

품사 정보를 활용한 이미지 캡션 생성

(Boosting Image Caption Generation with Parts of Speech)

강 필 구 [†] 임 유 빈 ^{**} 김 형 주 ^{***}
(Philgoo Kang) (Yubin Lim) (Hyoungjoo Kim)

요약 일상 생활 속에서의 스마트 기기와 AI에 대한 의존도가 높아지면서, 시각 장애인 보조, 인간 컴퓨터 상호 작용 등 다양한 분야에 접목 가능한 이미지 캡션 생성 기술의 중요성이 높아지고 있다. 본 논문에서는 캡션 생성 기능의 향상을 위해 명사, 동사와 같은 언어의 품사(POS) 정보를 이미지로부터 추출하여 활용하는 새로운 기법을 제안한다. 제안하는 모델은 복수의 CNN 인코더를 품사 별로 학습하여 품사 별 특징 벡터를 추출한 후, 추출한 품사 벡터를 LSTM에 입력하여 캡션을 생성한다. 제안한 모델은 Flickr30k, MS-COCO 데이터 셋에 대해 실험을 진행하며, 사람을 대상으로 2가지 설문 조사를 진행하여 결과물의 실질적인 유효성을 검증한다.

키워드: 이미지 캡션 생성, 인코더-디코더 구조, 품사, 컴퓨터 비전

Abstract With the integration of smart devices and reliance on AI into our daily lives, the ability to generate image caption is becoming increasingly important in various fields such as guidance for visually-impaired individuals, human-computer interaction and so on. In this paper, we propose a novel approach based on parts of speech (POS), such as nouns and verbs extracted from image to enhance the image caption generation. The proposed model exploits multiple CNN encoders, which were specifically trained to identify features related to POS, and feed them into an LSTM decoder to generate image captions. We conducted experiments involving both Flickr30k and MS-COCO datasets using several text metrics and additional human surveys to validate the practical effectiveness of the proposed model.

Keywords: image caption generation, encoder-decoder architecture, parts of speech, computer vision

[†] 비 회 원 : Tmax Data Tiberio 2-3팀 연구원
philgookang@gmail.com
^{**} 비 회 원 : 서울대학교 컴퓨터공학부 학생(Seoul Nat'l Univ.)
yblim@idb.snu.ac.kr
(Corresponding author임)
^{***} 종신회원 : 서울대학교 컴퓨터공학부 교수
hjk@snu.ac.kr

논문접수 : 2020년 7월 8일
(Received 8 July 2020)
논문수정 : 2020년 11월 10일
(Revised 10 November 2020)
심사완료 : 2020년 12월 30일
(Accepted 30 December 2020)

Copyright©2021 한국정보과학회 : 개인 목적이거나 교육 목적인 경우, 이 저작물의 전체 또는 일부에 대한 복사본 혹은 디지털 사본의 제작을 허가합니다. 이 때, 사본은 상업적 수단으로 사용할 수 없으며 첫 페이지에 본 문구와 출처를 반드시 명시해야 합니다. 이 외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 얻고 비용을 지불해야 합니다.
정보과학회논문지 제48권 제3호(2021. 3)

1. Introduction

In recent years, research on image caption generation, the process of generating a description for an image, has garnered more attention from researchers due to its wide range of applications – guidance for the visually impaired and interaction between human and computer to name a few.

Image captioning essentially relies on visual and linguistic understanding, requiring precision in detection of visual features and selection of appropriate words for semantically and syntactically correct captions. The whats and the hows of these actions have their respective challenges. Therefore, combining them to cooperate as a single entity engenders a whole new set of problems. For example, an extracted visual feature, despite its importance to the whole image, may not align with what is considered useful for the language model to generate an effective sentence. This type of misdetection would trigger a misguided selection thereby formulating a poor caption.

Numerous models aiming to compensate for these limitations and challenges have been proposed. The most renown approach is neural image captioner (NIC)[1] which is based on the most widely used encoder-decoder architecture. CNN encoder for extracting object-related features is used in combination with RNN model, especially LSTM, for generating caption sentences. Based on the architecture of [1], several approaches exploiting additional features from images have also been proposed [2-7]. Commonality discerned among these methods is the lack of addressing for issues related to grammar. Aforementioned models rely solely on the language model to resolve all grammar-related issues.

