávaný výstup, výsledek modelu uU_{eusep}, výsledek modelu uU_{pfsep}.

6.2.5 Shrnutí a výsledky

Model	Jaccard	F1-skóre	FgPA
ARU_{pftsep}	**0,750**	**0,833**	0,752
ARU_{eusep}	0,564	0,639	**0,894**
RU_{pfsep}	0,668	0,756	0,627
RU_{eusep}	0,524	0,589	0,063
U_{pfsep}	0,663	0,748	0,568
U_{eusep}	0,621	0,701	0,589
uU_{pfsep}	0,678	0,762	0,689
uU_{eusep}	0,626	0,702	0,552

Tabulka 6.2: Naměřené hodnoty pro validační data z datové sady Portafontium. Metriky: průměrná hodnota Jaccardova koeficientu, průměrná hodnota F1-Skóre a přesnost pixelu v popředí $FgPA$.

Tabulka 6.2 ukazuje hodnoty naměřených výsledků pro validační data z datové sady Portafontium. Nejlepší přesnosti pixelů v popředí ($FgPA$) dosahuje model ARU_{eusep}, který ale zároveň dosahuje nejhoršího F1-skóre a Jaccardova koeficientu. Pro dotrénovaný model ARU_{eusep} bylo dosaženo druhého nejlepšího výsledeku $FgPA$ a zároveň bylo dosaženo nejlepších hodnot F1-skóre a Jaccardova koeficientu. Tento kontrast můžeme vidět na obrázku 6.12, kde je vidět, že dotrénováním došlo k odstranění šumu v obrázku, zároveň však označené čáry nejsou tolik výrazné. Přestože vizuálně druhých nejpoužitelnějších výsledků dosahuje model RU_{pfsep}, naměřené statistiky jsou významně horší než pro modely ARU-Net. Lepších statistických výsledků není dosahováno ani oproti modelu uU_{pfsep}, přestože model RU_{pfsep} lépe označuje krátké oddělovače uprostřed textu (viz. obrázky 6.15 a 6.21). Horší výsledky lze vysvětlit tím, že správná detekce krátkého oddělovače uprostřed textu je pro statistiku méně významná, než správné označení oddělovače vedoucího přes celou stránku. Výsledky modelu RU_{pfsep} jsou také ovlivněny tím, že se ve výsledných segmentacích vyskytuje lehký šum. Vzhledem k tomu, že se jedná většinou o osamocené pixely, je možné dalšími operacemi tento šum odstranit a tím tak docílit lepších statistických výsledků. Jak bylo zmíněno výše, modely sítí U-Net a upravený U-Net označují jako oddělovače některé nadpisy v textu, což může být daleko větší chybou pro následné rozdělování stránky, než nevýrazné označení skutečného oddělovače. Z tohoto důvodu se jeví jako lepší model RU_{pfsep}. Z vizuálních i statistických výsledků nejlépe vychazí modely ARU-Net. Přestože model ARU_{eusep} dosahuje nejlepší

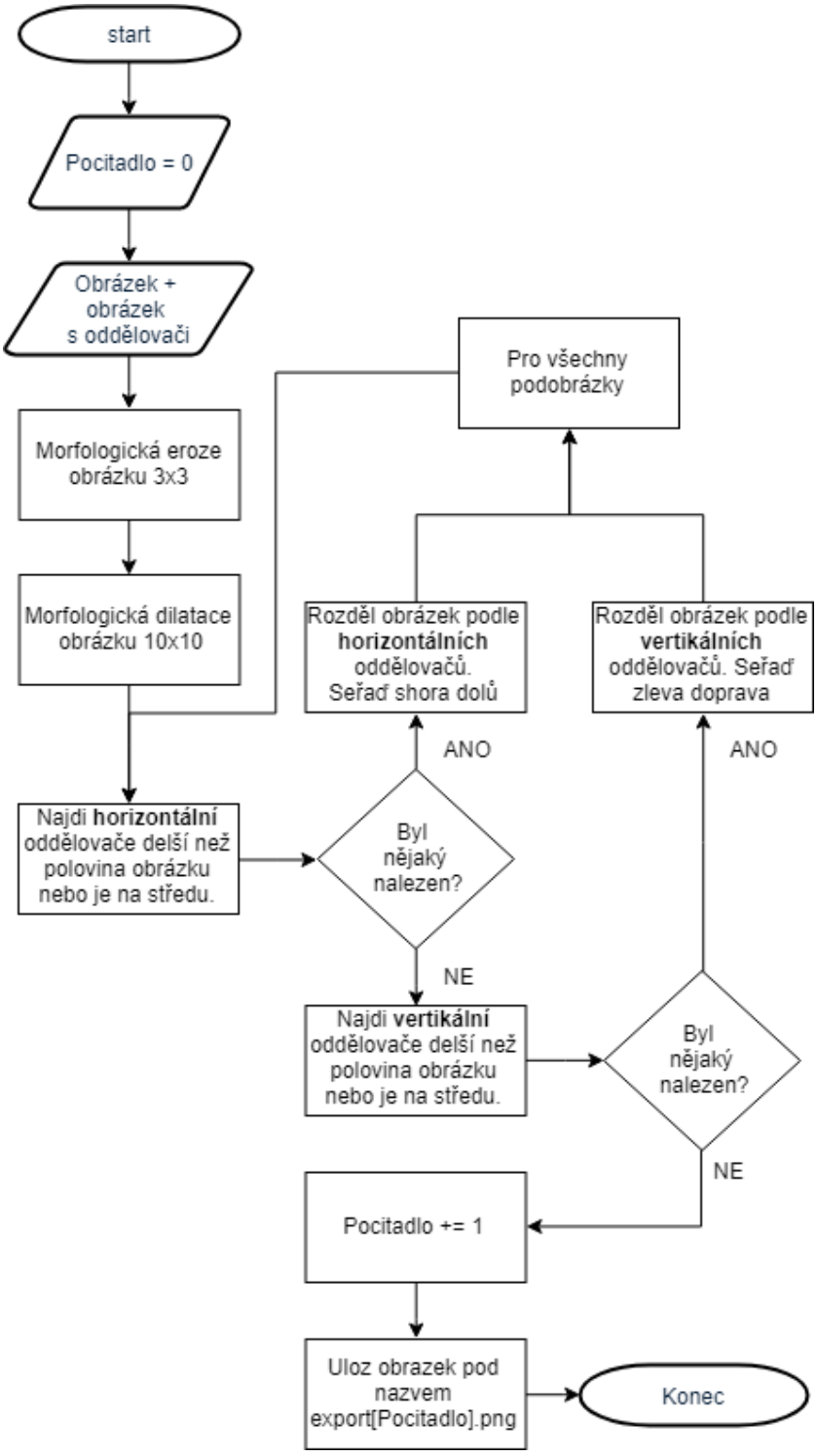
$FgPA$, je pro detekci oddělovačů jako nejlepší model zvolen ARU_{pftsep}, v jehož výsledcích je podstatně méně šumu, jehož odstranění není triviální.

