FAKULTA APLIKOVANÝCH VĚD
ZÁPADOČESKÉ UNIVERZITY
V PLZNI

STUDENTSKÁ
VĚDECKÁ KONFERENCE 2022

Transformer co-attention for word segmentation in image data

Tomáš Zítka[1]

1 Introduction

Despite the huge progress in the past years, optical character recognition (OCR) of handwritten text is still a hard task. A prerequisite to it is extracting parts of the image that contain the text – line segmentation. In this work, we presume that line segmentation is available and are interested in splitting the lines into individual words – word segmentation. Such word segmentation can have many uses e.g. highlighting relevant words on the page when performing full text search or supplementing the full text search with search based on visual similarity between handwritten words.

Recently machine learning OCR algorithms have drawn inspiration from models used in natural language processing called transformers [2]. In our work, we employ the transformer model based on [1] and use its internal co-attention mechanism to segment an image of a line of handwritten text into words.

2 Co-attention in transformer models

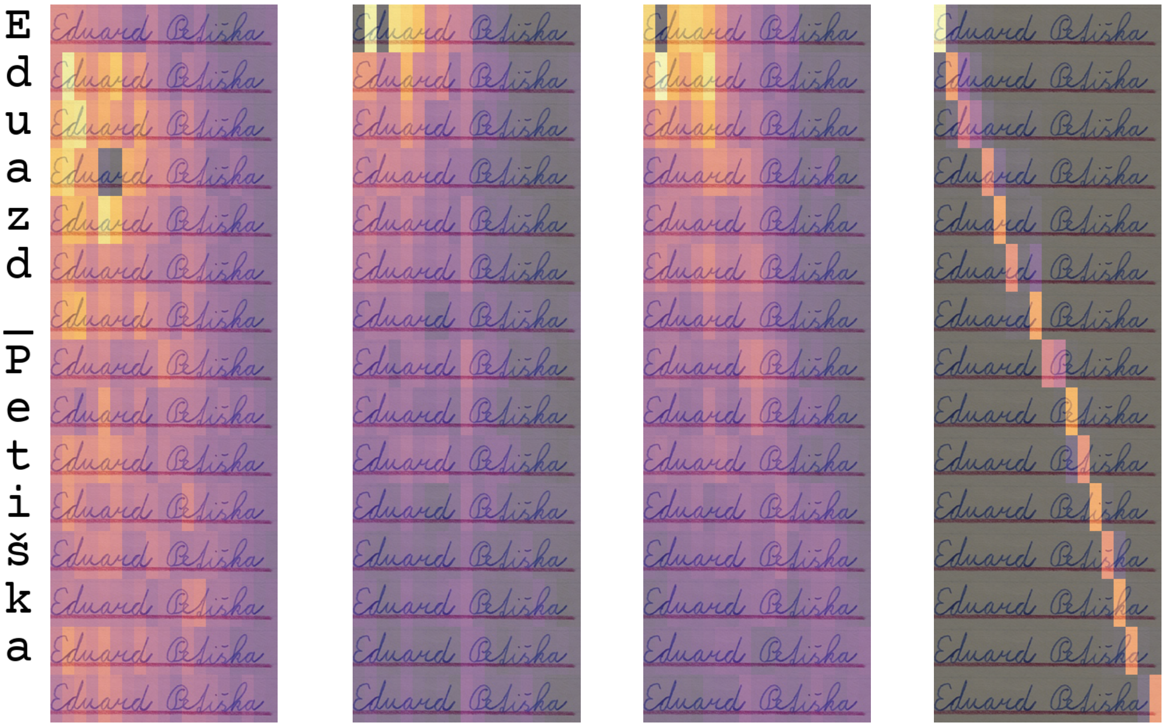
The transformer model consists of two parts the encoder and the decoder [2, p. 3]. The encoder takes an input sequence, in our case, this is a picture projected by convolutional neural network, flattened into a sequence of vectors, and through a sequence of self-attention heads and multi-layer perceptrons, it processes it, preserving the shape. The decoder part of the transformer then recursively predicts the output sequence from a starting token combining previous predictions and the output of the encoder using the co-attention mechanism.

Let $M \in \mathbb{R}^{w\times c}$ be output of the encoder, sometime called memory, and $X \in \mathbb{R}^{l\times d}$ decoder output sequence so far, the co-attention between decoder and encoder $A \in \mathbb{R}^{w\times l}$ is computed as [2, p. 4]

$$A = \text{softmax}\left(\frac{1}{\sqrt{d}} MQ \cdot (XK)^T\right), \tag{1}$$

where the softmax function is applied row-wise and $Q \in \mathbb{R}^{c\times d}$ and $K \in \mathbb{R}^{d\times d}$ are learnable projection matrices, $q = MQ$ and $k = XK$ are called query and key respectively. Example co-attention for each of four layers in decoder averaged over all heads in the layer can be seen in Figure 1. Co-attention of the layer 4 used to predict the space character (represented by underscore ”_” in the figure) can be used to find word separators. We have tested this approach on our dataset of czech hanwriten texts achieving 0.821 mean intersection over union.

[1] student navazujícího doktorského studijního programu Aplikované vědy a informatika, obor Kybernetika, specializace Multimodální interakce člověka-stroj, e-mail: zitkat@students.zcu.cz



(a) Layer 1 **(b)** Layer 2 **(c)** Layer 3 **(d)** Layer 4

Figure 1: Mean co-attentions for each predicted character (displayed in left column).

3 Conclusion

We demonstrated that decoder co-attention can be used for word segmentation. In future work, we plan to further study word segmentation, testing our approach on a public handwritten dataset to allow reproducible evaluation. Further, we want to explore other uses of co-attentions. For example, we observed that more diffused attention in the last decoder layer might be indicative of incorrect character prediction and could serve as a confidence measure. We also observed cases where the co-attention did not move monotonically from left to right again indicating potential failures.

Acknowledgement

The work has been supported by the grant of the University of West Bohemia, project No. SGS-2022-017. Computational resources were supplied by the project ”e-Infrastruktura CZ” (e-INFRA LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures.

References

[1] Lei Kang, Pau Riba, Marçal Rusiñol, Alicia Fornés, and Mauricio Villegas. Pay Attention to What You Read: Non-recurrent Handwritten Text-Line Recognition. *arXiv:2005.13044 [cs]*, May 2020.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.