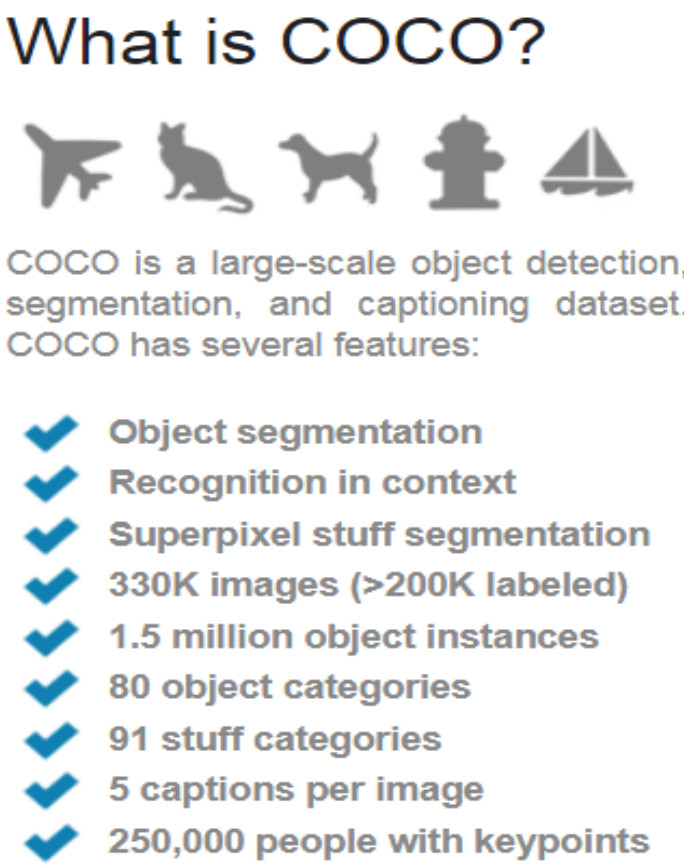
Image Captioning

Introduction

- Image captioning is a popular research area of Artificial Intelligence (AI) that deals with image understanding and a language description for that image. Image understanding needs to detect and recognize objects. It also needs to understand scene type or location, object properties and their interactions. Generating well-formed sentences requires both syntactic and semantic understanding of the language.
- Image captioning is important for many reasons. For example, they can be used for automatic image indexing. Image indexing is important for Content-Based Image Retrieval (CBIR) and therefore, it can be applied to many areas, including biomedicine, commerce, the military, education, digital libraries, and web searching. Social media platforms such as Facebook and Twitter can directly generate descriptions from images.
- Understanding an image largely depends on obtaining image features. The techniques used for this purpose can be broadly divided into two categories: (1) Traditional machine learning based techniques and (2) Deep machine learning based techniques.
- In this project we will use a Deep learning based technique using Encoder-Decoder Architecture.