6.2.6 Použití oddělovačů

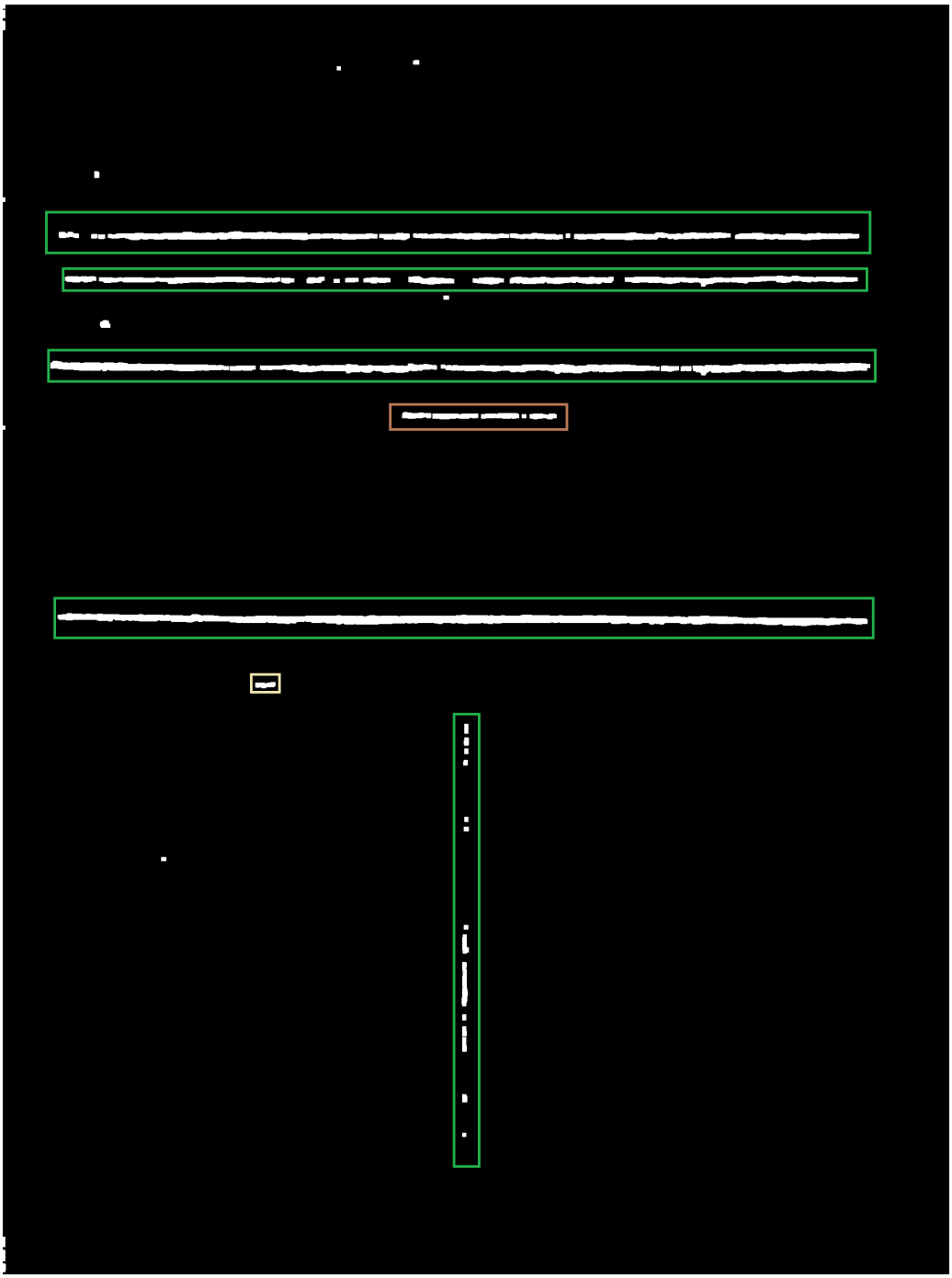
Jakmile máme ve stránce nalezené oddělovače, můžeme stránku podle nich rozdělit. Na obrázku 6.22 můžeme vidět vývojový diagram celého algoritmu. Označme vstupní obrázek jako I_{orig} a obrázek s označenými oddělovači jako I_{sep} (černá barva reprezentuje pozadí, bílá barva reprezentuje pixely oddělovače). Předpokládá se, že jsou oba obrázky binarizované. Tyto obrázky jsou vstupem algoritmu. Nejprve se na obrázek I_{sep} aplikuje morfologická eroze s maticí o velikost 5×5. Díky tomu jsou z obrázku odstraněny především osamocené body, které pravděpodobně nebudou tvořit horizontální, nebo vertikální čáru. V případě použití matice vyšších rozměrů docházelo již i k odstraňování pixelů tvořících čáru. Dále se na obrázek I_{sep} aplikuje morfologická dilatace. Velikost operátoru je 10×10. Bylo vypozorováno, že oddělovače v `ground-truth` nejsou moc široké a navíc jsou čáry v původním obrázku často přerušované. Rozměry operátorů byly voleny na základě experimentů. Dilataci se tak v obrázku I_{sep} docílí zvýraznění předpovězené čáry a navíc je docíleno toho, že jsou některé přerušované čáry spojeny v jednu.

V obrázku I_{sep} jsou následně hledány horizontální čáry. Jedním způsobem pro hledání horizontálních čar je určení horizontálního profilu. Pro každý řádek obrázku se vytvoří histogram, kde hodnoty jednotlivých sloupců představují počet bílých pixelů. Pokud si tento histogram představíme jako graf, kde x představuje číslo řádku a y počet bílých pixelů, hledané horizontální čáry budou ve vrcholech tohoto grafu.

Dalším způsobem je využití Houghovy transformace pro detekci čar [13]. Zjednodušeně, tento algoritmus prochází postupně všechny pixely v obrázku. Pokud je pixel v našem případě černý, je přeskočen. Pokud je bílý, je skrz tento bod proloženo několik přímek pod různým úhlem. Pro každou přímku se zkoumá, kolik bílých pixelů na této přímce leží. Na základě takto vytvořené statistiky jsou získání kandidáti pro čáry. Knihovna `OpenCV` má tuto funkci implementovanou. Mimo jiné je možné si zde parametry nastavit, jaká je minimální délka čáry, či jaká je maximální mezera mezi čárami, aby byla považována za jednu. Nalezené čáry Houghovou transofrmací je nutné ještě vyfiltrovat podle jejich orientace (horizontální, vertikální).

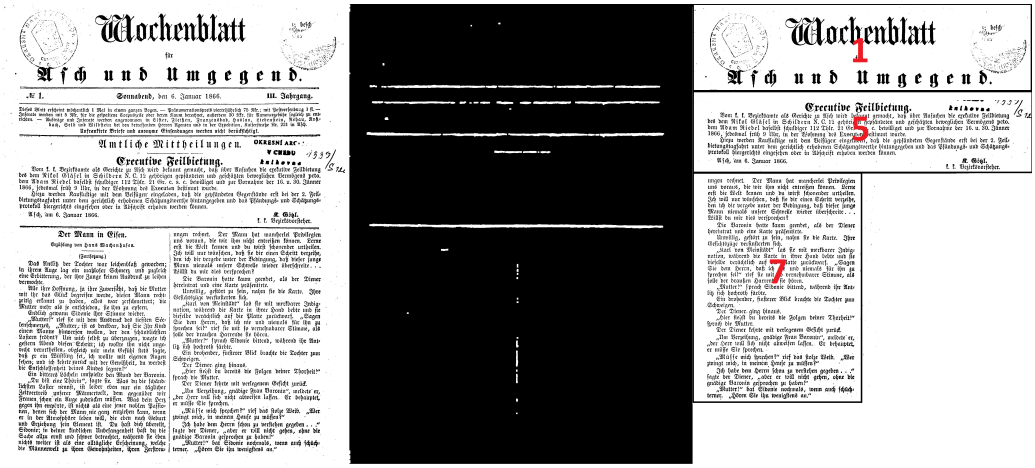


Obrázek 6.22: Vývojový diagram navrženého algoritmu pro rozdělení stránky podle oddělovačů.



Obrázek 6.23: Ukázka kandidátních. Neorámované bílé pixely nejsou při dělení stránky použity.

