

## PROJECT PROPOSAL

### Image captions generation

Mikhail Ershov

In this project, we try to use a generative model based on a deep recurrent architecture that combines recent advances in computer vision and machine translation and that can be used to generate natural sentences describing an image. The model is trained to maximize the likelihood of the target description sentence given the training image. Experiments on several datasets show the accuracy of the model and the fluency of the language it learns solely from image descriptions. Our model is often quite accurate, which we verify both qualitatively and quantitatively. For instance, while the current state-of-the-art BLEU-1 score (the higher the better) on the Pascal dataset is 25, our approach yields 59, to be compared to human performance around 69. We also show BLEU-1 score improvements on Flickr30k, from 56 to 66, and on SBU, from 19 to 28. Lastly, on the newly released COCO dataset, we achieve a BLEU-4 of 27.7, which is the current state-of-the-art.

### Introduction

Being able to automatically describe the content of an image using properly formed English sentences is a very challenging task, but it could have great impact, for instance by helping visually impaired people better understand the content of images on the web.

### Project Goals

1. Try to reproduce image encoding NN model.
2. Train the model on different datasets.
3. Measure results on different metrics, suggested in paper (starting from BLEU score metric, and diving into what other metrics they used)