News Image Caption

Fuxiao Liu*
fl3es@virginia.edu
University of Virginia

Dexuan Zhang*
dz5se@virginia.edu
University of Virginia

KEYWORDS

CNN, RNN, News, Entity, Embedding Feature, Attention.

ACM Reference Format:
Fuxiao Liu and Dexuan Zhang. 2020. News Image Caption. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In computer vision community, given images, image classification and object detection are two largely explored tasks. In addition, machine translation, given a source sentence translating into an output sentence in desired languages, is a focused topic in natural language processing research.

Generally, we want to be able to generate English description for given images, not only detecting objects contained in images but also portraying interaction between objects and events shown by images. What's more, we want our output captions are more human-like sentences instead of just rigid templates. To do that, we will follow the encoder-decoder structural used in NLP machine translation[2]. While keeping recurrent neural network(RNN) as our language decoder, we use convolution neural network(CNN), commonly used method in CV tasks, as the image encoder by extracting the fixed-length vector representation from the last hidden layer of a pre-trained image classification model. One example of our task shows in Figure 1.

2 DATASET

Our work is inspired by Vinyals et al.'s work[8]. The work has only been done with some rather less narrative datasets such as Pascal dataset or COCO dataset. And some previous works can at best perform at the description level but they can't integrate prior world knowledge, like News caption (Figure 1) including real person, locations, date and so on. So, what we want to do is to achieve better performance on News datasets. We collected the datasets from USA TODAY, WashingtonPost and BBC. Each data has one image then its corresponding caption and article. In our experiments, these articles are only utilized as post-processing for entity inserting.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

- **Raw**: Serena Williams celebrates her win over Angelique Kerber
- **Template gt**: PERSON_ celebrates her win over ORG_.

Figure 1: News Caption

3 ARCHITECTURE

3.1 Strategy

As we known, huge vocabulary size problems are very common in machine translation field as well as image caption. In order to solve this kind of problems, we always set a threshold like 3 or 4 and the words with frequency less than this threshold will be replaced as <unk>. This strategy is very efficient in coco-liked dataset but may not for news dataset. This is because in news dataset, many less-frequency words like Obama, Trump, White House are very important in news. Without these words, a news would miss its core information. So in order to not lose this important information, we utilized Name Entity recognizer (NER) to replace these less-frequency words with 18 name entity tags including PERSON, ORG, DATE, GPE, LOC, FAC and so on. The name entity recognizer here we used is spaCy tool. After this pre-processing, we have two groundtruth: raw groundtruth and template. One example is shown in Figure 1. Our strategy is to produce template caption in the first step. Then in second step, tags in templates will be replaced with real name entities from corresponding articles. Therefore, the final output captions can be improved in this two steps individually.

3.2 Model

Our baseline model is popular encoder-decoder architecture, which has been widely used in semantic segmentation, machine translation and image caption. In order to save training time we just replaced the encoder with a pretrained Resnet18 and then the decoder is a LSTM, which takes the CNN output feature as input and then predict the caption. To be specific, the CNN output feature is the last layer before the softmax layer in Resnet18, which originally aims to classify images.

Figure 2 is just a basic model and it's easy to see that the connection between encoder and decoder is so weak since there is only one link between them. As a result, when predicting the new word, the LSTM may not be able to find its focused features from images. For instance, when predicting the new word "dog", it should focused on the area which has dog in the image. So, based on the baseline, we add visual attention mechanism between encoder and decoder to enhance their connection.

Then for the name entity inserting part,each image has a corresponding article, which tells details about its caption. In other word, captions are summarized sentences of the articles so we can get the name entities from the corresponding article. So, we design two methods: Random Insert and Sort Insert. Random insertion means that we randomly select a name entity of certain tags in articles. As you can see, this method may not perform well. So another method called Sort insertion, which means we first sort entities of every tags according to their frequency and then select the most frequent one.

