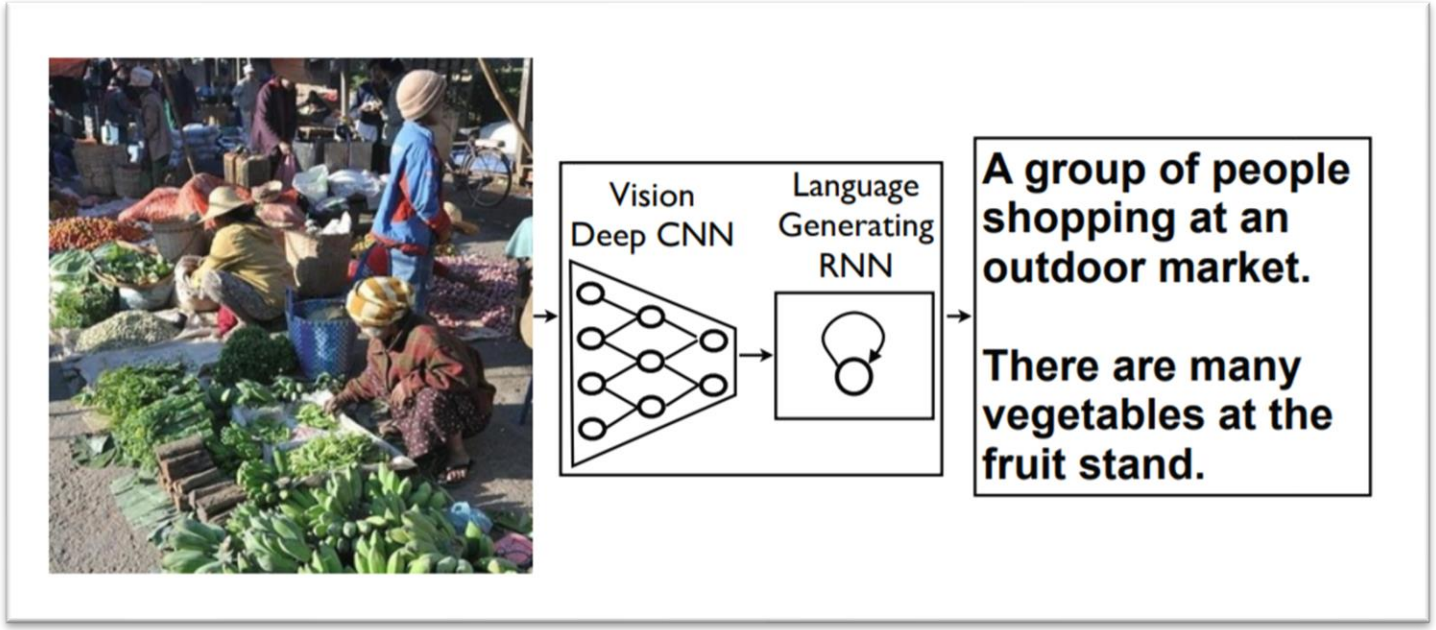
Image Captioning

Machine Learning Study
JinHo Kim

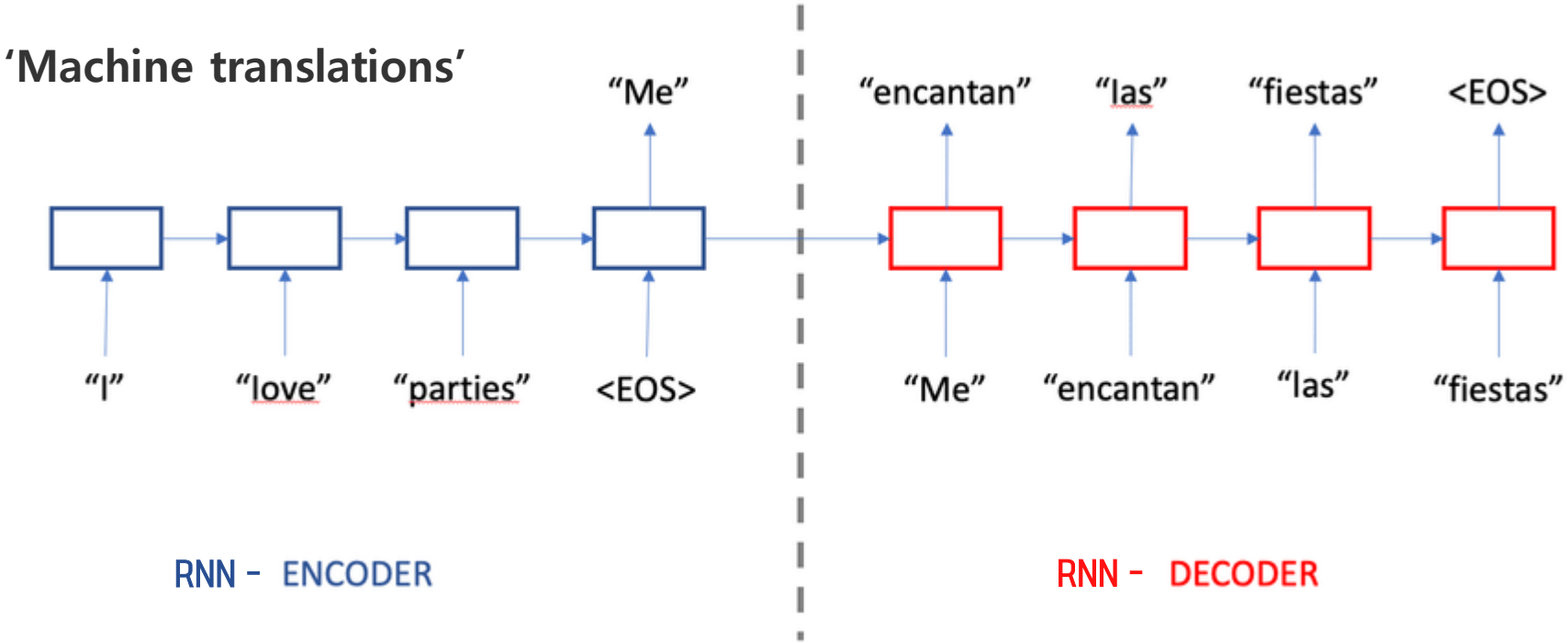
Contents

What is Image Captioning?



"Computer Vision + Natural Language Processing"

What is Image Captioning?



What is Image Captioning?

