Szöveges leírás megadása képekről, neurális hálóval (2018/19 ősz)

Textual summarization about images using neural networks

Sághy Dániel, Markos Péter

Absztrakt - **A munkánk során arra a kérdésre kerestük a választ, hogy milyen lehetőségek vannak jelenleg a képből szöveg generálásra, illetve hogy a gépi tanulás hogyan illeszthető be ebbe a képbe. Megvizsgáltuk a jelenlegi legfrissebb kutatásokat, és néhány alapvető elméleti művet, majd az egyik leghasználhatóbbnak tűnő keretrendszert, a TensorFlow-t alkalmazva, megnéztük, hogy hogyan kaphatnánk minél pontosabb leírást egy képről a betanított neurális hálótól.**

Abstract - **Throughout our project, we were searching for opportunities and methods of how to generate images from text and how to combine these with machine learning. We examined the latest researches and publications and have chosen the framework which seemed the easiest to handle: We applied TensorFlow to see how to get as accurate description of an image as possible from the neural network.**

Index Terms - **Deep learning, neural networks, tensorflow, CNN encoder, RNN decoder, attention, Inception V3.**

Introduction

In the era of the increasing popularity of Deep-learning or Machine-learning, the possibility to generate textual descriptions based on images automatically and within only seconds is becoming more and more accessible. The results of the topic can be useful regarding several scientific or public issues, such as semantic visual searching, visual intelligence in chatbots, sharing photos and videos on social media, the automatic labeling of the previous examples, or a kind of support for those who have visual disorder. However, a common problem can be the fact that different people tend to give different descriptions of the same image.

The thoughts summarized above motivated us to deal with this topic in our project for this semester.

Introducing the topic

Former solutions

The ‘image to text’ is a field where the purpose is assigning descriptions to images. The point is to imagine without pictures, based on only descriptions, what the original input could have been. This can be approached from several views: First, we acknowledge again that based on a given description, different individuals would have slightly different pictures in their mind, as they have different backgrounds. However, if their task was to match three different images to three descriptions, it is likely that they would do this correctly. The ‘image to text’ method we have chosen is aiming this second task: Creating bijection between images and text.

Looking closer to our point of view in connection with neural networks and based on some background research, we found similar problems with the following solutions: In most of the resources, they used RNN Decoder to generate text and CNN Encoder to resolute images. In order to make descriptions more efficient, we read about (and also used later) a method called ‘Attention’ which helps identifying which part of the image is used to learn and predict the words of the description in the neural system.

Systemplan

Considering the field of our topic and the resources we found, we have chosen a model of ‘Inception V3’ as a network. Here the output layer is the last convolutional layer, because we are using attention in this example. The shape of the output of this layer is 8x8x2048. We extract the features from the lower convolutional layer.

type	**patch size/stride** or remarks	**input size**
conv	3×3/2	299×299×3
conv	3×3/1	149×149×32
conv padded	3×3/1	147×147×32
pool	3×3/2	147×147×64
conv	3×3/1	73×73×64
conv	3×3/2	71×71×80
conv	3×3/1	35×35×192
3×Inception	As in figure 5	35×35×288
5×Inception	As in figure 6	17×17×768
2×Inception	As in figure 7	8×8×1280
pool	8 × 8	8 × 8 × 2048
linear	logits	1 × 1 × 2048
softmax	classifier	1 × 1 × 1000

