Show and Tell :

A Neural Caption Generator

김경환

NLP boot camp

Contents

- Introduction
- Model
- Experiments
- Results

Introduction

- 이미지를 설명하는 문장을 자동으로 만들어 내는것은 굉장히 어려운 문제다. (challenging task)
- image classification과 object dection보다 훨씬 어렵다.

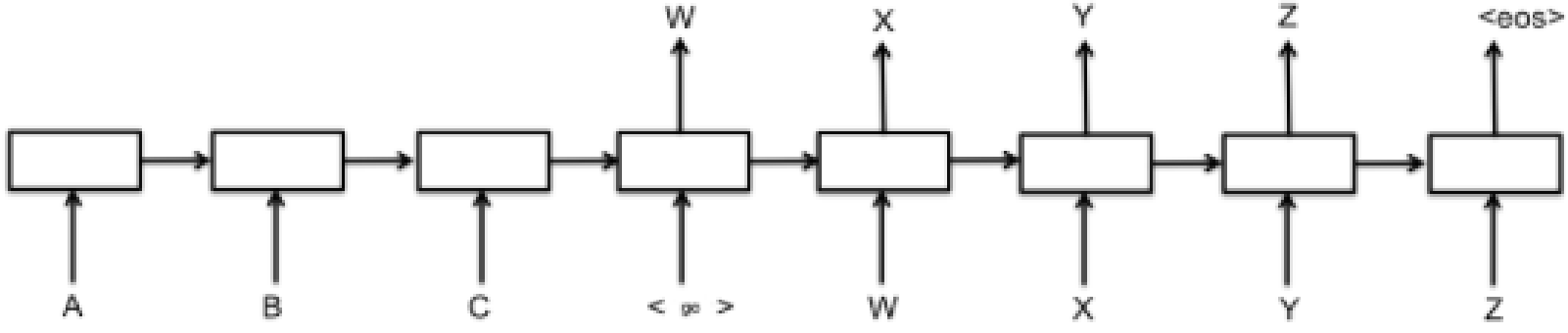
  $\rightarrow$ **이미지 인식 + 자연어 표현까지 학습해야 되기 때문**

Introduction

Idea

- **encoder** RNN - **decoder** RNN
- Source language S - Target language T

maximizing $p(T|S)$ 하도록 학습

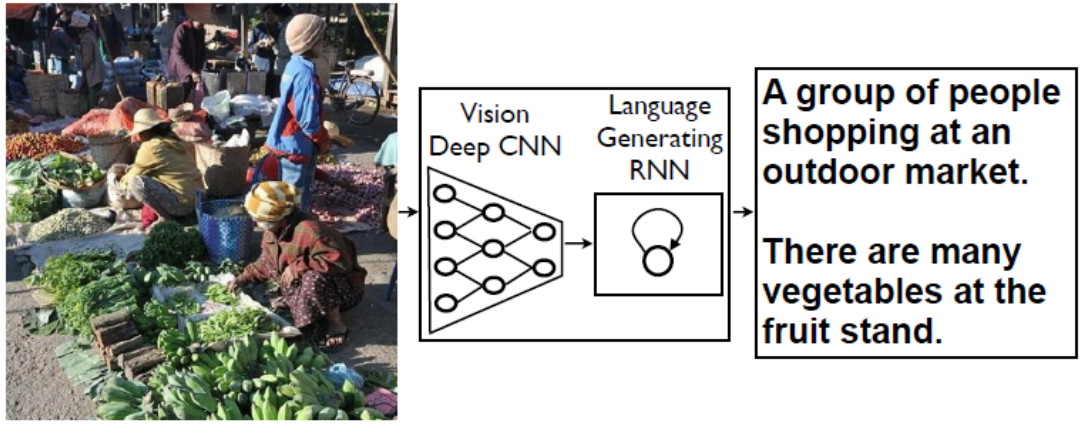


Introduction

Idea

- **encoder** RNN $\rightarrow$ **encoder CNN**
- image I
- target sequence of words $S = \{S_1, S_2, ...\}$
- **Neural Image Caption** model **(NIC)**