In this paper, we propose a novel approach that utilizes several Part-Of-Speech (POS) based CNN encoders. We first train several CNN encoders which extract visual features related with each POS. Then we combine POS features into a single vector and finally feed it into a language model. By detecting the different POS contained in images, we could generate captions which are similar in quality to human generated captions. To validate our approach, we evaluate our model on Flickr30k and MS-COCO

2014 dataset with the widely-used metrics including BLEU[8], CIDEr[9], and ROUGE[10]. For BLEU-4 metrics on MS-COCO 2014 dataset, our model scored a value of 34.27 which outperforms previous works. Furthermore, we perform two types of human surveys to evaluate our model qualitatively. Through these evaluation methods, we attempt to prove that our approach has capability of generating high quality captions utilizing POS features.

This paper is organized as follows: Section 2 we explain prior research methods and juxtapose their approach. Section 3 we give an overview of our model and explain our POS CNN image caption model, and Section 4 we evaluate our model and in Section 5 we conclude our research.

2. Related Works

A large number of deep neural network have been proposed for image captioning. The most well-known and popular network is neural image captioner (NIC) [1] which exploits encoder-decoder architecture. In NIC, image features are encoded through CNN encoder and encoded features are fed into LSTM decoder to generate a caption.

There have been lots of efforts to enrich the features for captions based on encoder-decoder architecture. Most of them utilizes additional features from image to guide the recurrent neural network to generate better caption. [2] added an attention mechanism to make the model focuses on the region corresponding with text. [3] extracted both features and attributes from the image to boosting LSTM prediction. They also proposed several feeding combinations. [4] employed a semantic attention model and a feedback loop to guide the recurrent neural network language model during each iteration when it creates the sentence. [5] also guided the language model by adding a bias to words that are semantically linked to the content of the image. Aforementioned methods only focus on image-based additional features without regard to grammar.

Some researches proposed methods which utilizing POS information, one of the basic grammatical feature. [6] extracted the POS tag for each word and

fed POS tag sequences to language model with visual feature vectors. [7] summarized an image with quantized POS tag sequences and generate captions conditioned on it. These approaches enhance syntactical completeness of caption. However, these approaches only exploits annotated POS tag sequences to guide language model. In our work, we directly extract visual features of each POS from image based on encoder-decoder architecture.

3. Proposed Model

Our model is based on encoder-decoder architecture which is effective and most widely used in image captioning. The whole framework of our model is shown in Fig. 1. Following previous approaches, our model consists of two main modules: several CNN encoders and a single LSTM decoder.

3.1 Problem Setup

The goal of our model is to utilize POS features which are extracted from several CNN encoders to enrich the semantics of a sentence S . A caption $S = \{w_1, w_2, \dots, w_{N_s}\}$ consists of a sequence of words w_i , $i \in \{1, \dots, N_s\}$ which accurately describe given image I . For each caption S , w_i is obtained from a fixed vocabulary built on ground-truth annotations. Additionally, we use an automatic POS tagger from python NLTK module to annotate POS tag p for each words, where $p \in P$. P includes only 5 POS tags - noun, verb, adjective, conjunction and preposition. This is because number of instances of some POS tags are somewhat insufficient for proper training. The details about POS tags in whole dataset are shown in Table 1.

Table 1 Frequency of each POS tag in dataset

Part-of-Speech	Flickr30k	MS-COCO
Noun	478,411	2,086,418
Pronoun	52,117	5
Verb	192,518	515,896
Adjective	398,137	532,193
Adverb	37,472	85,347
Conjunction	232,697	722,159
Preposition	201,438	721,964
Interjection	5	5

To train our encoders for each POS tag, we split dataset into 5 according to POS. We train encoder E_p using split $D_p = \{(I, w_{i_p})\}$ which contains data (I, w_{i_p}) composed of image I , and word w_{i_p} where $i_p \in \{1, \dots, N_s\}$ is index of word corresponding to POS p .

3.2 Encoder

Training a CNN model to detect POS features from an image is a hard task. Approaches such as multiple instance learning or unsupervised learning showed less accuracy than supervised learning. Hence, we simply adopt ResNet-152 network for our encoders E_p which showed best performance in LSVR Challenge.

