DEVELOPING A Seq2Seq NEURAL NETWORK USING VISUAL ATTENTION TO TRANSFORM MATHEMATICAL EXPRESSIONS FROM IMAGES TO LaTeX

P. Vyaznikov, I. Kotilevets
Federal State Budget Educational Institution of Higher Education «MIREA - Russian Technological University», Moscow, Russia

sha.cehca@yandex.ru

I. INTRODUCTION

In the modern world, Optical Character Recognition technology finds an incredible number of applications (text recognition quickly scanning document). Progress in this area is due to the emergence of advanced deep learning algorithms and neural network models, which, learning from a huge number of examples, can make very accurate predictions.

The task of im2latex, known thanks to the OpenAI company, is to create an OCR neural network capable of converting an image with mathematical expressions into a similar expression in the LaTeX markup language with high accuracy. This problem belongs to the Image Captioning type - the neural network scans the image and, based on the extracted features, generates a description in natural language.

The proposed solution uses the seq2seq architecture, which contains the Encoder and Decoder mechanisms, as well as Bahdanau Attention [1]. This approach makes it possible to achieve high efficiency and accuracy in generating captions of mathematical expressions. The principle of operation of the neural network is shown in Figure 1.

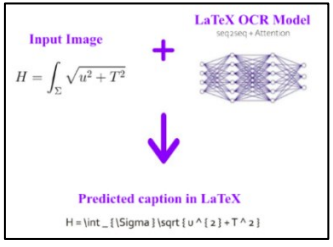

Figure 1. Principle of im2latex

II. DATASET

The training data for im2latex contains 100000 images with mathematical expressions, paired with their true LaTeX label. Before training, all captions are being tokenized (represented as numbers) and then a dictionary is assembled from all individual tokens (i.e., individual LaTeX words), which is used during training process.

III. NEURAL NETWORK ARCHITECTURE

Sequence to Sequence architecture consists of two main components: Encoder and Decoder. These are two connected neural networks training simultaneously, but performing different tasks.

Encoder is the mechanism, that contains a set of convolutional layers and is designed to extract features from photos, which can then be used by Decoder and Attention to generate an output sequence. The number of layers and their settings are very important and may vary depending on the task. After extracting the features, Encoder summarizes it in a form called the Internal State Vector.

After extracting the features, the Decoder begins to train, which in the future will be able to generate caption for any image. The mechanism is based on a recurrent neural network (RNN), which is able to identify and remember important information and ignore irrelevant information.

An important element of Decoder is the Attention mechanism. Its essence is to generate “importance weights”, called the Context Vector, for the output sequences of the Encoder (image features), which are then combined with the input data of the Decoder, which allows the network to learn much more efficiently.

The input data of the Decoder are the real captions corresponding to the dataset images passed through the Encoder. The whole training loop of the neural network is shown on Figure 2.

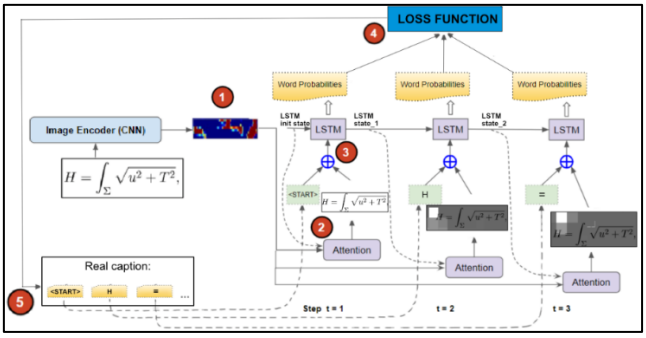

Figure 2. Seq2Seq training loop

IV. HYPERPARAMETERS

The neural network uses SparseCategoricalCrossentropy as a loss function and Adam as an optimizer, “batch_size” is equal to 24 and epochs is set to 15. With such parameters, the network trained for about 14 hours using Nvidia RTX 3080 and 32 gigabytes of RAM. The lowest average loss obtained is 0.025, which is an excellent result for Image Captioning neural networks. The loss graph is shown of Figure 3.

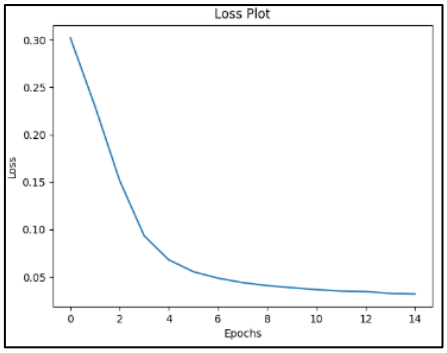

Figure 3. Training Loss graph

V. EVALUATION

Measured BLEU [2] score is about 70% and Levenshtein distance metric is 31. In comparison with similar works [3], the obtained metric measurements are quite high. The competing solution has 40% for BLEU against 70% of the reviewed and 44 for Minimal Edit Distance against 31. Based on the above data, it can be argued that the developed neural network has high efficiency.

VI. CONCLUSIONS

The presented neural network with the seq2seq architecture and Attention mechanism successfully solves the im2latex problem, which is confirmed by the results of measuring metrics. Generated captions for images with equations are quite accurate and, in most cases, coincide with the real ones.

Such a solution can be used in mathematical programs to automatically translate images into LaTeX and further solve and analyze the resulting equations or expressions.

The scope of use can be expanded by training the network on images with handwritten equations. A similar technology is used in the PhotoMath application, however, it has low accuracy and a small set of supported mathematical symbols. The described solution is devoid of both problems.

REFERENCES

[1] NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE. // https://arxiv.org/pdf/1409.0473.pdf

[2] BLEU: a Method for Automatic Evaluation of Machine Translation // https://aclanthology.org/P02-1040.pdf

[3] Im2Latex // https://github.com/luopeixiang/im2latex.