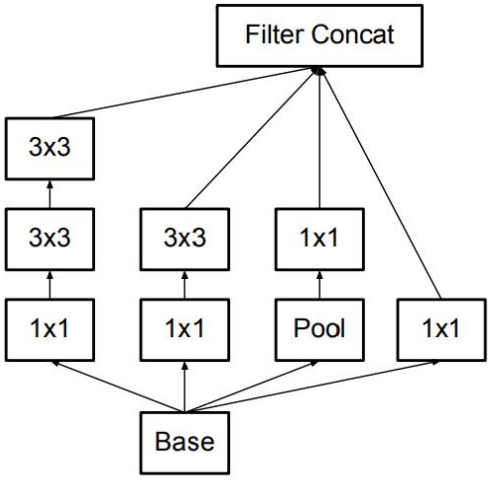


figure 5

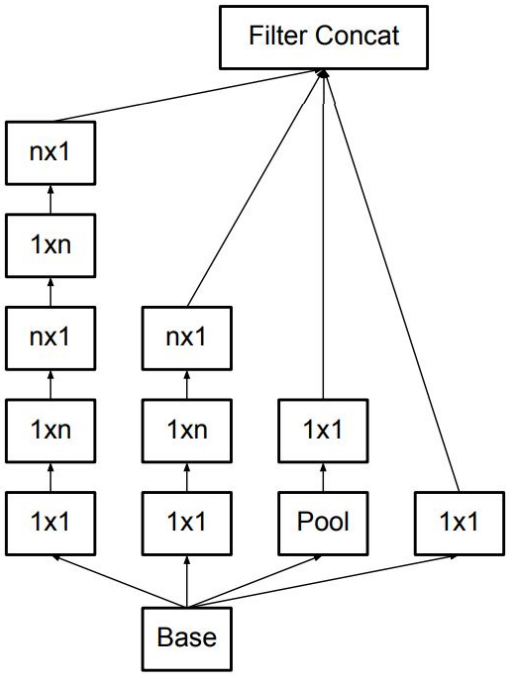


figure 6

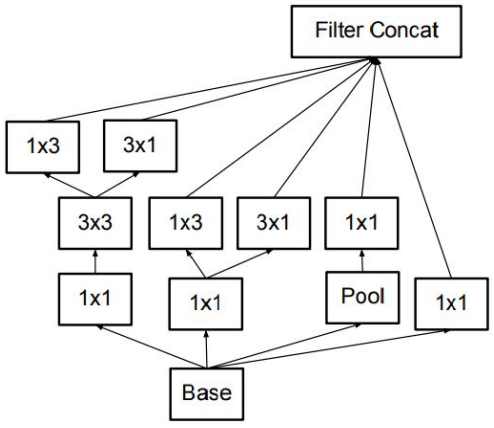
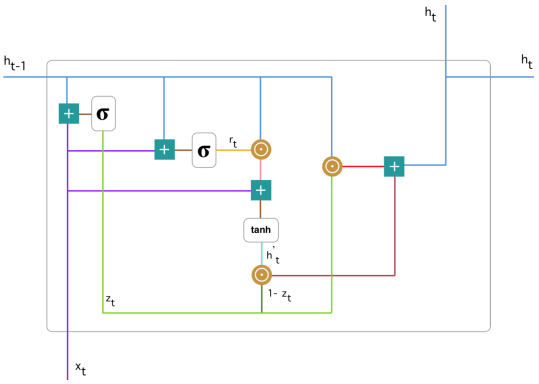


figure 7

Thus, for processing images, we used a fully connected layer, and we applied recurrent layers to decoding and predicting purposes. In our case, this recurrent layer is GRU which is also used to solve the vanishing gradient problem which comes with a standard recurrent neural network.



GRU

We use Soft Attention at the output layer which highlights the important parts appearing in the pictures.

Realization

I. Data acquisition and preparation

Acquiring proper data and preparing it to be able to teach is one of the critical points of machine teaching.

We used Google's Conceptual Captions-t (https://ai.google.com/research/ConceptualCaptions) as a database. The dataset consists of approximately 3,3 millions of data in .tsv files where the description and the links to the pictures can be found. We downloaded the images with a Python script and named the files according to their order of their occurence. Simultaneously, we assigned their descriptions to the pictures in a .txt file.

Next, we generated JSON files with another Python script which then we prepared for the neural network with InceptionV3.

II. Teaching

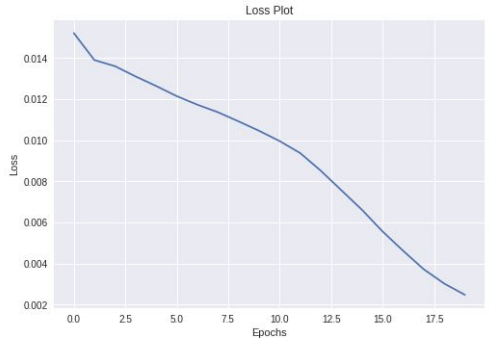
Throughout the teaching process, we used Adam optimization and softmax cross entropy to decrease the probability of an occuring mistake. Based on our researches, these seemed the most efficient optimizing formulas for our purposes.

The available quantity of data is quite low because of personal technical limitations. Moreover, the generation of numpy files were failed in case of several images, so it also reduced the number of data. Therefore, during the teaching with 64 batch-size we had many unsuccessful batches.

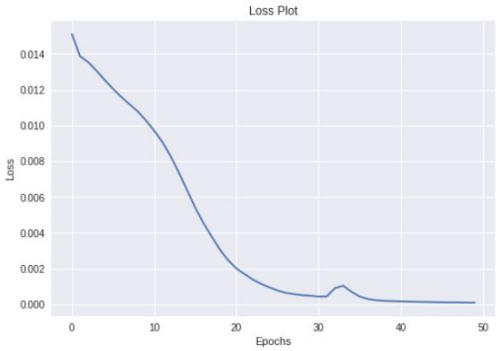
III. Evaluation

The evaluation of data and our results is not an easy task in any part of deep learning. For the reasons, mentioned earlier above, the results are relative. Because of the short amount of time in the semester, the different predictions created by the neural network were examined based on our own judgement. We concentrated on the way of connecting the descriptions to the pictures (attention) and on the extent of clearness. Throughout the whole process we counted the occurrences of loss.

We got to the surprising conclusion that from only a few thousands of data and running only 20 epochs, our neural network has learnt very effectively. Consequently, the measure of loss decreased significantly as well. We tested the teaching with 50 epochs, as well and we got these results. In the last 5 epochs, the loss is not really changed at all.