Data

- **MS COCO Dataset.**
- url: http://cocodataset.org/#download



Approach - Encoder-Decoder Architecture-Based Image captioning

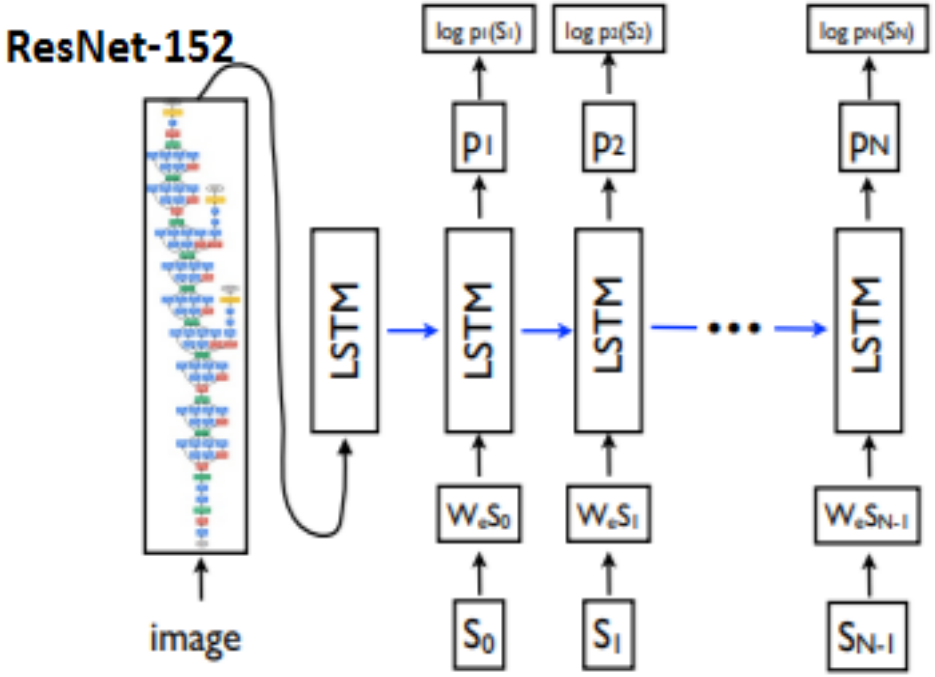
- We use a Deep learning based technique using Encoder-Decoder Architecture. The neural network-based image captioning methods are very similar to the encoder-decoder framework-based neural machine translation. In this network, global image features are extracted from the hidden activations of CNN and then fed them into an LSTM to generate a sequence of words.

A typical method of this category has the following general steps:

- A vanilla CNN is used to obtain the scene type, to detect the objects and their relationships.
- The output of Step 1 is used by a language model to convert them into words, combined phrases that produce an image captions.

Approach - Architecture

CNN-RNN (EncoderCNN-DecoderRNN) model is based on the model proposed in the paper "Show and Tell: A Neural Image Caption Generator" (https://arxiv.org/pdf/1411.4555.pdf). Below figure adapted from the paper shows a LSTM model combined with a CNN (ResNet152) image embedder and word embeddings. The unrolled connections between the LSTM memories are in blue and they correspond to the recurrent connections. All LSTMs share the same parameters.



Approach: Evaluation Metrics

BLEU (Bilingual evaluation understudy) is a metric that is used to measure the quality of machine generated text. Individual text segments are compared with a set of reference texts and scores are computed for each of them. In estimating the overall quality of the generated text, the computed scores are averaged. However, syntactical correctness is not considered here. The performance of the BLEU metric is varied depending on the number of reference translations and the size of the generated text. BLEU is popular because it is a pioneer in automatic evaluation of machine translated text and has a reasonable correlation with human judgements of quality. However, it has a few limitations such as BLEU scores are good only if the generated text is short.

A perfect match results in a score of 1.0, whereas a perfect mismatch results in a score of 0.0.

Evaluation Metrics - BLEU Score

```
Evaluation performance on MS-COCO validation
dataset

BLEU-1: 0.6508
BLEU-2: 0.4527
BLEU-3: 0.3099
BLEU-4: 0.2145
```

To evaluate the skill of the model using BLEU scores:
For reference, below are some ball-park BLEU scores for skillful models when evaluated on the test dataset (taken from the 2017 paper - Where to put the Image in an Image Caption Generator) :

BLEU-1: 0.401 to 0.578
BLEU-2: 0.176 to 0.390
BLEU-3: 0.099 to 0.260
BLEU-4: 0.059 to 0.170

Sample Image/Caption

A group of people walking down a street.

Conclusions and Future Work

- In this Image Captioning project, we implemented a Deep learning-based technique using Encoder-Decoder Architecture. We have shown our calculated Evaluation Metric - BLEU scores, outperformed the reference BLEU scores (Ball-park BLEU scores for skillful models when evaluated on the test dataset, taken from the 2017 paper - Where to put the Image in an Image Caption Generator)
- In their paper - BLEU: a Method for Automatic Evaluation of Machine Translation, the authors quote -
- The BLEU metric ranges from 0 to 1. Few translations will attain a score of 1 unless they are identical to a reference translation. For this reason, even a human translator will not necessarily score 1. […] on a test corpus of about 500 sentences (40 general news stories), a human translator scored 0.3468 against four references and scored 0.2571 against two references.