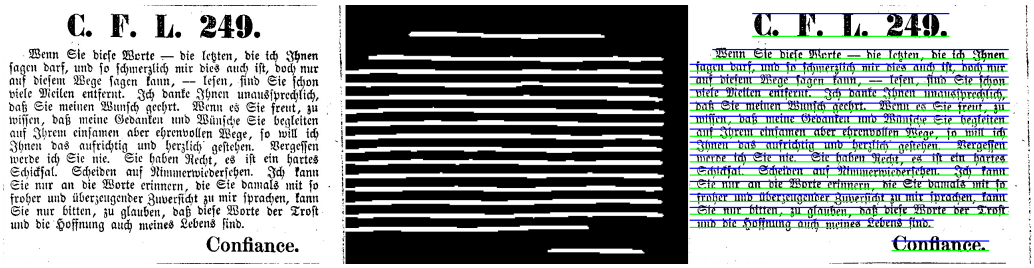
Z nalezených čar jsou následně vyfiltrovány ty, které jsou kratší než je polovina obrázku, nebo oddělovač není uprostřed stránky. Na obrázku 6.23 můžeme vidět kandidáty oddělovačů. V zelených rámečcích jsou označeny oddělovače, které jsou ve svém bloku delší než je polovina obrázku. Hnědou barvou jsou pak označeny oddělovače, které nejsou delší než je polovina obrázku, ale jsou na středu. Pokud i přesto nějaké oddělovače zůstaly nevyfiltrované, je obrázek podle nich horizontálně rozdělen na podobrázky a celý proces se opakuje pro každý podobrázek. Pokud nebyla žádná horizontální čara nalezena, pokusíme se najít stejným způsobem čáru vertikální. Opět pokud byly oddělovače nalezeny, obrázek se dělí vertikálně a proces se znovu opakuje. Pokud ne, zkoumaná část obrázku se exportuje do souboru. Pro shrnutí, obrázek je rekurzivně dělen podle horizontálních a vertikálních čar do té doby, dokud jsou v podobrázcích nalézány oddělovače. Rekurze navíc v navrženém algoritmu zajišťuje, že se jako první vyexportuje podobrázek, který se vyskytuje nejvíce vlevo nahoře a jako poslední podobrázek vlevo dole. Tato posloupnost přesně odpovídá pořadí čtení, na které jsme zvyklí. Výsledek dělení můžeme vidět na obrázku 6.24.



Obrázek 6.24: Ukázka rozdělení stránky podle oddělovačů. Vlevo je originální obrázek, uprostřed vidíme obrázek reprezentující oddělovače a vpravo můžeme vidět ukázku oddělených částí. Červená čísla v blocích představují určené pořadí čtení.

6.3 Detekce řádků

Aby bylo možné rozdělit text na jednotlivé řádky a ty následně předložit procesu OCR, je potřeba řádky nejprve v obrázku detekovat. K tomu je



Obrázek 6.25: Vlevo můžeme vidět vstupní obrázek, uprostřed označené základní čáry sítí ARU-Net, vpravo nalezené horní a spodní hranice jednotlivých řádků.

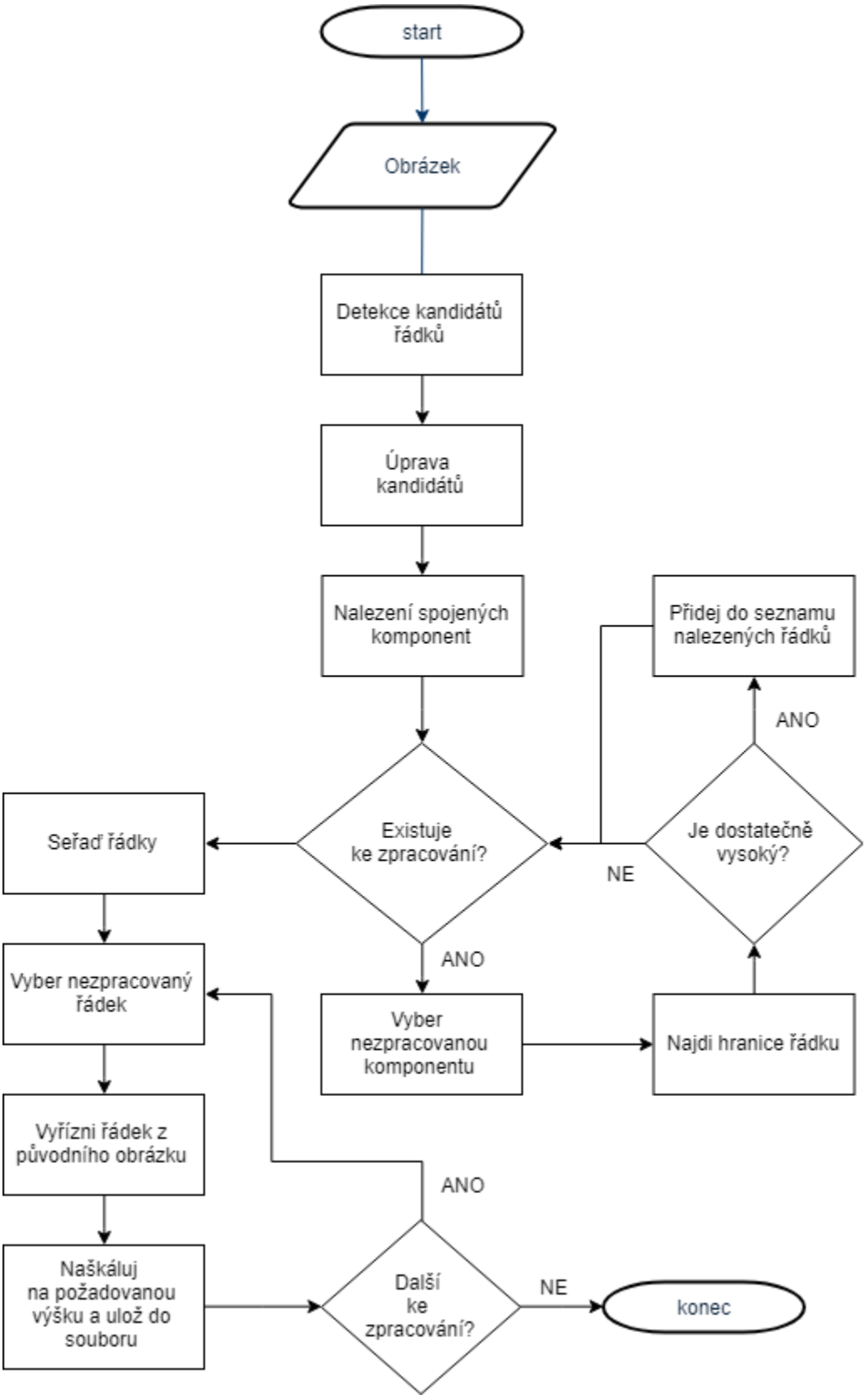
využito předtrénované sítě ARU-Net vytvořené v rámci výzkumu [16] na datové sadě READ-BAD [15], která obsahuje celkem 256 označených obrázků. Tuto síť je možné získat na `GitHub` stránkách projektu [1]. Tento předtrénovaný model je využit z důvodu, že datové sady Europeana a Portafontium neobsahují `ground-truth` označující řádky textů a navíc nebyly nalezeny takto označené datové sady, které by se strukturou stránek a použitým jazykem podobaly datové sadě Portafontium. Prostřední obrázek v 6.25 ukazuje výstup z této sítě. Vpravo pak můžeme vidět označené bloky řádků, kdy zelená barva reprezentuje spodní hranici a modrá horní. Nyní bude popsáno, jakým způsobem jsou tyto hranice určeny.