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	-	-

Dataset

Flickr8k Sample Data

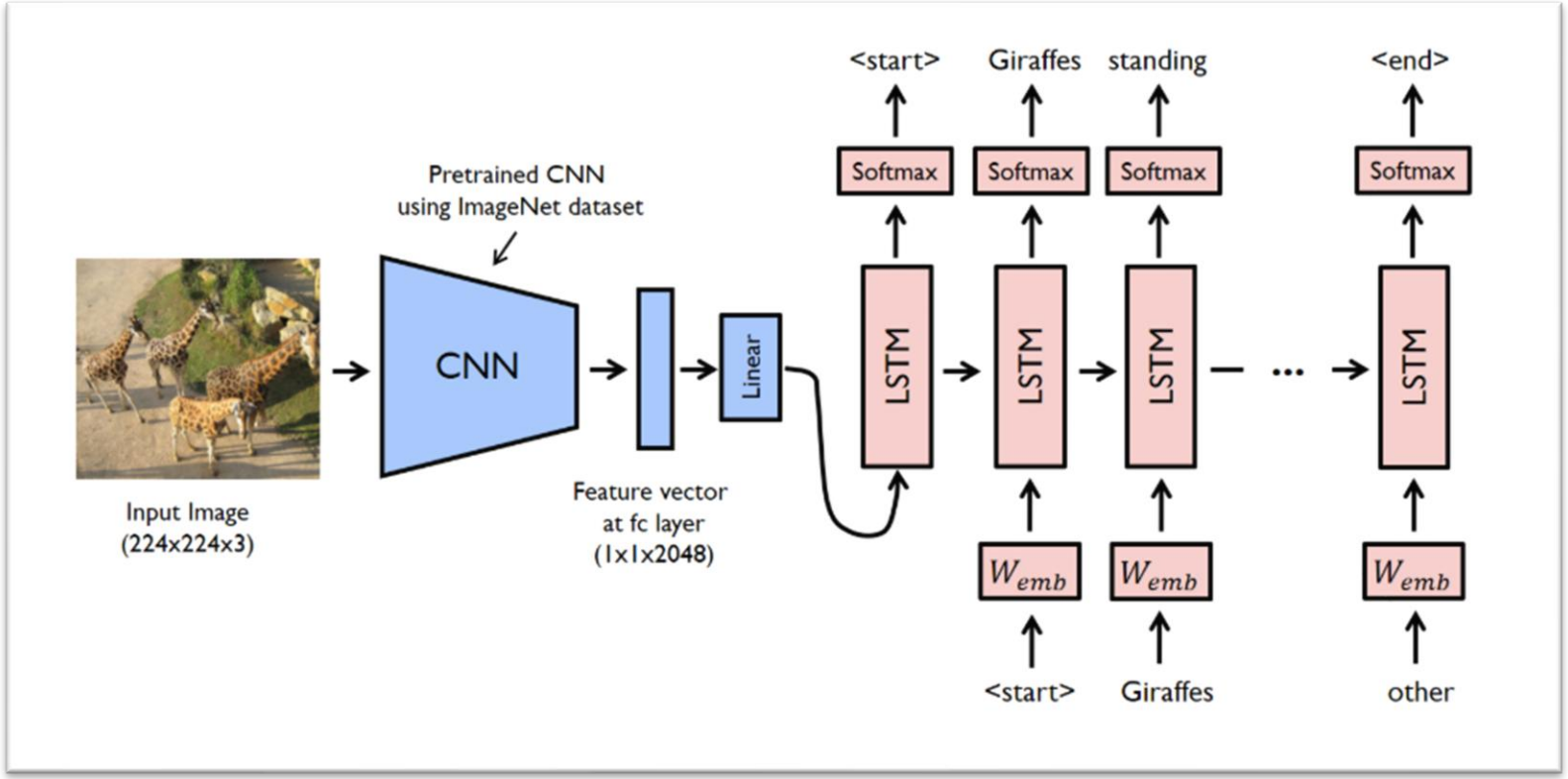
1. A dog is attempting a turn by a nearby picnic bench and metal object .
2. A white dog running next to a bench .
3. The big white dog is running in the grass .
4. The white dog is running around in the grass .
5. White dog with collar running in fenced in grassy area .