maximizing $p(S|I)$ 하도록 학습

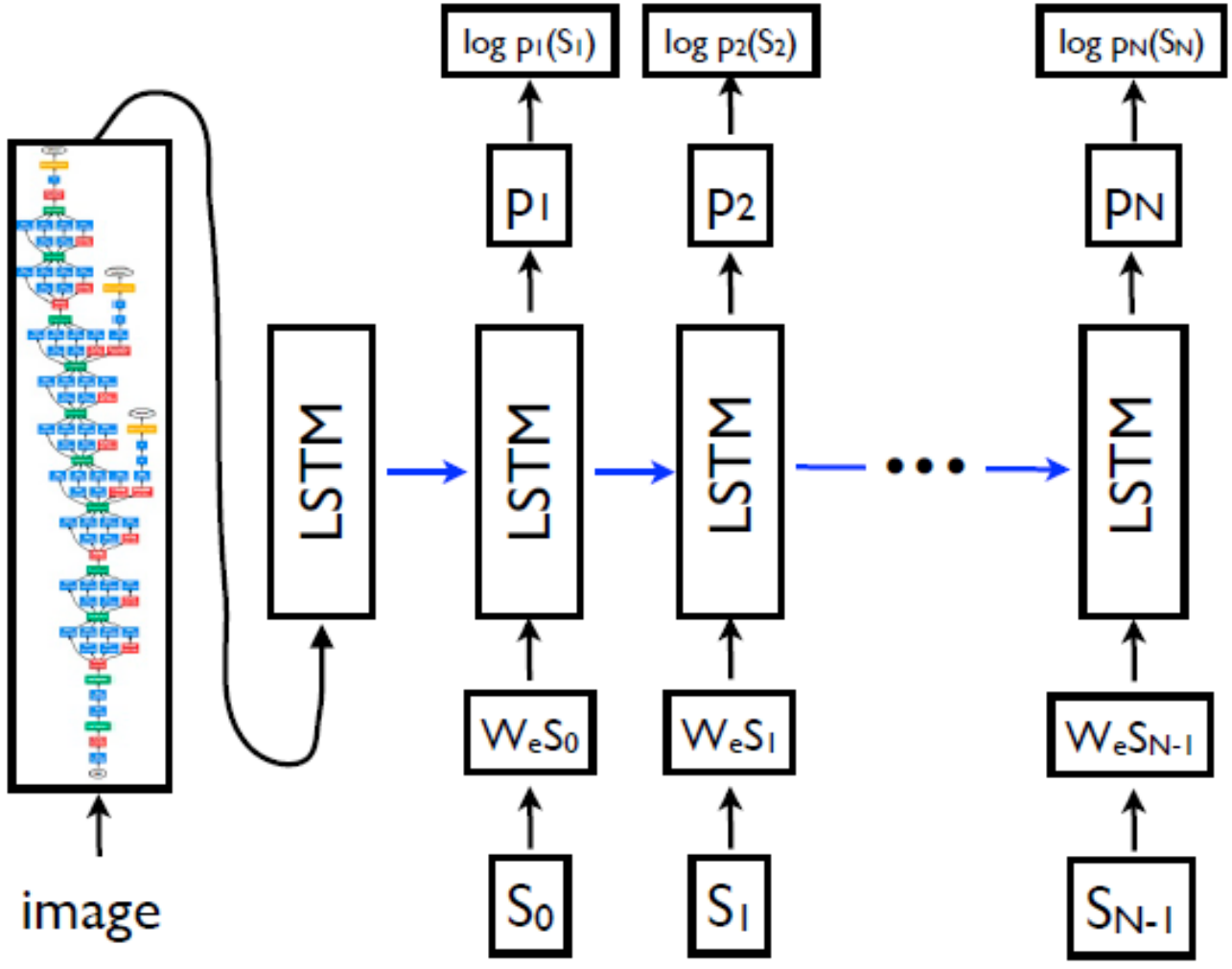


Model

- Machine translation 모델과 같이
- input **Image**가 주어졌을 때 정답 output **Sequence**의 확률을 **maximizing**
- correct transcription S, input image I, parameter θ

