Image captioning Bot

Semyon Semenov, Daniil Merkulov

Optimization Class Project. MIPT

Introduction

Image captioning problem has a great applied significance and until recently could not be satisfactorily solved. Now the accumulated knowledge, advanced deep learning technologies and modern methods of stochastic optimization can significantly improve the result. All the source data of the project can be found at the link: `https://github.com/miptstudent/Image-captioning-Bot`

Problem statement

The task is to build a mapping from the $\mathbb{R}^{n\times m\times 3}$ space of the images to the space of their text descriptions:

$$f : \mathbf{X} \to \mathbf{Y}$$

To find the function, the problem of unconditional optimization is posed:

$$L\big(\mathbf{Y}, \hat{\mathbf{Y}}\big) \to \min$$

As a loss function we use binary cross entropy, as an optimizer, we use Adam.

Model architecture

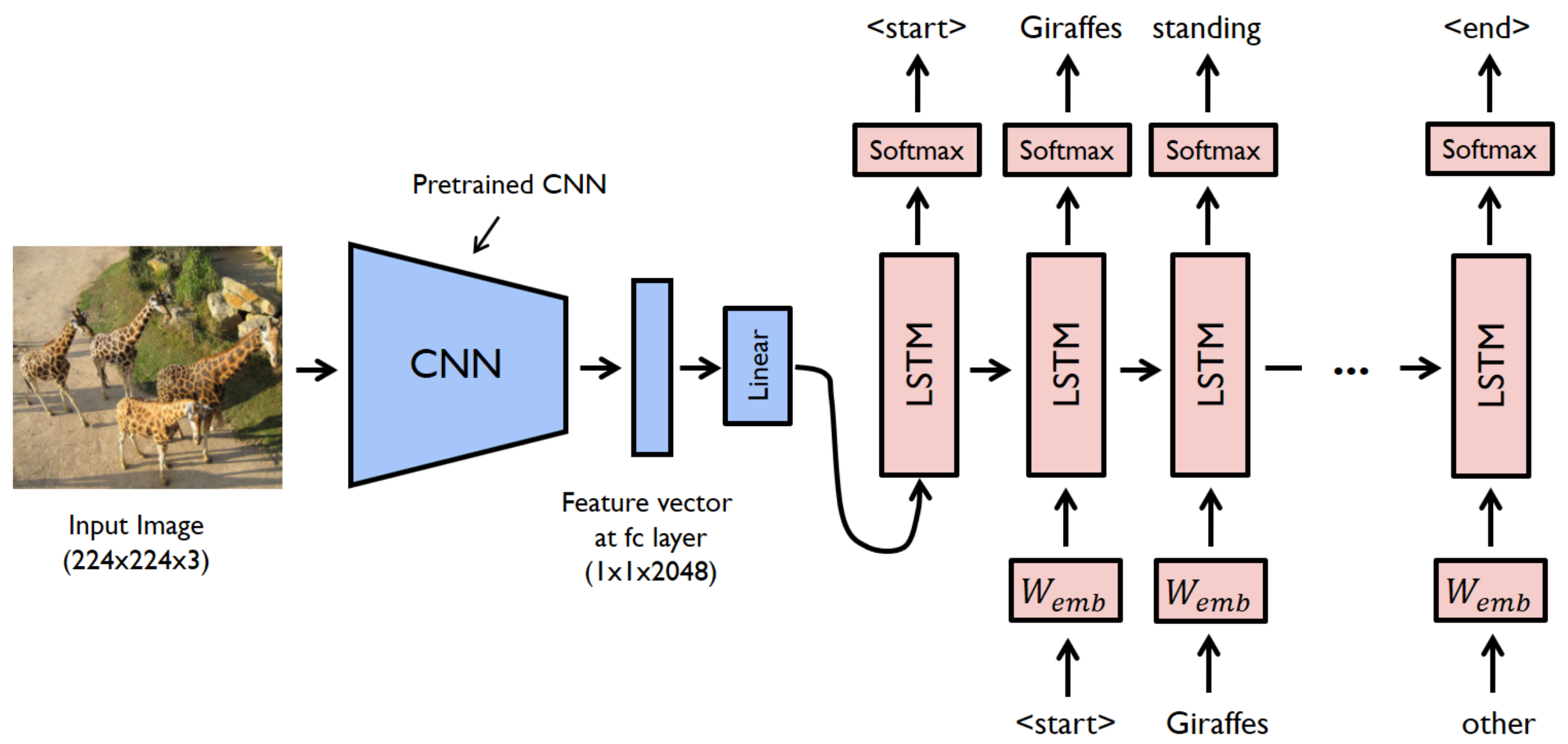


Figure 1: Architecture.