Image Captioning Papers

1. Show and Tell
2. Show, Attend and Tell

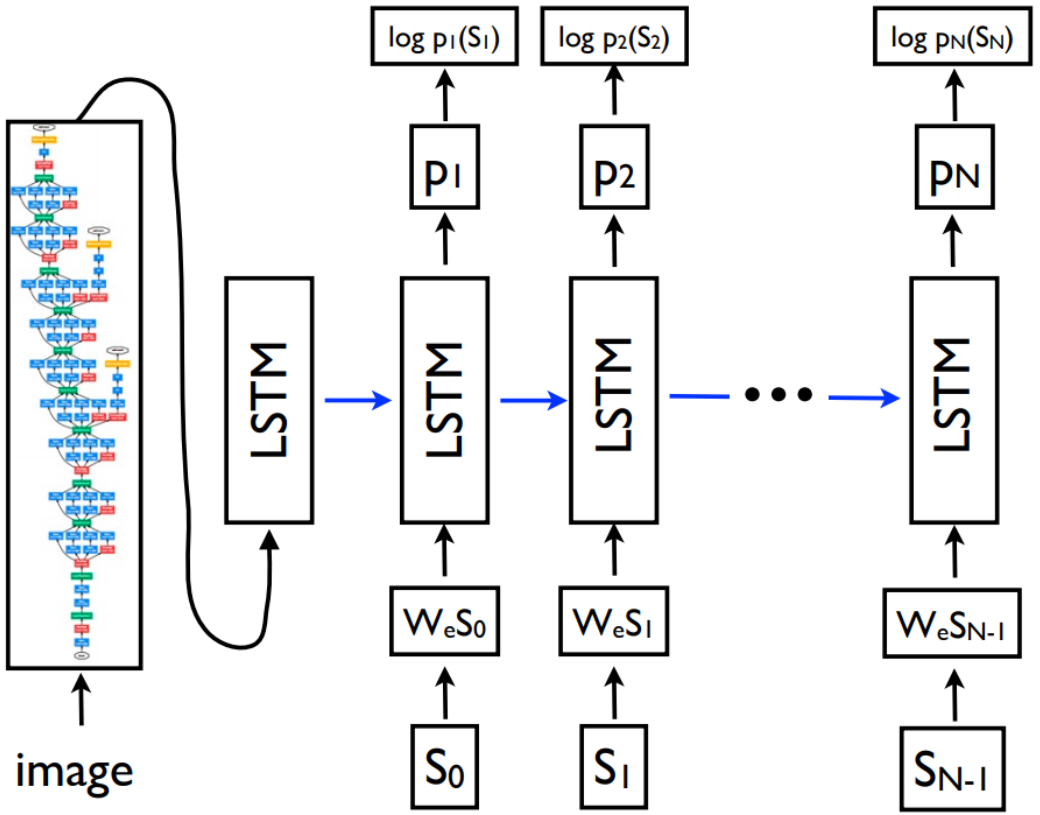
Show and Tell

Neural Image Caption(NIC) Model



Encoder -> pre-trained CNN Decoder → RNN(LSTM)

Show and Tell



For.

