SSY340 Planning report - Group 35

Pasquale Bianco, Jacob Billvén, Mattia Carlino, Francesca Zambetti

October 9, 2024

Project

We plan to do image captioning using our own RNN fed with the feature vector coming from a Convolution Neural Network. For this purpose we will evaluate pros and cons about different CNN: VGG16, ResNet and GoogleNet. Based on these features our RNN captioner will output a caption character by character to generate its text. The exact design for our RNN architecture has to be evaluated further (and is what this project mostly is about), but GRUs are a viable option for character by character text generation. Also, LSTMs provide a good ground for RNNs and have to be considered too. To compare the performance of our model we will also employ Microsoft GIT (GenerativeImage2Text) [1] to see how our "homemade" model sticks up to the larger models. Moreover, the employment of word based output instead of character based is of interest and should be explored to see if it gives easier implementation.

Since we are building a custom model with lower computational power compared to most state-of-the-art image captioners, we expect it to perform reasonably well, although not exceeding the benchmarks set by models like Microsoft GIT. Nonetheless, this will provide a valuable opportunity to identify the strengths and weaknesses of our approach, allowing us to understand which aspects the model handles effectively and which areas still need improvement.

Datasets

We plan to use Flickr30k [2] dataset containing 30 thousand images with 5 different captions per image (making 150k samples for our purpose). The captions are provided by human annotators and contain complete sentences as well as more ragged descriptions.

Evaluation

Evaluating this type of task is quite challenging, so we will apply the following heuristics and select the one that yields the best final performance. Additionally, we plan to test the idea of combining these heuristics to potentially achieve better results on the validation dataset:

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation): can help evaluate how well captions capture essential information from the images.
- CIDEr (Consensus-based Image Description Evaluation): it is specifically designed for image captioning and evaluates the quality of generated captions based on their consensus with multiple human annotations.

Time plan

- `week 6`:
  - preparation and processing of the datasets
  - find new dataset for further evaluation
- `week 7`:
  - building and training our model;
  - training and tuning the benchmark model;
- `week 8`:
  - results evaluation;
  - preparation of final report and poster.

Relevant papers and available code

- paper – VGG Image Caption [3]: example of usage of the dataset to do image captioning employing KANs+LSTMs
- code – GIT: A Generative Image-to-text Transformer for Vision and Language [1]: paper related to the benchmark model we will employ for performance comparison
- code – Image captioning with visual attention [4]: tutorial code to build a model based on self-attention for image captioning

References

[1] J. Wang, Z. Yang, X. Hu, *et al.*, "Git: A generative image-to-text transformer for vision and language," *arXiv preprint arXiv:2205.14100*, 2022.

[2] *Flickr 30k dataset.* [Online]. Available: `https://www.kaggle.com/datasets/adityajn105/flickr30k`.

[3] *Vgg image caption.* [Online]. Available: `https://www.kaggle.com/code/pes1ug22am124/vgg-image-caption`.

[4] *Image captioning with visual attention.* [Online]. Available: `https://www.tensorflow.org/text/tutorials/image_captioning`.