$$\theta^* = \arg\max_{\theta} \sum_{(I,S)} logp(S|I;\theta)$$

$$logp(S|I) = \sum_{t=0}^{N} logp(S_t|I, S_0, ..., S_{t-1})$$


log p1(S1)
log p2(S2)
log pN(SN)
p1
p2
pN
LSTM
LSTM
LSTM
LSTM
WeS0
WeS1
WeSN-1
S0
S1
SN-1
image

Model

Training

- Encoder로 Deep CNN 사용.
- Decoder로 LSTM 사용.

$$x_{t-1} = CNN(I)$$

$$x_t = W_e S_t, t \in \{0, ..., N-1\}$$

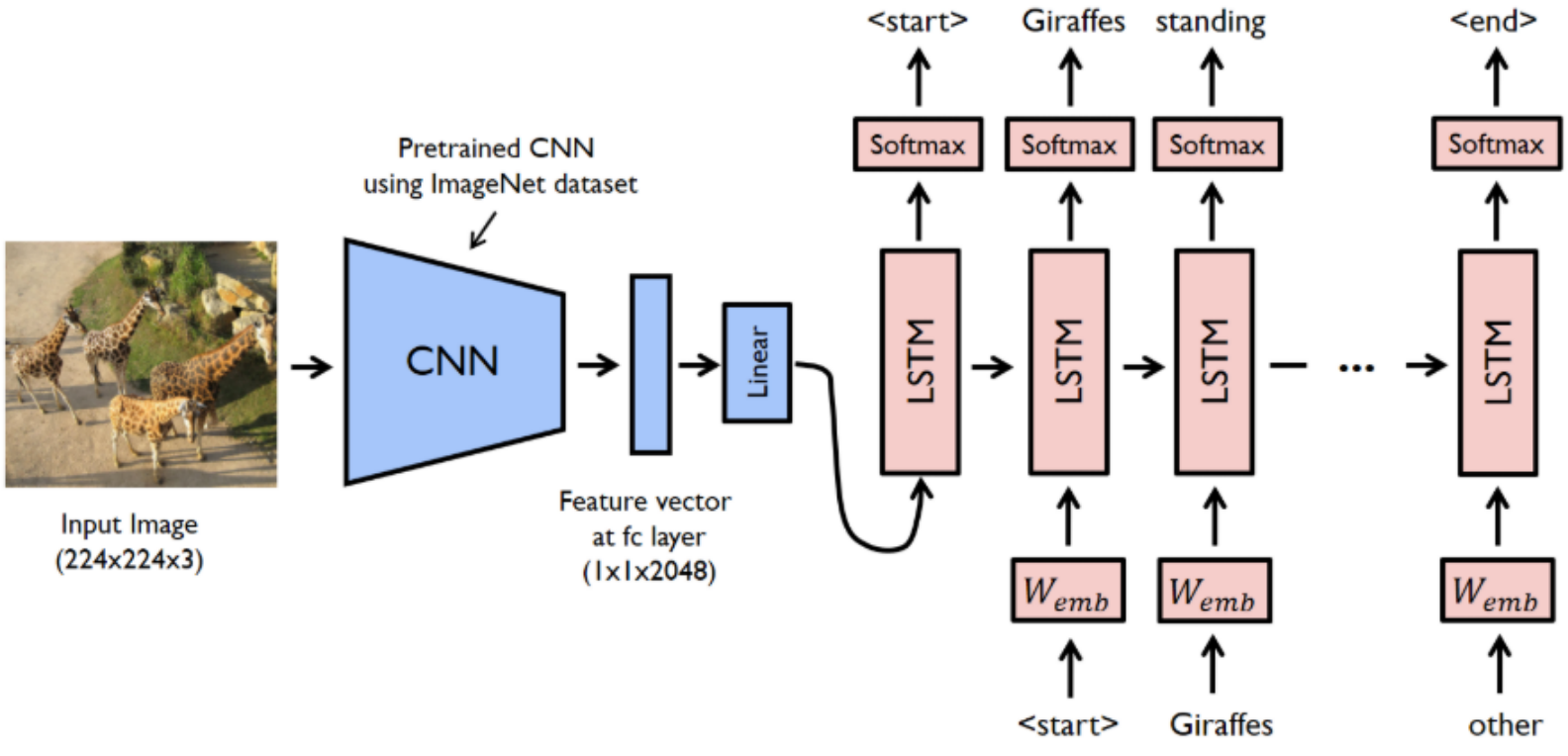
$$p_{t+1} = LSTM(x_t), t \in \{0, ..., N-1\}$$