gradient vanishing

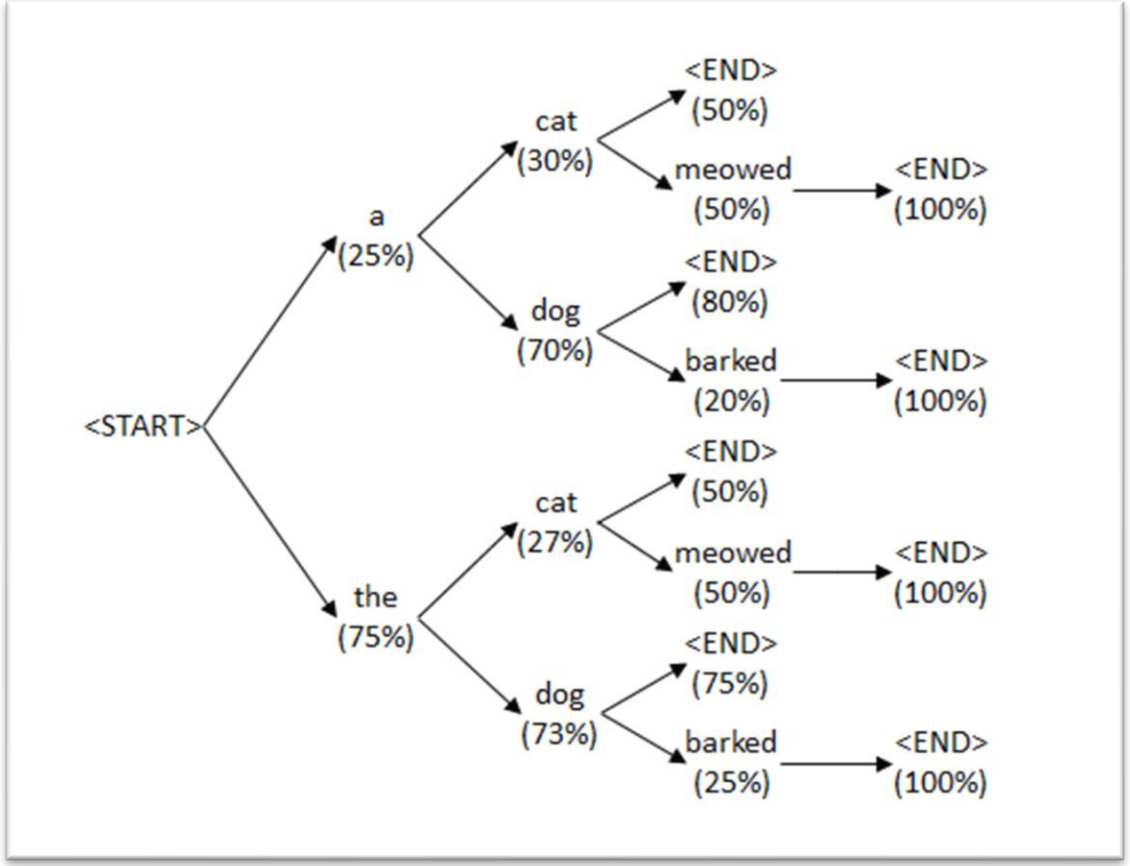
gradient exploding

Show and Tell

Generate sentence approach

1. Sampling : 가장 확률이 높은 값(단어)을 고르는 방법
2. BeamSerarch : k개의 후보를 뽑아서 단어와의 조합 고르는 방법