We first pretrain encoder E_p , which is pretrained on ImageNet, using dataset D_p separately to capture the intrinsic features of each POS. In case of adjective encoder E_{adj} , for example, we feed dataset $D_{adj} = \{(I, w_{i_{adj}})\}$ consisting of image and all adjective-tagged words. Once pretrain process finish, we repeat the same process on other encoders. We confirm whether this approach works by conducting a simple experiment. We solve top-1 classification

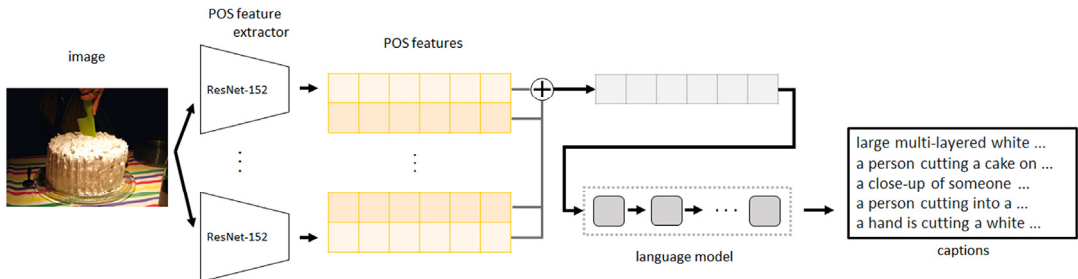


Fig. 1 Illustration of proposed POS CNN encoder-decoder model

Table 2 Top-1 classification accuracy of POS

	Verb	Adj.	Conj.	Prep.
Top-1 (%)	61.84	99.80	85.14	85.97

task with all pretrained POS encoders with MS-COCO dataset and it shows quite good performance as described in Table 2.

It can be inferred that each POS encoder successfully captures features of each POS group.

After pretraining, we remove last fully connected layer and softmax layer of pretrained encoders and utilize 2,048-way *pool-5* layer from ResNet-152 encoder as image representation.

Our model consists of several encoders with a single LSTM decoder. Hence, it is another challenge to combine POS feature vectors into single vector. Among various approaches such as concatenation, element-wise addition, multiplication or fully connected layer, we empirically select element-wise multiplication. Given image I , we formulate input to decoder, i.e. X , as:

$$X = \frac{1}{|P|} \sum_{p \in P} E_p(I), \quad (1)$$

3.3 Decoder

Decoder is a word generator. Given POS feature representation X produced from encoder, as indicated by Eq. (1), the decoder generates sentence one by one in a recurrent manner. In our model, we train decoder to maximize likelihood $p(S|X; \theta)$ by using the following formulation:

$$\theta^* = \operatorname{argmax}_{\theta} \sum \log p(S|X; \theta) \quad (2)$$

where θ are parameters of our model, X is POS representation as indicated by Eq. (1), and θ^* are optimal parameters. This formulation is further factorized into:

$$\log p(S|\theta) = \sum_{t=0}^{N_s} \log p(w_t | X, w_0, \dots, w_{t-1}) \quad (3)$$

where we assume that the sentence is generated one by one depending on POS feature vector X , and previously generated words $w_{0:t-1}$.

Hence, it is natural to model the decoder with recurrent network, especially LSTM. The formulations for our LSTM decoder are summarized as below. For time step t , x^t , h^t , and c^t are the input,

hidden and cell state respectively. Given inputs x^t , h^{t-1} , and c^{t-1} , LSTM decoder updates for time step t as following:

$$i^t = \sigma(W_i x^t + R_i h^{t-1} + b_i), \quad (4)$$

$$f^t = \sigma(W_f x^t + R_f h^{t-1} + b_f), \quad (5)$$

$$o^t = \sigma(W_o x^t + R_o h^{t-1} + b_o), \quad (6)$$

$$z^t = \phi(W_z x^t + R_z h^{t-1} + b_z), \quad (7)$$

$$c^t = i^t \odot z^t + f^t \odot c^{t-1}, \quad (8)$$

$$h^t = o^t \odot \phi(c^t). \quad (9)$$

where i^t , f^t , o^t , and z^t are input gate, forget gate, output gate, cell input respectively. W , R , and b represent input weight matrices, recurrent weight matrices and bias vectors respectively.

3.4 Training

Based on all the computational details of encoders and decoder, the whole procedure of our model is briefly described as:

$$x^{-1} = X = \frac{1}{|P|} \sum_{p \in P} E_p(I), \quad (10)$$

$$x^t = W_e w_t, \quad t \in \{0 \dots N_s - 1\}, \quad (11)$$

$$p_{t+1} = LSTM(x^t), \quad t \in \{0 \dots N_s - 1\} \quad (12)$$

where W_e represents word embedding matrix. According to Eq. (1), loss function of our model is the negative log likelihood at each step t which is formulated as:

$$L(X, S) = - \sum_{t=1}^{N_s} \log p_t(S_t). \quad (13)$$

The above loss is minimized w.r.t all parameters of encoders and decoder of proposed model.