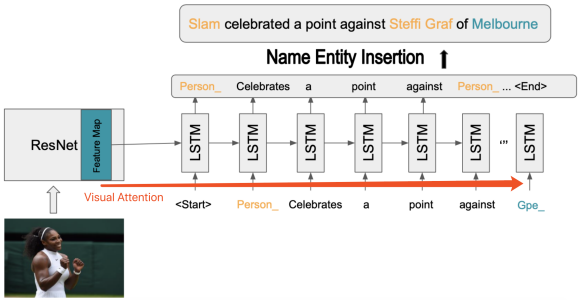

Figure 2: Our Baseline Model

4 EXPERIMENTS

4.1 Preprocessing

Our datasets come from three sources so the according captions are stored in three different .json files, every data in which has one caption, its image id, article id and caption length. Then our preprocessing includes two parts: image preprocessing and caption preprocessing. For image parts, we deleted repeated images, and ones which have open problems and so on. Then as we known, tiny images may not have much information so we only kept the images with both length and width larger than 180. As for news caption operations, we first filtered html labels, brackets, un-ASCII characters and some special tokens. What's more, we designed our own stop words based nltk and then filtered them. Apart from this, we only kept the captions of length between 3 and 30. This is simply because if the caption is too short, it tells nothing while if it's too long, it's much difficult for LSTM to learn and predict.

4.2 Implemention Details

After preprocessing, we divided each of our three datasets into training, validation and testing parts by 100000, 10000 and 10000. Our model was trained with bath size 128 and learning rate 0.0001 for 50 epochs. What's more, we used simple technique to train the model. First, the hyperparameter finetune is false, meaning that only the LSTM is training during this period. This period is stopped when the CIDer score is no longer increased. Then we reset the finetune value as true to finetune both the CNN and LSTM parts. The training process is shown in Figure 3. We want to evaluate our model with BLEU[5], METEOR[3], CIDEr[7] and Rouge-L. These evaluation metrics are implemented with different focus, correlation on the sentence level, heavily weighted recall, and downweights in n-gram common respectfully.

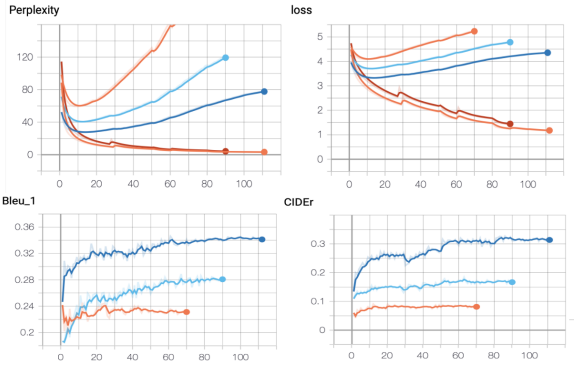

Figure 3: Loss, Perplexity, BLEU1, CIDer

4.3 Experiment Results

First of all, we show the vocabulary statistic for three datasets. We choose threshold as 4 to build our vocabulary and frequency less than 4 will turn into <unk> token.

Threshold	>=0	>=4	>=6	>=10
USA	61127	20669	15959	-
WashingtonPost	65821	20706	15633	-
BBC	60703	19562	14752	-

As you can see, there are a large amount of words which occur just once or twice. So if we just ignore them all and treat them as <unk> tokens, the predicted caption must lose many important information. So how to deal with these out-of-vocabulary words is a main point we can improve. Before that, let's see the performance of baseline model:

Dataset	CIDer	BLEU1	BLEU2	METEOR	ROUGE
USA	3.7%	9.8%	3.5%	2.8%	9.4%
Wash	2.5%	9.6%	3.1%	2.7%	8.8%
BBC	1.4%	8.3%	2.6%	2.1%	8.9%

From the above table, we can see the basic model did not perform well. So in order to deal with the out-of-vocabulary words, we try to reduce the whole vocabulary size and categorize the low frequency words or phrases.

4.4 Entity-Aware Captions

We first use SpaCy's name entity recognizer to recognize all the name entities in the original captions and switch them with their according tags like PERSON, ORGANIZATION, DATE and so on.The

new vocabulary are shown below. Obviously, the vocabulary reduced a lot.

Threshold	>=0	>=3	>=4	>=10
USA	38259	15680	12943	-
WashingtonPost	40011	16053	13258	-
BBC	38065	15948	13261	-

And then use the baseline CNN+LSTM to predict the Template. Finally, try two name-entity inserting skills (Random Insertion and Sort Insertion) to output the full captions. The improved scores are shown below. The first chart is Random Insert and Sort Insertion is the second one. Obviously, you can see Sort Insertion performs better than Random Insertion. Finally we visualize output captions of Figure 1.