The architecture is designed as follows: the pretrained neural network InceptionV3 is used as a encoder and performs image convolution. The input transformed by the encoder is fed to the decoder, which is the only trainable part of the our network and represents several numbers of LSTM layers and converts the convoluted vector to a text description.

During the training, the data bypasses the convolutional block and is fed directly to the decoder, so the encoder is not involved in the training process.

Dataset

We use MSCOCO dataset which contains of the images and their descriptions. Each image is embedded by 2048 unit-vector and provided with its text caption.

BLEU details

BLEU is computed using a couple of ngram modified precisions. Specifically,

$$\text{BLEU} = \text{BP} \cdot \exp\big(\sum_{n=1}^{N} w_n \log p_n\big)$$

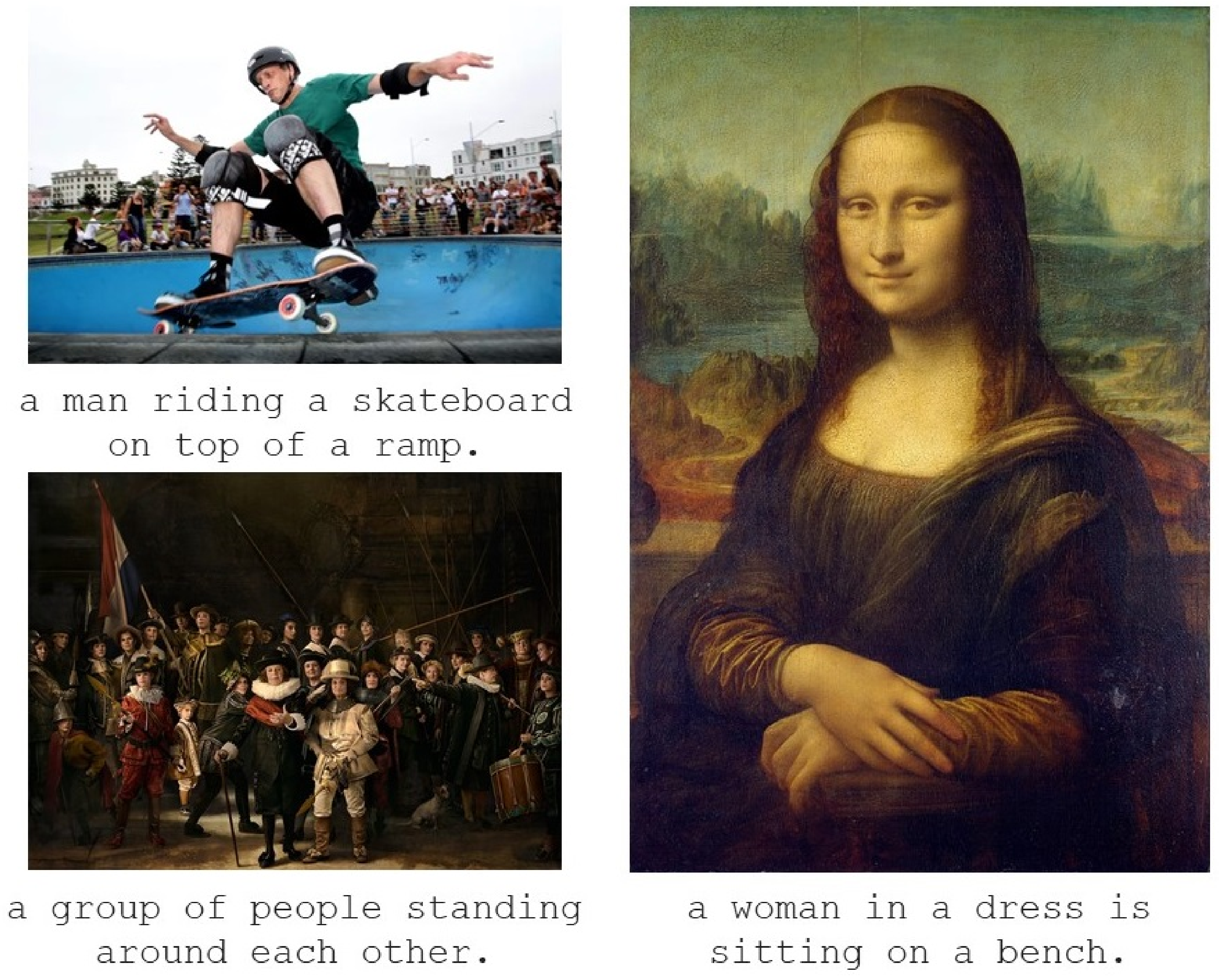
where p_n is the modified precision for ${}_n gram$, $\sum\limits_{n=1}^{N} w_n = 1$ and and BP is the brevity penalty to penalize short machine translations.