Vývojový diagram tohoto algoritmu můžeme vidět na obrázku 6.26. Vstupní obrázek I celého algoritmu je přiveden na vstup síte ARU-Net. Výstupem této sítě je černobílý obrázek I_l, kde bílá barva reprezentuje detekované kandidáty řádků. Přesněji čáru, která prochází pod textem (tzv. `baseline`). Na tento obrázek je následně aplikována dilatace, následovaná erozí s rozměrem 15×1. Díky tomu dojde k odstranění malých děr. V takto upraveném obrázku I_l jsou nalezeny spojené komponenty. K tomu bylo využito knihovny pro Python `OpenCV`, konkrétně funkce `connectedComponentsWithStats`. Z návratových hodnot lze pro každou komponentu zjistit, jaké největší a nejmenší y souřadnice nalezená komponenta dosahuje. Stejně tak i pro x souřadnici. Když máme určeno, kde nejníže je řádek umístěn, vyřízne se z původního obrázku oblast o velikosti 170 pixelů nad touto hranicí (v datové sadě Portafontium bylo vypozorováno, že nejvyšší nadpis dosahuje výšky 155 pixelů).

Pixely této oblasti jsou invertovány, tedy z bílých pixelů se stanou černé a naopak. Pro tuto oblast se spočítá horizontální profil – pro každý řádek

[1] https://github.com/TobiasGruening/ARU-Net

je sečtena intenzita všech pixelů a vydělena šířkou řádku. Díky tomu je získána průměrná intenzita řádku. Následně je ve vyříznuté oblasti postupně iterováno od nejspodnější pozice směrem nahoru a to do té doby, dokud je průměrna intenzita řádku větší než 30. Tato hodnota byla zvolena na základě experimentů, kdy byly zkoumány horizontální profily několika řádků. Pokud průměrná intenzita klesne pod 30, je považováno, že se v tomto řádku již nevyskytuje text a dosáhli jsme horní hranice textu. Takto je získána minimální a maximální x a y hodnota, čímž získáme obdélník, ve kterém je text obsažen. Z popisu vyplývá, že takto určené obdélníky budou mít různou výšku. Pro následné zpracování OCR je ale žádoucí, aby všechny řádky textu byly stejné výšky. Z tohoto důvodu je spočítán poměr p mezi žádanou výškou a výškou obdélníku. Na obrázku 6.26 vpravo si můžeme všimnout, že modrá čára reprezentující horní hranici obdélníku často nesahá až na vrchol písmen (například poslední text „Confiance“). Stejně tak zelená čára, reprezentující spodní hranici obdélníku, nepokrývá části písmen, které zasahují pod základní čáru (například písmeno „g“) Z tohoto důvodu výška tohoto obdélníku rozšířena na obě strany o polovinu původní výšky. Takto rozšířený obdélník je následně výříznut z původního obrázku. Tomuto vyříznutému podobrázku je následně upravena velikost podle poměru p a výsledek uložíme do souboru. Díky tomu je docíleno, že všechny vyříznuté bloky řádků budou mít velmi podobnou velikost (záleží na přesnosti určení horní a spodní hranice).

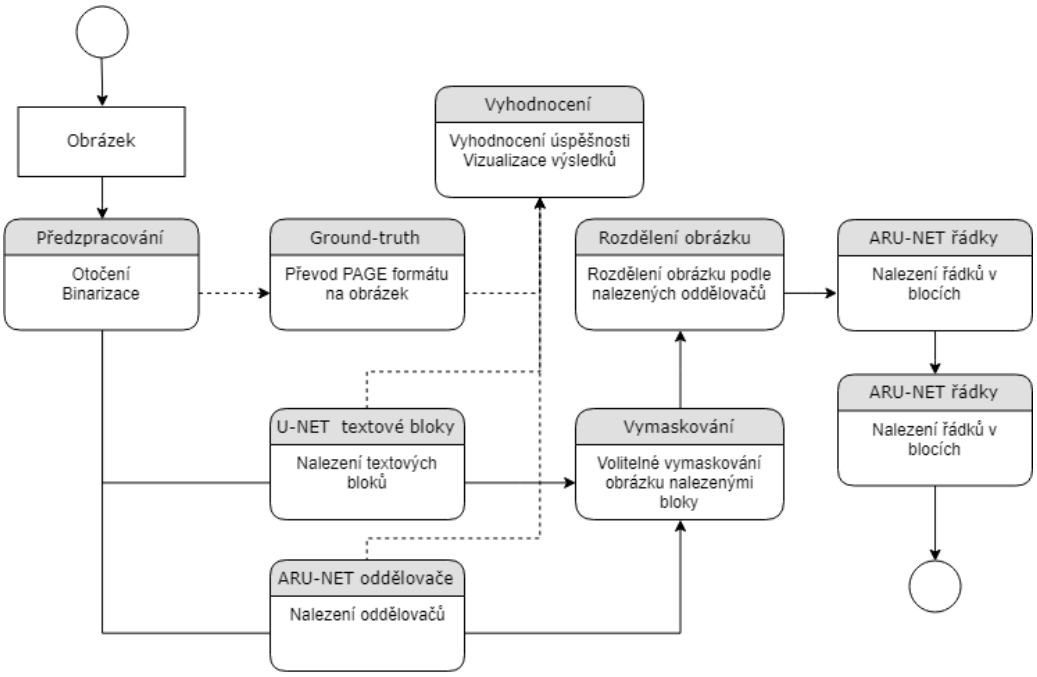


Obrázek 6.26: Vývojový diagram získání řádků textu z obrázku.

7 Architektura aplikace

V diagramu 7.1 můžeme vidět průběh zpracování vstupního obrázku v navržené architektuře. Vstupní obrázek I je nejprve předzpracován stejným způsobem jako při tvorbě `ground-truth` (viz sekce 4.2.4). V tomto obrázku jsou následně nalezeny bloky obsahující text s využitím sítě upravený U-Net, čímž získáme obrázek I_{text}. Dále jsou v obrázku I nalezeny oddělovače sítí ARU-Net natrénované na segmentaci separátorů. Výstupem je obrázek I_{sep}. Původní obrázek I může být vymaskován obrázkem I_{text}, kdy v obrázku I zůstanou pouze ty pixely, které jsou součástí I_{text}. Tím je docíleno odstranění šumu v oblastech mimo textové oblasti, který by mohl způsobovat nepřesnosti při hledání řádků textu nebo při zpracování OCR systémem. Zároveň však může dojít k situaci, kdy text v obrázku I nebude označen v I_{text} a dojde tak k tomu, že text v obrázku I bude odstraněn. Následuje rozdělení obrázku podle nalezených oddělovačů v obrázku I_{sep} popsané v sekci 6.2.6. Vstupní obrázek se nám rozdělí na podobrázky $I_{frac_1}, I_{frac_2} \cdots, I_{frac_n}$. Obrázky I_{frac} jsou dále vstupem sítě ARU-Net, která je tentokrát natrénovaná pro hledání textových řádků. Nalezené řádky jsou následně z obrázku extrahovány a uloženy do souboru. Proces extrakce řádků textu je popsán v sekci 6.3.

Pro vyhodnocení úspěšnosti klasifikátorů pro nalezení textových bloků a oddělovačů, je potřeba mít ke zkoumanému obrázku `ground-truth` data s informacemi, kde se jednotlivé textové bloky a oddělovače vyskytují. Z nich jsou vytvořeny očekávané výsledky klasifikátorů. $I_{gt_{text}}$ pro textové bloky a $I_{gt_{sep}}$ pro oddělovače. Výsledky klasifikátorů I_{text} a I_{sep} jsou na základě metrik – přesnosti pixelů v popředí, Jaccardova koeficientu a F1-skóre (popsáno v sekci 5.1) vyhodnoceny právě s obrázky $I_{gt_{text}}$ a $I_{gt_{sep}}$. Dále jsou tyto obrázky vstupem nástroje pro vizualizaci zmíněnem v sekci 5.1.1.