4. Experiment

4.1 Dataset

We evaluate our model using two publicly available dataset which consist of images and captions describing images: Flickr30k and MS-COCO 2014, which are commonly used in evaluating image caption generation tasks.

Flickr30k is a dataset collected by crowd-sourcing on Flickr web site. The majority of the content of the image is based on human daily activities. It contains 31,783 images and 158,915 English captions (5 captions per image). As Flickr30k has no standard

split for train, validation, and test, we hold out 1,000 images for validation and test and train on the remainders.

MS-COCO 2014 contains 123,287 images and 616,767 English captions. In the case of MS-COCO, we validate and test using 5,000 images each and train with remainders.

4.2 Implementation and Setup

Our model consists of several pretrained CNN encoders with a single LSTM decoder. Table 3 provides the details about decoder depending on dataset. We use Adam optimizer[11] with learning rate 0.0005 for entire end-to-end training with batch size 64. We also utilizes batch normalization and early stopping techniques to prevent over-fitting and optimize training. For caption generation in inference stage, we empirically select beam search with beam size of 5.

We conduct all experiments on a server with Intel E5-2650 2.20GHz, 128GB of main memory, and four NVIDIA GeForce TITAN Xp GPUs.

Table 3 Details of decoder architecture

	Flickr30k	MS-COCO
Dim. of input	256	512
Dim. of hidden	256	512
No. layers	2	3

4.3 Evaluation Metrics

Our goal is to generate semantically and syntactically well-formed caption. To validate our model, we perform an automatic evaluation on generated captions with several metrics and also conduct human surveys due to the ambiguity of natural languages.

4.3.1 Automatic evaluation

Several metrics, such as BLEU, CIDEr, and ROUGE,

have been proposed to evaluate generated captions and widely used in image captioning. These metrics can be computed automatically by comparing with the ground truths. We compare the results with those of previous methods.

4.3.2 Human survey

It is still difficult to evaluate whether a caption is good enough in terms of human perception or not with aforementioned metrics. Hence, we conducted human surveys to demonstrate the effectiveness of our model intuitively. We conducted 2 types of surveys with 20 human respondents; one is to score caption quality and the other one is to compare with ground truth. Our respondents are randomly chosen from a wide variety of backgrounds and age groups.

For the first survey, we randomly sampled 40 test examples from MS-COCO dataset following the method conducted in [1]. Each item consists of 1 image and 1 caption with 4 options; 1) no error 2) minor error 3) little related 4) wrong.

The purpose of second survey is to evaluate the quality of captions in comparison with ground truth captions. The survey consists of 25 questions excluding the images used in first survey. Each item contains 1 image and 3 options. One is ground truth caption, another is generated caption and the last one is an option which denotes that 2 captions are similar. Respondents select the best among the 3 options.

4.4 Evaluation Results

We compare our proposed model with several image caption models on Flickr30k and MS-COCO 2014 datasets. Table 4 shows the performance of our model in comparison to other existing models on test dataset. The performance of our model is competitive enough on both datasets.

Table 4 Performance of proposed model and other image caption models in Flickr30k and MS-COCO 2014

Model	Flickr30k				MS-COCO					
	B1	B2	B3	B4	B1	B2	B3	B4	CD	RG
NIC [1]	62.7	42.3	27.7	18.3	71.6	53.2	41.3	33.5	79.6	52.3
LSTM-A ₅ [3]	-	-	-	-	73	56.5	42.9	32.5	98.6	53.8
ATT-FCN [4]	64.7	46.0	32.4	23.0	70.9	53.7	40.2	30.4	-	-
He et al. [6]	63.8	44.6	30.7	21.1	71.1	53.5	38.8	27.9	88.2	-
Deshpande et al. [7]	-	-	-	-	74.4	57.0	41.9	30.6	101.4	53.1
Proposed Model	60.6	44.3	33.8	27.3	71.9	55.7	44.0	36.2	90.6	53.0



Fig. 2 Sample survey results evaluating caption quality generated from the proposed model

On Flickr30k dataset, our model outperforms other models in terms of both BLEU-3 and BLEU-4. Our model shows a big increase compared with the baseline NIC[1]. Though the results on BLEU-1 and BLEU-2 is somewhat poorer than other models, they are still competitive.