Model

Training Details

- Over fiiting을 막기 위한 techniques
  i. Pre-Train 된 Deep CNN 사용. (e.g. ImageNet)
  ii. Pre-Train 된 Word embedding vetor 사용. (효과 적음)
  iii. Dropout & Ensembling

Model



ref : https://github.com/yunjey/pytorch-tutorial/tree/master/tutorials/03-advanced/image_captioning

Model

Inference

- Sampling : 가장 확률이 높은 값(단어)을 고른다.
- BeamSearch : k 개의 후보(단어)를 뽑아서 다음 t+1 에서의 단어와의 조합의 확률을 보고 높은 값을 고른다.

Experiments

Evaluation Metrics

- Amazon Mechanical Turk experiment
- BLEU
- METHOR
- CIDER

Experiments

Datasets

- 이미지마다 5개의 문장 (SBU 제외)
- Pascal VOC 2008은 test로만 사용 (실험에서는 학습은 MSCOCO 로 함)
- Flickr는 사진과 사진에 대한 글을 올리는 사이트
- SBU는 Flicker에 올라온 사진과 글을 그대로 데이터로 사용

Dataset name	size		
	train	valid.	test
Pascal VOC 2008 [6]	-	-	1000
Flickr8k [26]	6000	1000	1000
Flickr30k [33]	28000	1000	1000
MSCOCO [20]	82783	40504	40775
SBU [24]	1M	-	-

Results

- How dataset size affects generalization
- What kinds of transfer learning it would be able to achieve
- How it would deal with weakly labeled examples

Results

Generalization

- 좋은 데이터가 10만개 정도
- 데이터가 많아지면 더 좋은 결과가 나올 것이라 예상
- 데이터가 부족하기 때문에 generalization(overfitting 방지)을 하기 위해 노력함.

Results

Generation Results

- 여러 metric으로 평가해봄.
- 사람보다 점수가 높은 경우가 있지만 실제 결과는 그렇지 않음.
- metric에 대한 연구도 더 필요할 것으로 보임.

Generation Results

Metric	BLEU-4	METEOR	CIDER
NIC	**27.7**	**23.7**	**85.5**
Random	4.6	9.0	5.1
Nearest Neighbor	9.9	15.7	36.5
Human	21.7	25.2	85.4

Table 1. Scores on the MSCOCO development set.

Approach	PASCAL (xfer)	Flickr 30k	Flickr 8k	SBU
Im2Text [24]				11
TreeTalk [18]				19
BabyTalk [16]	25			
Tri5Sem [11]			48	
m-RNN [21]		55	58	
MNLM [14][5]		56	51	
SOTA	25	56	58	19
NIC	**59**	**66**	**63**	**28**
Human	69	68	70	

Table 2. BLEU-1 scores. We only report previous work results when available. SOTA stands for the current state-of-the-art.

Results

Transfer Learning

- 다른 dataset 간의 transfer가 가능한지 실험.
- Flickr30k -> Flickr8k (유사 데이터, 데이터 차이 4배)
  - BLEU 4 증가
- MSCOCO -> Flickr8k (다른 데이터, 데이터 차이 20배)
  - BLEU 10 감소, but 만든 문장은 괜찮음.
- MSCOCO -> SBU
  - BLEU 16 감소

Results

Generation Diversity Discussion

- generating model 모델이 새롭고 다양하고 높은 퀄리티의 문장을 만들어내는지 확인.