Obrázek 7.1: Data flow diagram navrženého programu.

7.1 Použité technologie

Pro realizaci diplomové práce byl použit programovací jazyk `Python`. Síť ARU-Net je naimplementována s využitím knihovny pro strojové učení `TensorFlow` [1]. Implementace sítí U-Net byla realizována s knihovnou `Keras` [2]. Pro morfologické operace s obrázkem jsem použil knihovnu `OpenCV` [3]. Pro předzpracování obrázku byl použit nástroj `ocropy` [4], který vyžaduje verzi `Pythonu` 2, 7. Knihovna `Keras` naopak v současné době vyžaduje verzi 3, 5 a tak byly v práci použity obě dvě verze.

[1] https://www.tensorflow.org/
[2] https://keras.io/
[3] https://opencv.org/
[4] https://github.com/tmbdev/ocropy

8 Dosažené výsledky

Pro testovací sadu obrázků z datové sady Portafontium byly spočítány statistiky zvolených modelů. Kromě statistik budou použity pro zhodnocení i obrázky vizualizující chyby označení. Jak bylo zmíněno v sekci 5.1.1, zelená barva reprezentuje správné označení. Červená barva říká, že model předpověděl pixel jako popředí, ale v `ground-truth` tento pixel není označen. Tyrkysová barva pak značí, že byl pixel v `ground-truth` označen jako popředí, ale model ho označil jako pozadí.

V tabulce 8.1 můžeme vidět výsledky modelu pro segmentaci textu uU_{pftext}. Ukázku vizualizace správnosti označení můžeme vidět na obrázku 8.1. Z vizualizace je vidět, že $FgPA$ je snižována horší schopností modelu označení některých jednořádkových textů. Co se týče větších bloků, model je označuje téměř bez problémů.

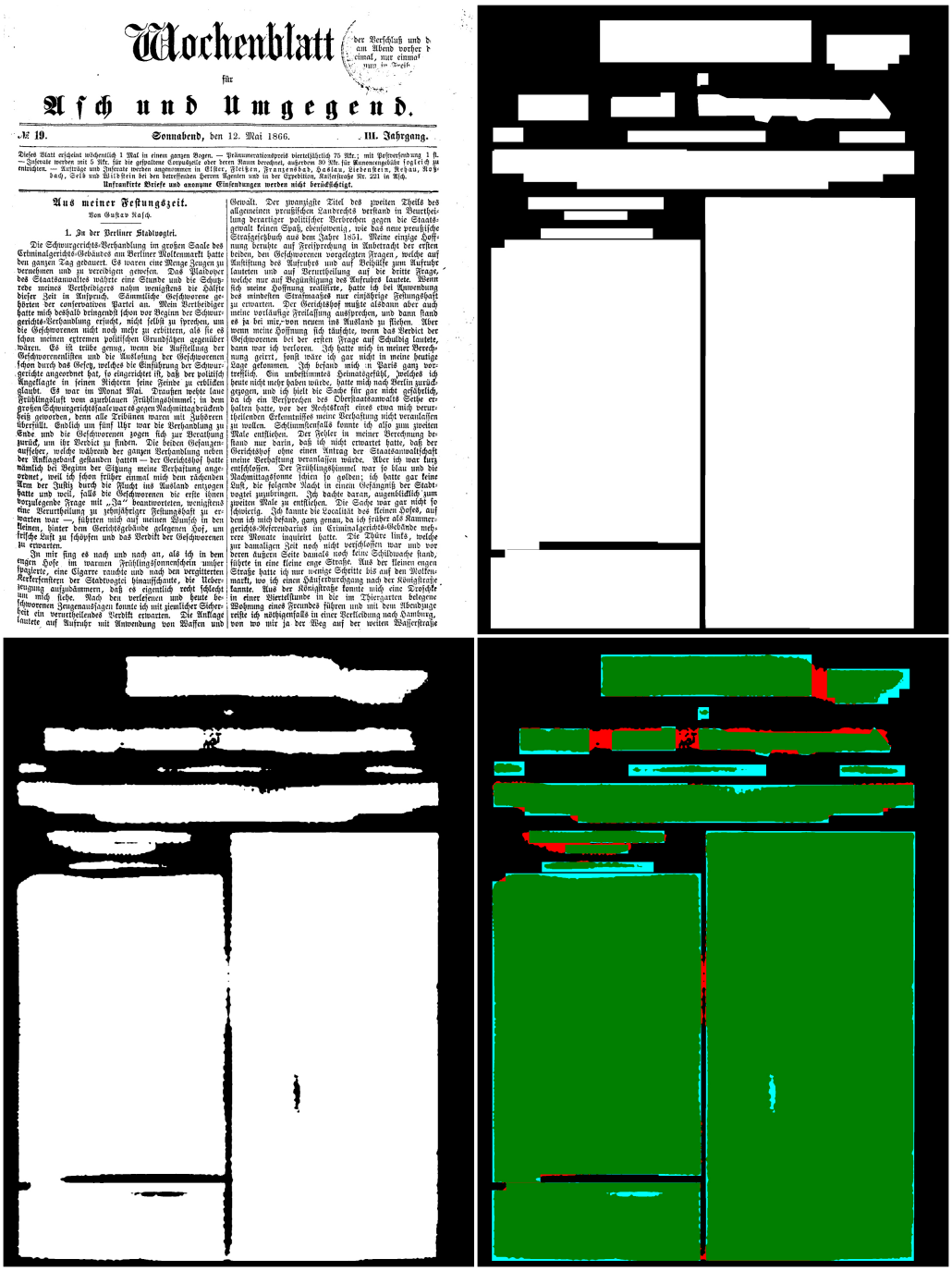
Model	Jaccard	F1-skóre	FgPA
uU_{pftext}	0,881	0,936	0,920

Tabulka 8.1: Výsledky segmentace textu. Metriky: průměr Jaccardova koeficientu, průměr F1-Skóre a přesnost pixelu v popředí $FgPA$.

Model	Jaccard	F1-skóre	FgPA
ARU_{pfsep}	0,565	0,625	0,768

Tabulka 8.2: Výsledky segmentace oddělovačů. Metriky: průměr Jaccardova koeficientu, průměr F1-Skóre a přesnost pixelu v popředí $FgPA$.

V druhé tabulce 8.2 jsou statistiky segmentace oddělovačů uU_{pftext}. Ukázka vizualizace je pak na obrázku 8.2. Na obrázku si můžeme všimnout, že model kromě oddělovačů označuje i spoustu pixelů tvořící text. Tento jev se projevuje i na horší hodnotě F1-skóre oproti hodnotám získaným při validaci. Většinou se však jedná o osamocené pixely, které budou během rozdělování obrázku odstraněny morfologickými operacemi (viz. sekce 6.2.6). Ukázka odstranění je na obrázku 8.3. V obrázku stále zůstávají některé špatně označené pixely, nicméně na nalezení oddělujících čar již nemají výrazný vliv. Hledané čáry jsou označeny dostačujícím způsobem pro jejich detekci v rámci navržených algoritmů (viz. sekce 6.2.6).



Wochenblatt

für

Asch und Umgegend.

№ 19. Sonnabend, den 12. Mai 1866. III. Jahrgang.

Aus meiner Festungszeit.

Von Gustav Rasch.