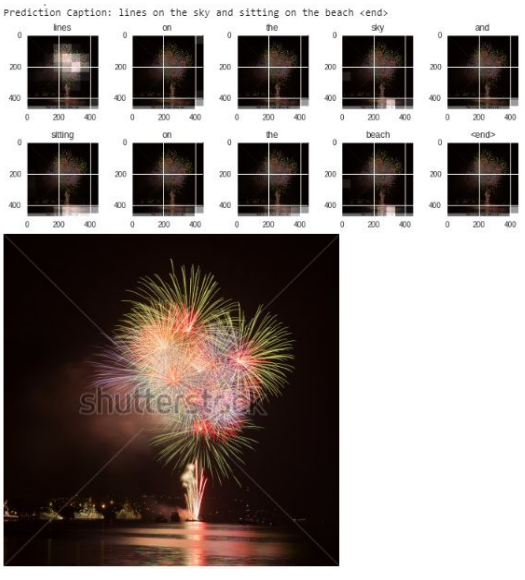
20 epoch



50 epoch

IV. Testing

The given prediction by the neural network was "*Lines on the sky and sitting on the beach*." Although it is not an undoubtedly accurate definition, it can be quite clearly identified. Attention helps us to see even more clearly that, for example, "lines" are made up by the lights of the firework and "beach" is probably from the lake at the bottom of the image.



Future plans, conclusion

It would be worthwhile to make further research on the topic and study it thoroughly. It would also be useful to teach our network through more epochs. Based on our present results, the basement of the network seems appropriate. The description was a bit too long sometimes, yet there was always a clear compliance between one image and its description. However, a database in which there would be more, rather short, definition to each image could be even more efficient as it would facilitate more reliable predictions. Nevertheless, in this case the preparation of data would take more time and more careful implementation.

We agree that making the method faster and more accurate could play a significant role in the future, whether we consider the field of social media, or we look at industry. Also, in the system of self-driving cars, throughout continuous image analyzation it can improve the connection with the passenger (eg. warnings). It is only a small step from this point that we can transform the written definitions into sounds as well which would lead to a compact audio-visual experience, while also facilitating more careful driving.

Although we are satisfied with our results, we acknowledge that there are undoubtedly remaining challenges and opportunities for further development in the project, which require higher computing power and better optimizing methods.

References

[1] C. Szegedy, V. Vanhoucke, S. Ioffe and J. Shlens, "Rethinking the Inception Architecture for Computer Vision", *Cv-foundation.org*, 2018. [Online]. Available: https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Szegedy_Rethinking_the_Inception_CVPR_2016_paper.pdf. [Accessed: 30- Nov- 2018].

[2] "Pretrained Inception-v3 convolutional neural network - MATLAB inceptionv3", *Mathworks.com*, 2018. [Online]. Available: https://www.mathworks.com/help/deeplearning/ref/inceptionv3.html;jsessionid=59e0385ffeaa9b18c15589978ab9. [Accessed: 04- Oct- 2018].

[3] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, "Show and Tell: A Neural Image Caption Generator", *Cv-foundation.org*, 2018. [Online]. Available: https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Vinyals_Show_and_Tell_2015_CVPR_paper.pdf. [Accessed: 13- Nov- 2018].

[4] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho and Y. Bengio, "Attention-Based Models for Speech Recognition", *Papers.nips.cc*, 2018. [Online]. Available: http://papers.nips.cc/paper/5847-attention-based-models-for-speech-recognition.pdf. [Accessed: 25- Nov- 2018].

[5] Z. Yang, X. He, J. Gao, L. Deng and A. Smola, "Stacked Attention Networks for Image Question Answering", *Cv-foundation.org*, 2018. [Online]. Available: https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Yang_Stacked_Attention_Networks_CVPR_2016_paper.pdf. [Accessed: 30- Oct- 2018].

[6] S. Kostadinov, "Understanding GRU networks – Towards Data Science", *Towards Data Science*, 2017. [Online]. Available: https://towardsdatascience.com/understanding-gru-networks-2ef37df6c9be. [Accessed: 17- Oct- 2018].

[7] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel and Y. Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention", *Arxiv.org*, 2016. [Online]. Available: https://arxiv.org/abs/1502.03044. [Accessed: 22- Nov- 2018].

[8] X. He and L. Deng, "Deep Learning for Image-to-Text Generation: A Technical Overview - IEEE Journals & Magazine", *Ieeexplore.ieee.org*, 2017. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8103169&tag=1. [Accessed: 30- Nov- 2018].

[9] D. P. Kingma and J. Lei Ba, "ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION", *Arxiv.org*, 2015. [Online]. Available: https://arxiv.org/pdf/1412.6980.pdf. [Accessed: 28- Oct- 2018].

[10] Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C. and Xu, W. (2018). *CNN-RNN: A Unified Framework for Multi-label Image Classification*. [online] Cv-foundation.org. Available at: https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Wang_CNN-RNN_A_Unified_CVPR_2016_paper.pdf [Accessed 3 Dec. 2018].