$$p_n = \frac{\sum\limits_{{}_n gram \in \hat{y}} Count_{clip}({}_n gram)}{\sum\limits_{{}_n gram \in \hat{y}} Count({}_n gram)}$$

$$\text{BP} = \begin{cases} 1, \ c > r \\ \exp\left(1 - \frac{r}{c}\right), \ c \leq r \end{cases}$$

where $Count_{clip}({}_n gram)$ is the maximum number of times that ${}_n gram$ appears in one of the references, $\hat{y}$ - machine translation candidate, c is the length of the candidate sentence and r is the best match lengths for each candidate sentence in the corpus.

Results


`a man riding a skateboard on top of a ramp.`

`a group of people standing around each other.`

`a woman in a dress is sitting on a bench.`

Figure 2: Examples.

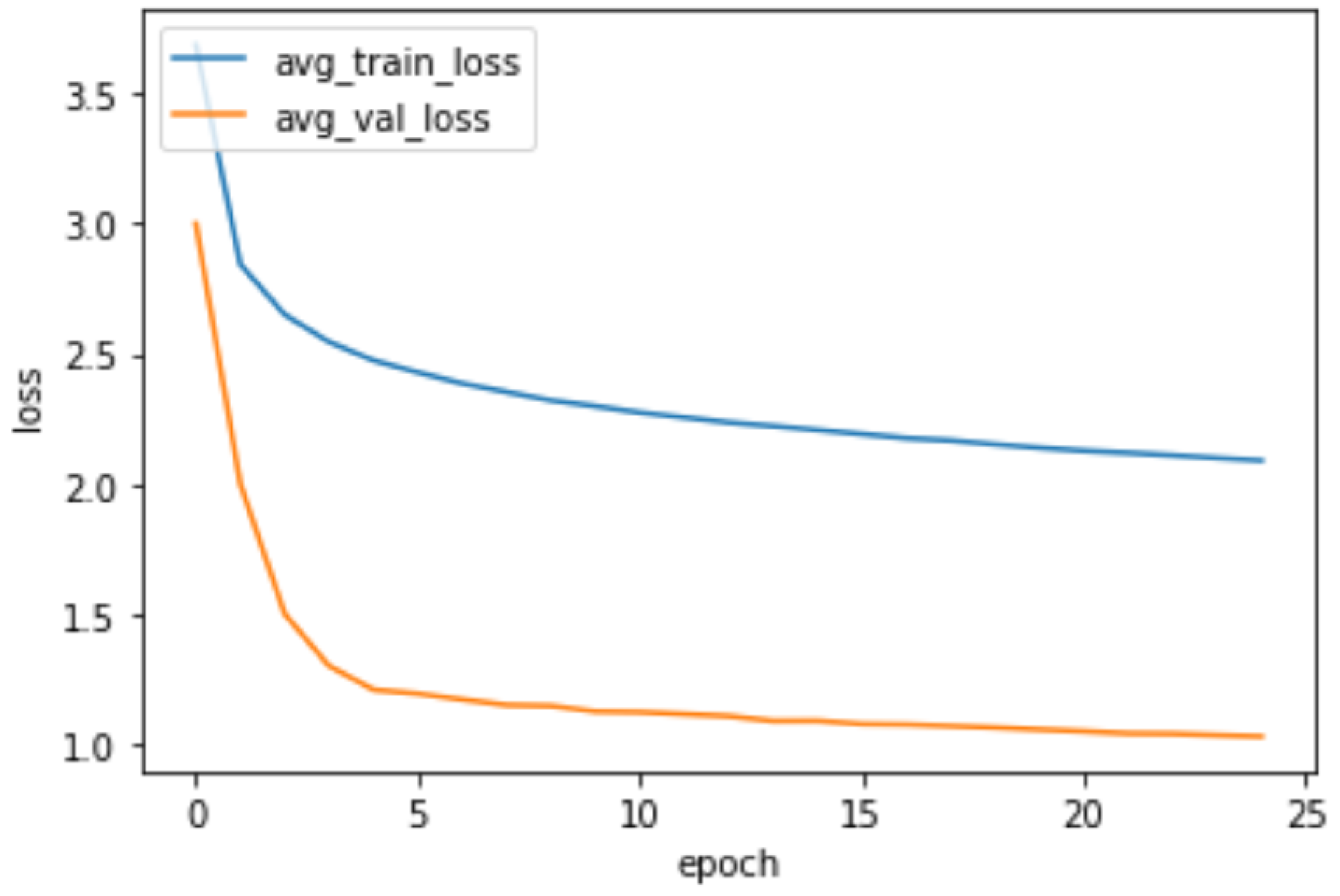


Figure 3: loss values.