Dataset	CIDer	BLEU1	BLEU2	METEOR	ROUGE
USA	14.8%	15.4%	6.6%	5.4%	13.3%
Wash	11.5%	13.5%	5.0%	4.5%	11.8%
BBC	14.1%	13.5%	4.8%	4.7%	12.2%

Dataset	CIDer	BLEU1	BLEU2	METEOR	ROUGE
USA	15.7%	15.5%	6.6%	5.6%	14.1%
Wash	12.7%	13.6%	4.9%	4.8%	12.6%
BBC	15.5%	14.3%	4.9%	5.1%	12.9%

- Slam celebrates her victory over Williams of Melbourne in the women singles match against Williams of Australia.
- Slam of New York celebrates a point against Steffi Graf of Melbourne.
- Williams of Melbourne reacts after winning a shot in the second round of the men golf tournament.

Figure 4: Output Caption Visualization

4.5 Visual Attention

Apart from the improvement in insertion part, we also trained a better model to improve template performance. Attention Mechanism is efficient not only in machine translation but also in computer vision field. In machine translation field, the translated word must have its corresponding part in original sentences instead of the whole sentence, which is the same in image caption. Each word corresponds certain part of the image. We alse tried several different methods including Show Tell(basic CNN + LSTM), Show Attend(Visual Attention), Up-down[1], Adaptive Att(Adaptive Attention)[4], Att2in2[6]. The results evaluated on USA TODAY are shown below. Clearly, it is the Visual Attention we select that performs best.

USA	CIDER	B_1	B_2	B_3	B_4	Meteor	Rouge	Spice
Show Tell	31.5%	34.0%	22.6%	14.4%	9.8%	16.3%	30.2%	20.1%
Show Attend	35.3%	34.9%	23.7%	15.5%	10.7%	16.8%	31.7%	20.8%
Up-Down	34.8%	32.7%	22.4%	14.9%	10.4%	16.2%	30.7%	20.9%
Adaptive Att	32.3%	31.7%	21.7%	14.3%	10.0%	15.4%	29.7%	20.4%
Att2in2	34.0%	32.4%	22.3%	14.9%	10.4%	16.0%	30.5%	20.7%
All_img	28%	31.9%	21.4%	13.8%	9.5%	15.7%	29.7%	19.9%

Then we used insertion by sorting to produce the final caption. So you can see the result is much better than the baseline model.

USA	CIDER	B_1	B_2	B_3	B_4	Meteor	Rouge	Spice
Show Tell	15.7%	15.5%	6.6%	3.3%	1.9%	5.6%	14.1%	4.6%
Show Attend	18.2%	15.3%	7.7%	3.8%	2.5%	5.9%	14.8%	5.3%
Up-Down	17.4%	14.7%	6.6%	3.6%	2.1%	5.7%	14.8%	4.9%
Adaptive Att	17.5%	14.9%	6.7%	3.6%	2.2%	5.7%	14.9%	4.9%
Att2in2	17.3%	14.7%	6.6%	3.5%	2.1%	5.6%	14.8%	4.8%
All_img	15%	14.2%	6.1%	3.1%	1.8%	5.4%	14.1%	4.3%

5 CONCLUSIONS

Generally speaking, we did a series of experiments in order to get better captions. From baseline model(CNN + LSTM) to Entity-aware model together with visual attention, which can produce better templates. Finally, we can use name entity insertion by sorting to produce predicted captions. The performance did improve a lot. However, there are still points to improve. For example, name entity tags like PERSON, LOC are special in vocabulary because they are much more frequent than other words. So, this imbalance problem limits the performance. How to figure it out is our future work.

REFERENCES

[1] P. Anderson et al. "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 6077–6086.

[2] Kyunghyun Cho et al. "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734. DOI: 10.3115/v1/D14-1179. URL: https://www.aclweb.org/anthology/D14-1179.

[3] Michael Denkowski and Alon Lavie. "Meteor Universal: Language Specific Translation Evaluation for Any Target Language". In: *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Baltimore, Maryland, USA: Association for Computational Linguistics, June 2014, pp. 376–380. DOI: 10.3115/v1/W14-3348. URL: https://www.aclweb.org/anthology/W14-3348.

[4] J. Lu et al. "Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 3242–3250.

[5] Kishore Papineni et al. "BLEU: A Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. ACL '02. Philadelphia, Pennsylvania: Association for Computational Linguistics, 2002, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: https://doi.org/10.3115/1073083.1073135.

[6] S. J. Rennie et al. "Self-Critical Sequence Training for Image Captioning". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 1179–1195.

[7] Ramakrishna Vedantam, C. Zitnick, and Devi Parikh. "CIDEr: Consensus-based Image Description Evaluation". In: (Nov. 2014).

[8] O. Vinyals et al. "Show and tell: A neural image caption generator". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015, pp. 3156–3164. DOI: 10.1109/CVPR.2015.7298935.