Conclusions and Future Work

Future improvements to investigate further:

- 1) Beam Search as an approximate search (often works better than the greedy approach)

- 2) Implement other models like attention based model ([7] - Show, Attend and Tell: Neural Image Caption Generation with Visual Attention), which introduces a visual attention mechanism based on the Encoder-Decoder structure, which can dynamically focus on the salient regions of the image during the process of generating descriptions in Decoder.

- 3) Researchers have also proposed various efficient improvement methods, but they have different focuses.

Conclusions and Future Work

Improvements in Encoder: [10] You et al., proposes a semantic attention model, in addition to using CNN's intermediate activation output as the global feature of the image v, and also using a set of attribute detectors to extract {Ai} the most likely to appear in the image.

[11] Fu et al., introduced advanced semantic information to improve image description based on attention.

[12] Yao et al., believes that the semantic relationship and spatial relationship between image objects are helpful for image description generation.

Improvements in Decoder: [13] Lu et al., believes that in the process of generating image description, visual attention should not be added to nonvisual words such as prepositions and quantifiers.

[14] Zhou et al., pointed out that in previous work, image features are only initially fed into LSTM, or on the basis of which attention mechanism is introduced to compute context vectors to input LSTM. Whether text context could be used to improve image description performance has not been solved yet, that is, the relationship between generated words and visual information was not involved. To explore this problem, they proposed a Text-Conditional attention mechanism, which allows attention to focus on image features related to previously generated words

Recommendations to the Client

Facebook can use Image captioning functionality to automatically generate captions for photos in the News Feed of people who can't see them. This can be used with text-to-speech engines that allow blind people to use Facebook in other ways

Image based web content can be made more accessible to the users by using Image Captioning to provide descriptions of the Images. Information from the visual content found in the image can be used to further analyze the image with tagging, domain-specific models, and descriptions in other languages.

For example: Recognize brands, celebrities and landmarks

Image Captioning can be used to describe Videos in real time.

References

[1] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr DollÃąr, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In European conference on computer vision. Springer, 740–755.

[2] Qi Wu, Chunhua Shen, Anton van den Hengel, Lingqiao Liu, and Anthony Dick. 2015. Image captioning with an intermediate attributes layer. arXiv preprint arXiv:1506.01144 (2015)

[3] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition. 3156–3164.

[4] Hamid Laga ...et al. A Comprehensive Survey of Deep Learning for Image Captioning, 2018 (https://arxiv.org/abs/1810.04020)

[5] Kishore Papineni, et al. BLEU: a Method for Automatic Evaluation of Machine Translation, 2002.

[6] Alfredo Canziani & Eugenio Culurciello, Adam Paszke,An Analysis of Deep Neural Network Models for Practical Applications. 2018.

[7] Kelvin Xu, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

[8] Diederik P. Kingma, Jimmy Ba. Adam: A Method for Stochastic Optimization

[9] Marc Tanti, et al. Where to put the Image in an Image Caption Generator

References

[10] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In CVPR, pages 4651–4659, 2016.

[11] K. Fu, J. Jin, R. Cui, F. Sha, and C. Zhang. Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts. IEEE Trans. Pattern Anal. Mach. Intell., 39(12):2321– 2334, 2017.

[12] T. Yao, Y. Pan, Y. Li, and T. Mei. Exploring visual relationship for image captioning. In ECCV, pages 711–727, 2018.

[13] J. Lu, C. Xiong, D. Parikh, and R. Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In CVPR, pages 3242–3250, 2017.

[14] L. Zhou, C. Xu, P. A. Koch, and J. J. Corso. Watch what you just said: Image captioning with text-conditional attention. In Proceedings of the on Thematic Workshops of ACM Multimedia, pages 305–313, 2017.

[15] Yiyu Wang, Jungang Xu, Yingfei Sun and Ben He,. Image Captioning based on Deep Learning Methods: A Survey