Our model shows better performance on MS-COCO dataset. In particular, it outperforms other models on BLEU-3 and BLEU-4 with score of 44.0 and 36.2 respectively. Though, the result on CIDEr is somewhat poor, other results are nearly on par with or better than best results. Overall, our model achieves best performance on most of metrics. It can be

inferred that our POS vector generalizes well and is effective for generating better captions.

In addition, Table 5 and Table 6 show the results of first and second human surveys respectively. The results of first survey shows that almost half of the captions have no error in view of human perception. Furthermore, almost 3 quarters of captions were understandable. It shows that our model is competitive enough as well as POS vectors successfully guide LSTM decoder to generate better captions. Some examples of survey results are shown in Fig. 2.

In second survey, the results are rather not good. Still, given that the comparison targets are human

Table 5 Result of survey evaluating caption quality

Option	Percentage(%)
1 No Error	46.42
2 Minor Error	27.02
3 Little Related	15.00
4 Wrong	11.54

Table 6 Result of survey compared with ground truth

Option	Percentage(%)
Ground Truth	46.32
Our Model	37.35
Similar	16.33




Image	Captions(Votes by percentage)
	Ours (100%): a person is taking a picture of a pizza GT (0%): a man taking a picture of his meal at a diner table
	Ours (65%): two giraffes are eating leaves off of a tree GT (35%): a couple of giraffes that are standing out in a field
	Ours (0%): a man holding a teddy bear in a crowd of people GT (100%): a soldier kneeling down next to little girls

Fig. 3 Sample survey results compared with ground truth

generated captions, the results can be considered as sufficiently competitive. Furthermore, over half of the captions are equal-to or better than human generated captions. Fig. 3 shows some examples of survey results.

5. Conclusion

In this paper, we proposed a novel approach in creating an image caption utilizing grammatical factors extracted from image. Our approach uses multiple CNN encoders to detect parts of speech related features to boost the quality of caption. Considering several evaluation results, we believe that our approach could improve image caption generation compared to prior works. In our future works, we

aim to utilize other forms of features that affects to the syntax.

References

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, Show and tell : A neural image caption generator, *CVPR*, 2015.

[2] K. Xu, J. Ba, R. Kiro, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, *In ICML*, 2015.

[3] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei. Boosting image captioning with attributes, *In CVPR*, 2016.

[4] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention, *In CVPR*, 2016.

[5] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars. Guiding the Long-Short Term Memory model for Image Caption Generation, *In ICCV*, 2015.

[6] X. He, B. Shi, X. Bai, G. S. Xia, Z. Zhang, and W. Dong, Image caption generation with part of speech guidance, *In PRL*, 2019.

[7] Deshpande, Aditya, et al. "Fast, diverse and accurate image captioning guided by part-of-speech," *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[8] K. Papineni, S. Roukos, T. Ward, and W. Zhu, Bleu: a method for automatic evaluation of machine translation, *In ACL*, 2002.

[9] R. Vedantam, C. L. Zitnick, and D. Parikh, CIDEr: Consensus-based image description evaluation, *In CVPR*, 2015.

[10] C. Lin, Rouge: a package for automatic evaluation of summaries, *In ACL Workshop*, 2004.

[11] Diederik P. Kingma and Jimmy Ba, Adam: A method for stochastic optimization, *International Conference on Learning Representations*, 2015.



강 필 구

2018년 한양대학교 컴퓨터공학부 학사
2020년 서울대학교 컴퓨터공학부 석사
2020년~현재 티맥스티베로. 관심분야는 데이터베이스



임 유 빈

2015년 한동대학교 전산전자공학부 학사
2015년~현재 서울대학교 컴퓨터공학부
박사과정 재학 중. 관심분야는 데이터베
이스, 데이터 마이닝



김 형 주

1982년 서울대학교 전산학과 학사. 1985
년 Univ. of Texas at Austin 석사
1988년 Univ. of Texas at Austin 박사
1988년~1990년 Georgia Institute of
Technology 부교수. 1991년~현재 서울
대학교 컴퓨터공학부 교수. 관심분야는
데이터베이스, XML, 시맨틱 웹, 빅데이터