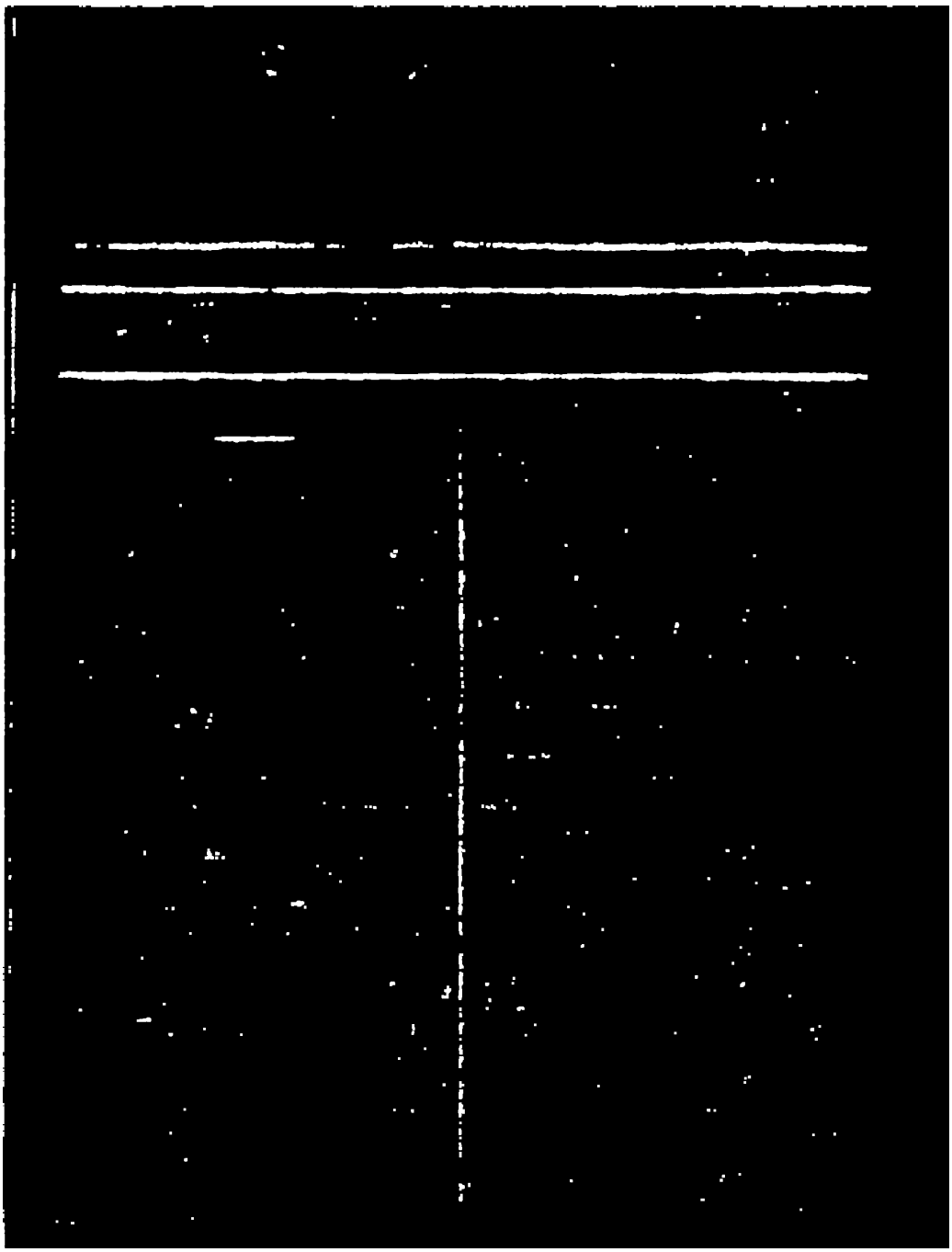
1. In der Berliner Stadtvogtei.

Die Schwurgerichts-Verhandlung im großen Saale des Criminalgerichts-Gebäudes am Berliner Molkenmarkt hatte den ganzen Tag gedauert. Es waren eine Menge Zeugen zu vernehmen und zu vereidigen gewesen. Das Plaidoyer des Staatsanwaltes währte eine Stunde und die Schlußrede meines Vertheidigers nahm wenigstens die Hälfte dieser Zeit in Anspruch. Sämmtliche Geschworene gehörten der conservativen Partei an. Mein Vertheidiger hatte mich deshalb dringendst schon vor Beginn der Schwurgerichts-Verhandlung ersucht, nicht selbst zu sprechen, um die Geschworenen nicht noch mehr zu erbittern, als sie es schon meinen extremen politischen Grundsätzen gegenüber wären. Es ist trübe genug, wenn die Aufstellung der Geschworenenlisten und die Auslosung der Geschworenen schon durch das Gesetz, welches die Einführung der Schwurgerichte angeordnet hat, so eingerichtet ist, daß der politisch Angeklagte in seinen Richtern seine Feinde zu erblicken glaubt. Es war im Monat Mai. Draußen wehte laue Frühlingsluft vom azurblauen Frühlingshimmel; in dem großen Schwurgerichtssaale war es gegen Nachmittag drückend heiß geworden, denn alle Tribünen waren mit Zuhörern überfüllt. Endlich um fünf Uhr war die Verhandlung zu Ende und die Geschworenen zogen sich zur Berathung zurück, um ihr Verdict zu finden. Die beiden Gefangenaufseher, welche während der ganzen Verhandlung neben der Anklagebank gestanden hatten — der Gerichtshof hatte nämlich bei Beginn der Sitzung meine Verhaftung angeordnet, weil ich schon früher einmal mich dem rächenden Arm der Justiz durch die Flucht ins Ausland entzogen hatte und weil, falls die Geschworenen die erste ihnen vorzulegende Frage mit „Ja“ beantworteten, wenigstens eine Verurtheilung zu mehrjähriger Festungshaft zu erwarten war —, führten mich auf meinen Wunsch in den kleinen, hinter dem Gerichtsgebäude gelegenen Hof, um frische Luft zu schöpfen und das Verdict der Geschworenen zu erwarten.