Show and Tell



BeamSearch

Show and Tell

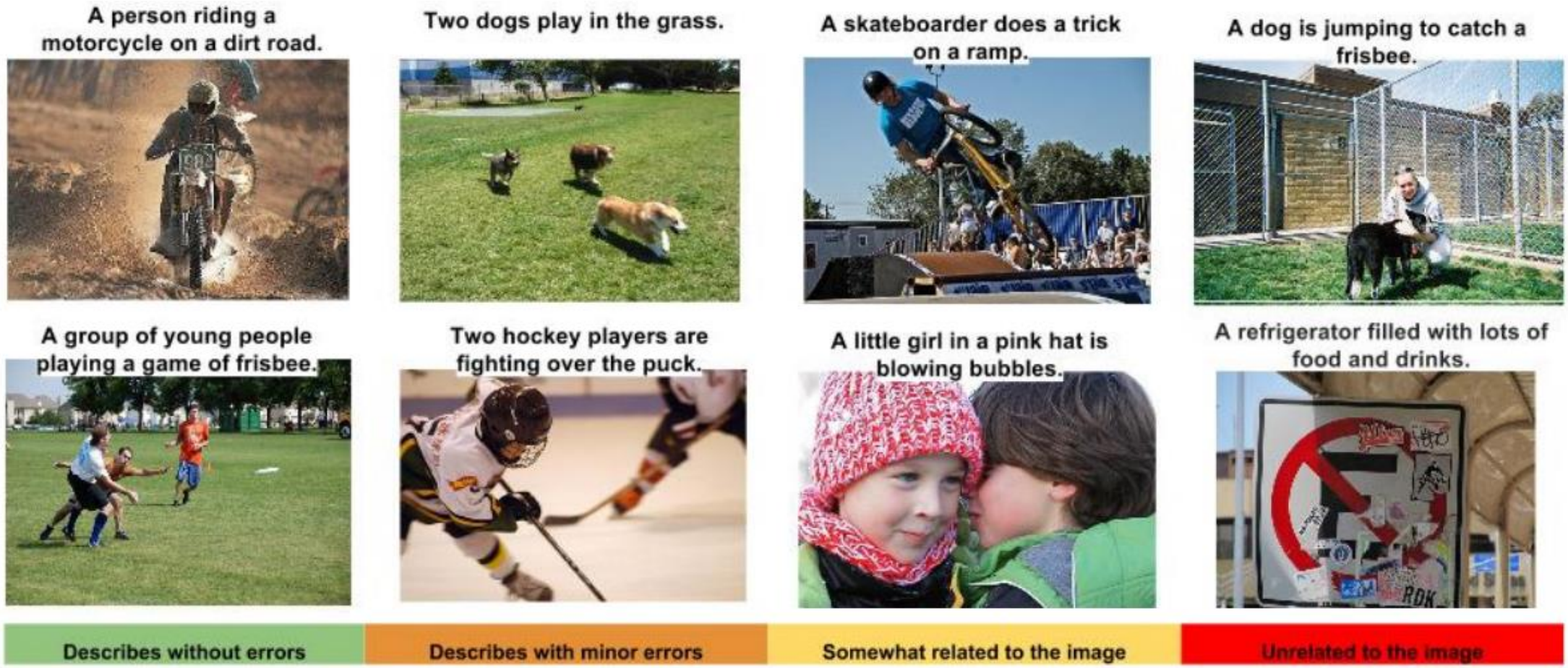
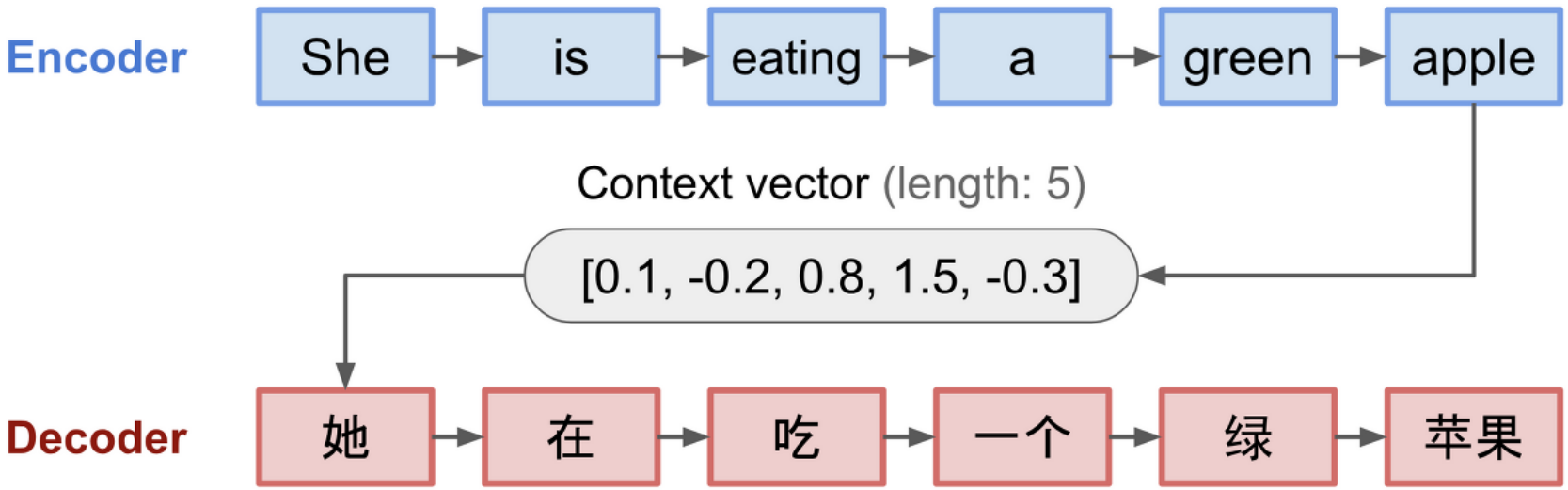


Figure 5. A selection of evaluation results, grouped by human rating.