A man throwing a frisbee in a park. **A man holding a frisbee in his hand.** **A man standing in the grass with a frisbee.**
A close up of a sandwich on a plate. A close up of a plate of food with french fries. A white plate topped with a cut in half sandwich.
A display case filled with lots of donuts. **A display case filled with lots of cakes.** **A bakery display case filled with lots of donuts.**

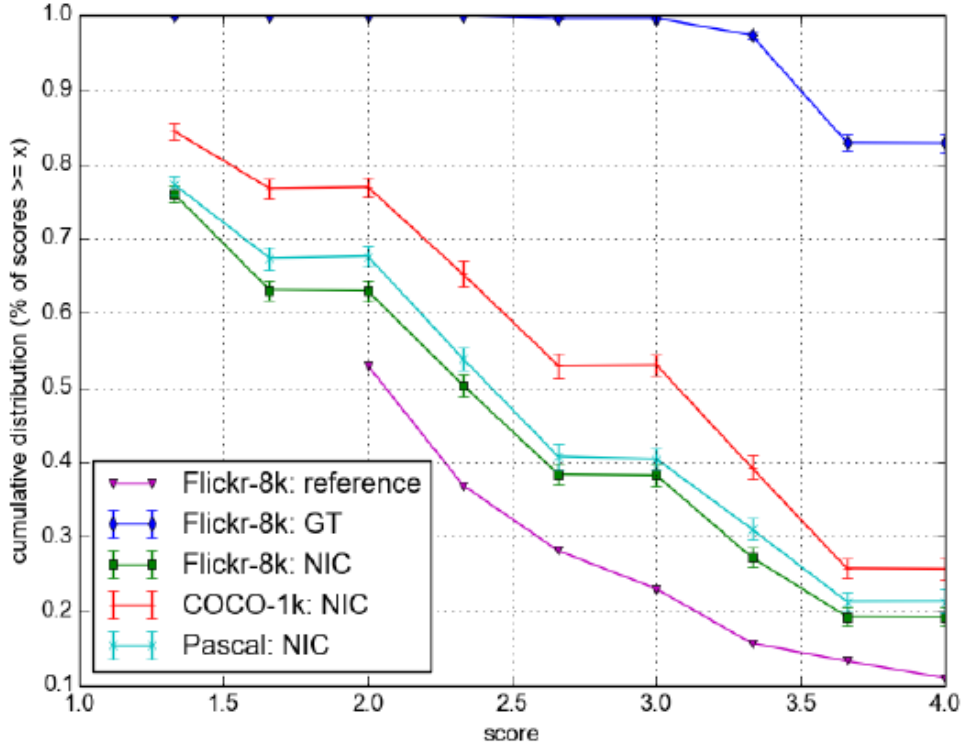
Approach	Image Annotation			Image Search		
	R@1	R@10	Med r	R@1	R@10	Med r
DeFrag [13]	13	44	14	10	43	15
m-RNN [21]	15	49	11	12	42	15
MNLM [14]	18	55	8	13	52	10
NIC	**20**	**61**	**6**	**19**	**64**	**5**

Table 4. Recall@k and median rank on Flickr8k.

Results

Human Evaluation

- 사람이 직접 평가한 지표를 보여줌.
- BLEU Score는 human보다 높았는데, 여기서는 낮다.
- BLEU 지표가 완벽한 지표는 아님을 보여줌.



Human Evaluation

Figure 5. A selection of evaluation results, grouped by human rating.

Results

Analysis of Embedding

- Word embedding vector도 유사한 단어들끼리 뭉쳐있도록 잘 학습됨을 확인.

Word	Neighbors
car	van, cab, suv, vehicule, jeep
boy	toddler, gentleman, daughter, son
street	road, streets, highway, freeway
horse	pony, donkey, pig, goat, mule
computer	computers, pc, crt, chip, compute

Table 6. Nearest neighbors of a few example words

감사합니다.