In mir fing es nach und nach an, als ich in dem engen Hofe im warmen Frühlingssonnenschein unser spazierte, eine Cigarre rauchte und nach den vergitterten Kerkerfenstern der Stadtvogtei hinaufschaute, die Ueberzeugung aufzudämmern, daß es eigentlich recht schlecht um mich stehe. Nach den verlesenen und heute beschworenen Zeugenaussagen konnte ich mit ziemlicher Sicherheit ein verurtheilendes Verdict erwarten. Die Anklage lautete auf Aufruhr mit Anwendung von Waffen und Gewalt. Der zwanzigste Titel des zweiten Theils des allgemeinen preußischen Landrechts verstand in Beurtheilung derartiger politischer Verbrechen gegen die Staatsgewalt keinen Spaß, ebensowenig, wie das neue preußische Strafgesetzbuch aus dem Jahre 1851. Meine einzige Hoffnung beruhte auf Freisprechung in Anbetracht der ersten beiden, den Geschworenen vorgelegten Fragen, welche auf Anstiftung des Aufruhrs und auf Beihülfe zum Aufruhr lauteten und auf Verurtheilung auf die dritte Frage, welche nur auf Begünstigung des Aufruhrs lautete. Wenn sich meine Hoffnung realisirte, hatte ich bei Anwendung des mindesten Strafmaaßes nur einjährige Festungshaft zu erwarten. Der Gerichtshof mußte alsdann aber auch meine vorläufige Freilassung aussprechen, und dann stand es ja bei mir, von neuem ins Ausland zu fliehen. Aber wenn meine Hoffnung sich täuschte, wenn das Verdict der Geschworenen bei der ersten Frage auf Schuldig lautete, dann war ich verloren. Ich hatte mich in meiner Berechnung geirrt, sonst wäre ich gar nicht in meine heutige Lage gekommen. Ich befand mich in Paris ganz vortrefflich. Ein unbestimmtes Heimatsgefühl, welches ich heute nicht mehr haben würde, hatte mich nach Berlin zurückgezogen, und ich hielt die Sache für gar nicht gefährlich, da ich ein Versprechen des Oberstaatsanwalts Seithe erhalten hatte, vor der Reichskraft eines etwa mich verurtheilenden Erkenntnisses meine Verhaftung nicht veranlassen zu wollen. Schlimmstenfalls konnte ich also zum zweiten Male entfliehen. Der Fehler in meiner Berechnung bestand nur darin, daß ich nicht erwartet hatte, daß der Gerichtshof ohne einen Antrag der Staatsanwaltschaft meine Verhaftung veranlassen würde. Aber ich war kurz entschlossen. Der Frühlingshimmel war so blau und die Nachmittagssonne schien so golden; ich hatte gar keine Lust, die folgende Nacht in einem Gefängniß der Stadtvogtei zuzubringen. Ich dachte daran, augenblicklich zum zweiten Male zu entfliehen. Die Sache war gar nicht so schwierig. Ich kannte die Localität des kleinen Hofes, auf dem ich mich befand, ganz genau, da ich früher als Kammergerichts-Referendarius im Criminalgerichts-Gebäude mehrere Monate inquirirt hatte. Die Thüre links, welche zur damaligen Zeit noch nicht verschlossen war und vor deren äußern Seite damals noch keine Schildwache stand, führte in eine kleine enge Straße. Aus der kleinen engen Straße hatte ich nur wenige Schritte bis auf den Molkenmarkt, wo ich einen Häuserdurchgang nach der Königstraße kannte. Aus der Königstraße konnte mich eine Droschke in einer Viertelstunde in die im Thiergarten belegene Wohnung eines Freundes führen und mit dem Abendzuge reiste ich nöthigenfalls in einer Verkleidung nach Hamburg, von wo mir ja der Weg auf der weiten Wasserstraße

Obrázek 8.1: Ukázka vizualizace označení textových bloků. Originální obrázek, `ground-truth`, výsledek segmentace a vizualizace označení.

Obrázek 8.2: Ukázka vizualizace označení oddělovačů. Originální obrázek, `ground-truth`, výsledek segmentace a vizualizace označení.



Obrázek 8.3: Ukázka odstranění osamocených pixelů. Podrobněji popsáno v sekci 6.2.6.

9 Závěr

V rámci diplomové práce byly na úvod prostudovány metody pro segmentaci historických obrazových dokumentů. Výstup procesu segmentace textu je typicky přiveden na vstup procesu optického rozpoznávání znaků. Konkrétně se jedná o jednotlivé řádky textu. Kromě pouhého převedení textu naskenovaného dokumentu do počítače, je žádoucí, aby bylo zachováno i pořadí čtení. Analýza dodaných dat ukázala, že textové bloky, které spolu logicky souvisí, jsou odděleny čárami. Díky těmto čarám je možné stránku na tyto bloky rozdělit a následnou detekcí řádků uvnitř bloků zachovat pořadí čtení. Z těchto požadavků tak vyplynulo, že bude potřeba vytvořit tři typy modelů: model pro označení textových bloků, model pro nalezení oddělujících čar a model pro označení řádků textu. Z prostudovaných materiálů tak byly vybrány plně konvoluční neuronové sítě ARU-Net, U-Net a její adaptace, která se osvědčila při segmentaci historických dokumentů.

Pro natrénování modelů a následné vyhodnocení je potřeba připravit datovou sadu, která bude tvořena dvojicemi obrázků: segmentovaný obrázek a očekávaný výstup segmentace. Datová sada Europeana, která je strukturou stránek a použitým jazykem podobná dodané datové sadě, tyto dvojice obsahuje. Vzhledem k tomu, že dodaná datová sada neměla označené očekávané výstupy, bylo potřeba je vytvořit. Bylo zjištěno, že pro uchování označených očekávaných výsledků segmentace se používají XML soubory se strukturou GEDI nebo PAGE, pro jejichž vytvoření pak slouží nástroje GEDI a Aletheia. Ukázalo se, že oba formáty umožňují uchovávat stejné typy informací. Formát PAGE je však oproti formátu GEDI rozšířenější a navíc byl použit pro anotaci datové sady Europeana, která je v této práci také použita. Celkem bylo označeno 17 stránek z dodané datové sady, z toho 10 z nich bylo použito pro trénování modelů, 3 byly použity jako validační a 4 pro testování.

Následně byly provedeny experimenty, které ukázaly použitelnost jednotlivých typů plně konvolučních neuronových sítí na zmíněné úlohy. Z experimentů tak vyplynulo, že pro označování textových bloků se pro dodanou datovou sadu nejlépe hodí adaptace sítě U-Net, pro hledání oddělujících čar síť ARU-Net a pro detekci řádků textu síť ARU-Net pro tento účel natrénovaná. Na základě těchto poznatků byl vyvinut program, který vstupní obrázek rozdělí na bloky podle oddělujících čar a následně z nich extrahuje řádky vhodné pro další zpracování OCR.

Výsledkem segmentace je obrázek s označenými pixely. Nalezením spojených komponent je možné nalézt souvislé bloky. Detekované bloky jsou definovány body tvořícími polygon. V rámci rozšíření práce by pak tyto polygony mohly být exportovány do XML PAGE souboru, čímž by se umožnila správa bloků v nástroji Aletheia. Dále by také mohl být pro vyhodnocení použit nástroj PRImA Performance Evaluation (viz. sekce 5.2.4), který pro vyhodnocování používá právě PAGE soubory. Aby vytvořený nástroj byl obecnější, nabízí se možnost práci rozšířit o segmentaci bílých oddělovačů. Detekce těchto oddělovačů však pro dodanou datovou sadu nebyla nutná. Dále se jeví, že se metriky používané pro vyhodnocení segmentace tolik nehodí pro vyhodnocení segmentace oddělovačů, což se ukázalo během experimentů. Síť ARU-Net detekovala vizuálně oddělovače velmi dobře, nicméně čísla metrik tomu tolik neodpovídají. Mohla by se sledovat například celistvost detekovaných čar.

10 Použité zkratky

GEDI – Groundtruthing Environment for Document Images

XML – eXtensible Markup Language

GT – Ground-truth

LBP – Local Binary Pattern

FCBF – Fast Correlation-based Filter

SVM – Support Vector Machine

CNN – Convolutional Neural Network

FCNN – Fully Convolutional Neural Network

RelU – Rectified Linear Unit

OCR – Optical Character Recognition

Literatura

[1] *Python-based tools for document analysis and OCR* [online]. Dostupné z: `https://github.com/tmbdev/ocropy`.

[2] Achanta, R. et al. Slic superpixels. Technical report, 2010.

[3] Ahonen, T. – Hadid, A. – Pietikainen, M. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*. 2006, , 12, s. 2037–2041.

[4] Alberti, M. et al. Open Evaluation Tool for Layout Analysis of Document Images. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, s. 43–47, Kyoto, Japan, nov 2017. doi: 10.1109/ICDAR.2017.311. ISBN 978-1-5386-3586-5.

[5] Alberti, M. et al. *DIVA-DIA* [online]. 2019. Dostupné z: `https://github.com/DIVA-DIA/DIVA_Layout_Analysis_Evaluator`.

[6] Antonacopoulos, A. – Ritchings, R. Representation and classification of complex-shaped printed regions using white tiles. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 2, s. 1132–1135. IEEE, 1995.

[7] Antonacopoulos, A. et al. ICDAR 2009 page segmentation competition. In *2009 10th International Conference on Document Analysis and Recognition*, s. 1370–1374. IEEE, 2009.

