CS6370: Natural Language Processing
Project Proposal

Using Deep Neural Networks for Image Captioning

R Pooja, Yogesh Tripathi
ch15b056@smail.iitm.ac.in, cs16b044@smail.iitm.ac.in

Indian Institute of Technology, Madras

1 Problem Statement

Generating natural language captions for images is at the intersection of Computer Vision, Natural Language Processing and Artificial Intelligence. This task has a number of interesting applications such as semantic visual search, visual intelligence in chatting robots, photo and video sharing in social media, and aid for visually impaired people to perceive surrounding visual content. Deep convolutional nets have brought about breakthroughs in processing images, video, speech and audio, and recurrent nets work well when processing sequential data such as text and speech. In our project, we attempt to provide syntactically correct and semantically meaningful caption to raw image data using an end-to-end framework. We implement the ideas for the image captioning task discussed in '*Deep Learning for Image-to-Text Generation*' by Xiaodong He and Li Deng [1].

2 Proposed Methodology

The challenges in the task is detecting detecting salient semantic concepts in the image and understanding the relationships between them, and composing a coherent description about the overall content of the image. First, we use an encoder-decoder framework for image captioning. A raw image is first encoded by a global visual feature vector which represents the overall semantic information of the image using deep convolutional neural networks(CNN). We use the AlexNet architecture for CNN because it has been very successful in large scale image classification. The activation values in second last fully connected layer is extracted as the global visual feature vector. After this extraction, it is fed into a recurrent neural network(RNN) based decoder for caption generation. We use a long-short memory network(LSTM) or gated recurrent unit(GRU) variation of RNN as they are more efficient in capturing long-span language dependencies than vanilla RNNs.

A second approach is using attention mechanism. Different from the simple encoder-decoder approach, the attention-based approach first uses the CNN to not only generate a global visual feature vector but also generate a set of visual vectors for subregions in the image. These subregion vectors can be extracted from a lower convolutional layer in the CNN. Then, in language generation, at each step of generating a new word, the RNN will refer to these subregion vectors and determine the likelihood that each of the subregions is relevant to the current state to generate the word. Eventually, the attention mechanism will form a contextual vector, which is a sum of subregional visual vectors weighted by the likelihood of relevance, for the RNN to decode the next new word.

3 Datasets

A very rich set of data is available for image captioning task. The following are the sources of data sets, some of which we plan to use in evaluating our model:

- Common Objects in Context(COCO) data set
- Flickr data set
- PASCAL sentence data set

4 Evaluation Measures

We propose the use of one or more of the following 4 automatic metrics to evaluate our models:

- BLEU[2]: Widely used in machine translation, measures the fraction of n-grams that are common to the output produced and reference texts.
- METEOR[3]: Measures precision and recall of unigrams, but does not just consider exact word matches but also similar words from WordNet synonyms of stemmed tokens.
- CIDEr[4]: It is the cosine similarity of n-gram matches (post stemming to root forms) between the proposed caption and reference text weighted by TF-IDF computed for all n-grams across all references put together.
- SPICE[5]: Overcomes limitations of n-gram based metrics by capturing semantic propositional content. Both candidate and reference captions are encoded into a graph-based semantic representation called a scene graph. The scene graph explicitly encodes the objects, attributes and relationships found in image captions, abstracting away most of the lexical and syntactic phrases. Caption quality is determined using an F-score calculated over tuples in the candidate and reference scene graphs.

5 References

[1] Xiaodong, HeLi Deng *Deep Learning for Image-to-Text Generation: A Technical Overview*

[2] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. *Bleu: a method for automatic evaluation of machine translation.* Proc. 40th Annu. Meeting Association Computational Linguistics, 2002, pp. 311–318

[3] S. Banerjee and A. Lavie. *METEOR: An automatic metric for MT evaluation with improved correlation with human judgments.* Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005.

[4] R. Vedantam, L. Zitnick, and D. Parikh *CIDEr: Consensus-based image description evaluation* Proc. European Conf. Computer Vision, 2015, pp. 4566–4575

[5] P. Anderson, B. Fernando, M. Johnson, and S. Gould. *SPICE: Semantic propositional image caption evaluation* Proc. European Conf. Computer Vision, 2016