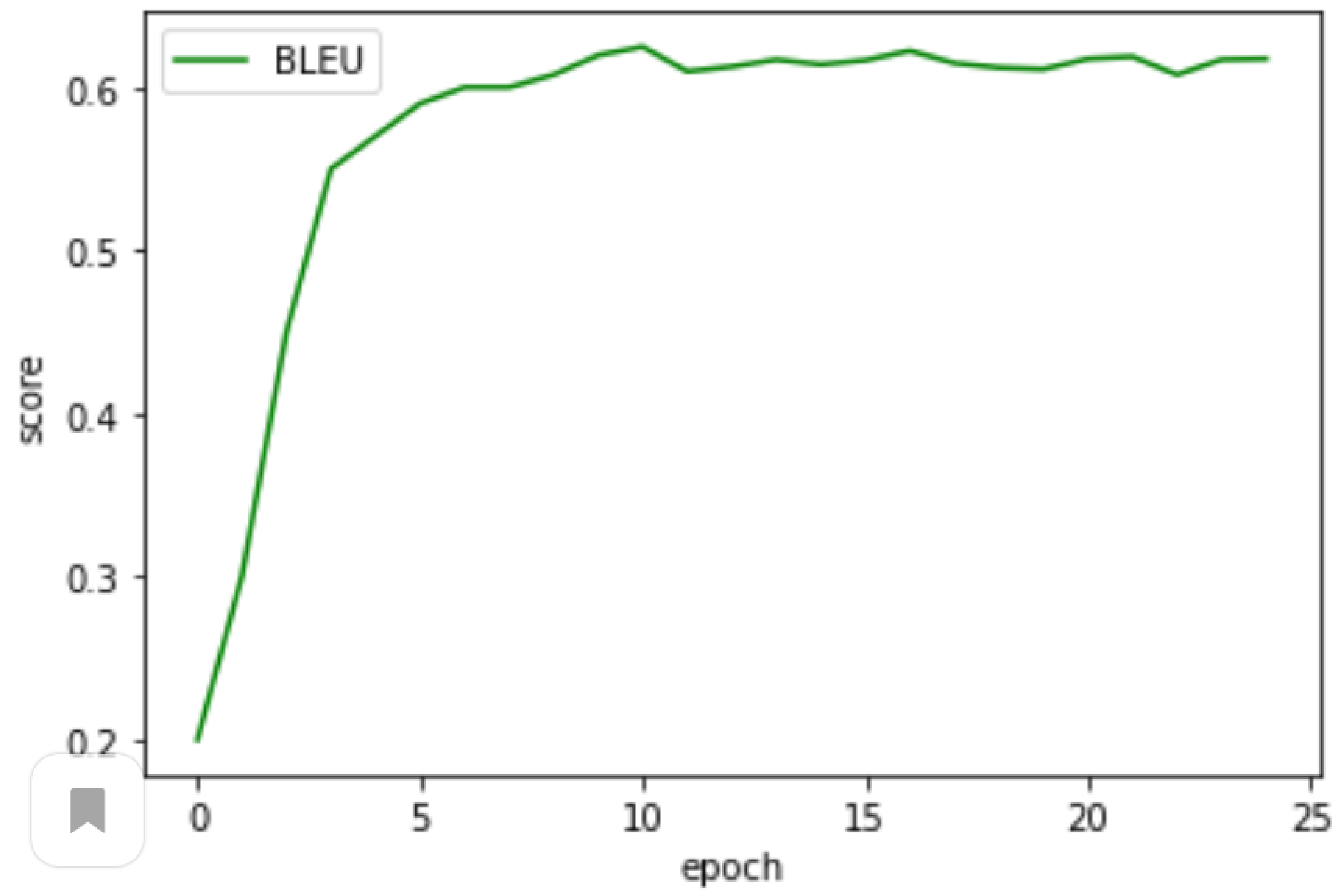


Figure 4: BLEU score.

Conclusion

We got an excellent result by writing a neural network from scratch. The BLEU score of 0.6 indicates that the neural network produces a description not by chance, has an acceptable quality and captures the essence.

Acknowledgements

This material is based upon work [1] and supported by the Daniil Merkulov.

References

[1] Vinyals, Oriol and Toshev, Alexander and Bengio, Samy and Erhan, Dumitru. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. IEEE transactions on pattern analysis and machine intelligence.