[8] Chen, K. et al. Page segmentation for historical handwritten document images using color and texture features. In *2014 14th International Conference on Frontiers in Handwriting Recognition*, s. 488–493. IEEE, 2014.

[9] Chen, K. et al. Convolutional neural networks for page segmentation of historical document images. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 1, s. 965–970. IEEE, 2017.

[10] Clausner, C. – Pletschacher, S. – Antonacopoulos, A. Scenario driven in-depth performance evaluation of document layout analysis methods. In *2011 International Conference on Document Analysis and Recognition*, s. 1404–1408. IEEE, 2011.

[11] Clausner, C. – Antonacopoulos, A. – Pletschacher, S. Icdar2017 competition on recognition of documents with complex layouts-rdcl2017. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 1, s. 1404–1410. IEEE, 2017.

[12] Cortes, C. – Vapnik, V. Support-vector networks. *Machine learning*. 1995, 20, 3, s. 273–297.

[13] Duda, R. O. – Hart, P. E. Use of the Hough transformation to detect lines and curves in pictures. Technical report, SRI INTERNATIONAL MENLO PARK CA ARTIFICIAL INTELLIGENCE CENTER, 1971.

[14] Garg, R. et al. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European Conference on Computer Vision*, s. 740–756. Springer, 2016.

[15] Grüning, T. et al. READ-BAD: A New Dataset and Evaluation Scheme for Baseline Detection in Archival Documents. 2017. Dostupné z: `http://arxiv.org/abs/1705.03311`.

[16] Grüning, T. et al. A two-stage method for text line detection in historical documents. *arXiv preprint arXiv:1802.03345*. 2018.

[17] Hamers, L. – others. Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula. *Information Processing and Management*. 1989, 25, 3, s. 315–18.

[18] LeCun, Y. et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*. 1998, 86, 11, s. 2278–2324.

[19] Pletschacher, S. – Antonacopoulos, A. The PAGE (page analysis and ground-truth elements) format framework. In *2010 20th International Conference on Pattern Recognition*, s. 257–260. IEEE, 2010.

[20] Portafontium. *Cíle* [online]. 2019. Dostupné z: `http://www.portafontium.eu/`.

[21] Research, P. *Layout Analysis Dataset A realistic contemporary document dataset* [online]. 2017. Dostupné z: `https://github.com/DIVA-DIA/DIVA_Layout_Analysis_Evaluator`.

[22] Ronneberger, O. – Fischer, P. – Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, s. 234–241. Springer, 2015.

[23] Rosebrock, A. *Intersection over Union (IoU) for object detection* [online]. 2016. Dostupné z: `https://www.pyimagesearch.com`.

[24] Srivastava, N. et al. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*. 2014, 15, 1, s. 1929–1958.

[25] Wei, H. et al. Evaluation of SVM, MLP and GMM classifiers for layout analysis of historical documents. In *2013 12th International Conference on Document Analysis and Recognition*, s. 1220–1224. IEEE, 2013.

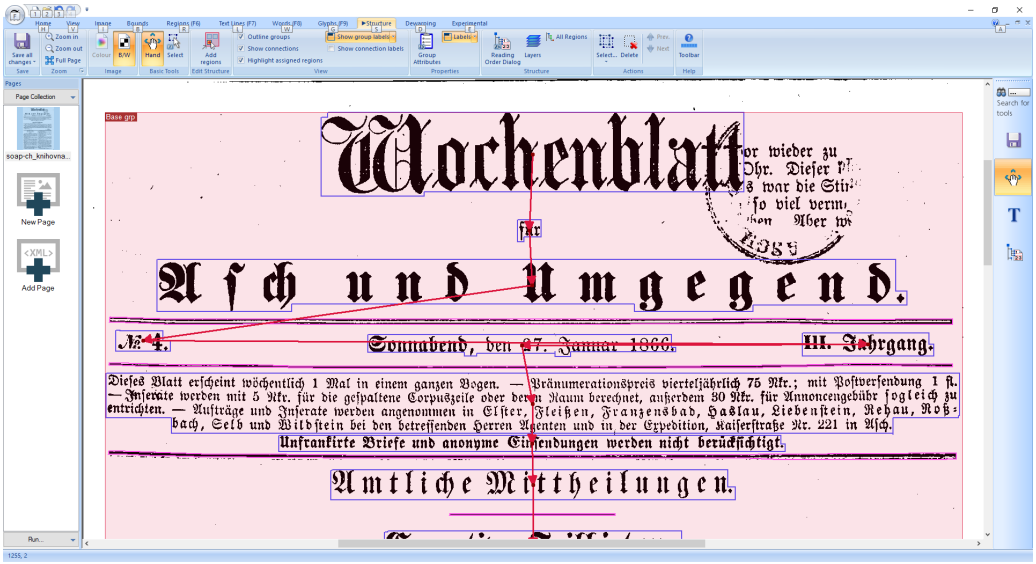
[26] Wick, C. – Puppe, F. Fully convolutional neural networks for page segmentation of historical document images. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, s. 287–292. IEEE, 2018.

[27] Yu, L. – Liu, H. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, s. 856–863, 2003.

A Přiložené obrázky

A.1 Ukázka nástrojů pro tvorbu Ground-truth

Obrázek A.1: Ukázka anotace v nástroji GEDI



Obrázek A.2: Ukázka anotace v nástroji Aletheia

B Obsah přiloženého DVD

Přiložené DVD obsahuje text práce ve formátu **pdf** spolu s jeho zdrojovými soubory ve složce **text**. Implementace diplomové práce je umístěna ve ve složce **src**. Ve složce **data** jsou umístěny všechny natrénované modely a také anotované obrázky tvořící datovou sadu Portafontium. Složka **poster** obsahuje zdrojový soubor s vytvořeným plakátem diplomové práce a jeho **pdf** verze. Struktura přiloženého DVD je podrobněji popsána v souboru **readme.txt**, který se nachází v kořenovém adresáři. Adresářová struktura zdrojových kódů je popsána v souboru **/src/Segmentation/readme.txt**

C Příručka

C.1 Instalace

Vytvořený program vyžaduje mít nainstalovaný Python verze alespoň 3.5.6. Pro předzpracování obrázků byl použit nástroj Ocropy. Jeho instalace je popsána na GitHub stránkách projektu [1]. Pro použití nástroje Ocropy je nutné mít nainstalovaný Python verze 2.7.

Další závislosti potřebné pro běh programu jsou uloženy v souboru **src/requirements.txt**. Závislosti se nainstalujcí s využitím příkazové řádky a balíčkového manažera PIP následovně:

```
pip install -r requirements.txt
```

Po doinstalování balíčků je potřeba ještě upravit ve zdrojovém souboru **src/Segmentation/Segmentation/Segmentation.py** cestu k nainstalovaným skriptům nástroje Ocropy a k Pythonu verze 2.7. Cesty se nastavují na začátku souboru do proměnných `binarizationFilePath` a `python2Path`.

C.2 Spuštění

Program, který slouží k nalezení řádků v obrázku je umístěn v souboru **src/Segmentation/Segmentation/Segmentation.py** a spouští se příkazem

```
python ./Segmentation.py
```

Program má povinné dva parametry:

- --imagePath - celá cesta k segmentovanému obrázku.
- --outputPath - cesta, kam budou exportovány výsledky.

Program do zvolené složky exportuje obrázek s označenými textovými bloky a oddělovači. Následně je obrázek rozdělen na bloky podle oddělovačů do složky, která se jmenuje jako vstupní obrázek. Rozdělené bloky jsou dále rozděleny na řádky. Další použité zdrojové soubory jsou podrobně okomentovány a jejich struktura je popsána v souboru **/src/Segmentation/readme.txt**.

[1] https://github.com/tmbdev/ocropy