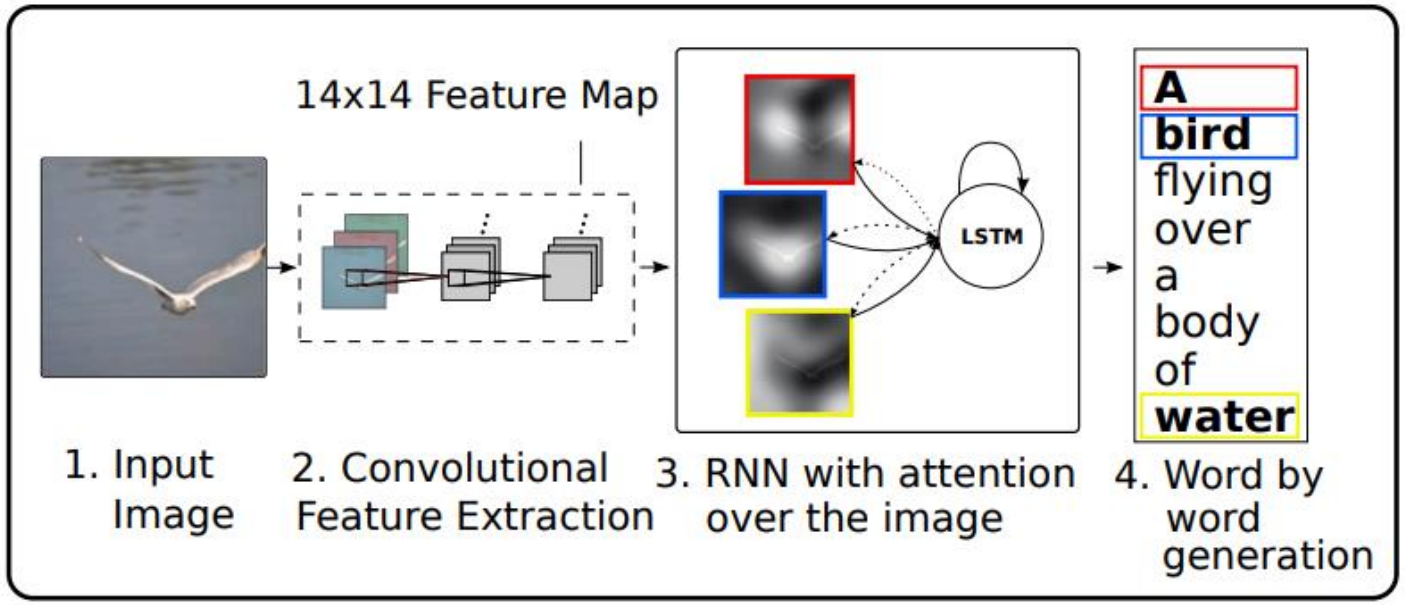
Show and Tell Problem → seq2seq model



Long-term dependency problem!

Show, Attend and Tell

Show and Tell의 후속작 → NIC 모델 + Attention

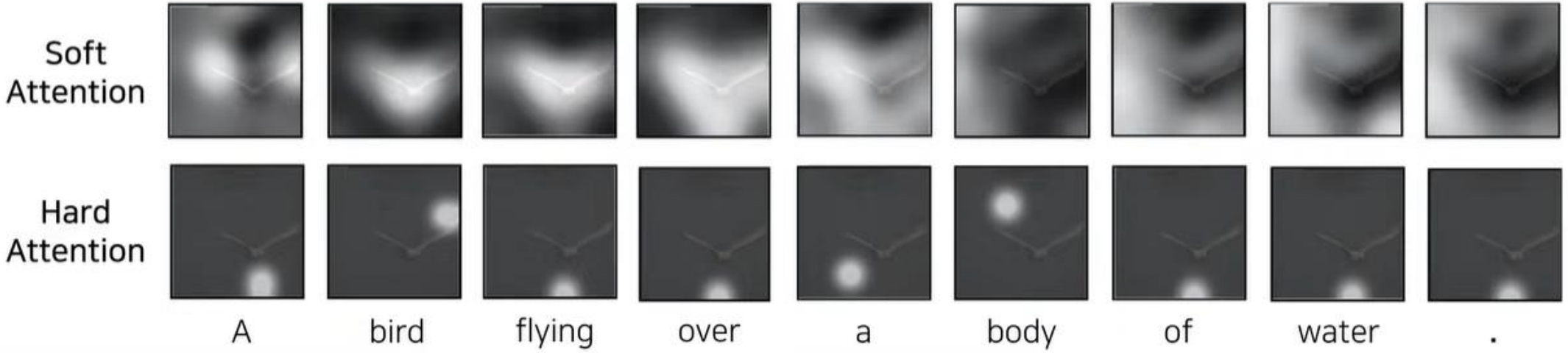


Attention의 기본 아이디어!
Decoder에서 예측하는 매 시점마다 Encoder를 다시 한 번 체크!
→ So, 이미지에 대한 caption 정확도 ↑

Show, Attend and Tell

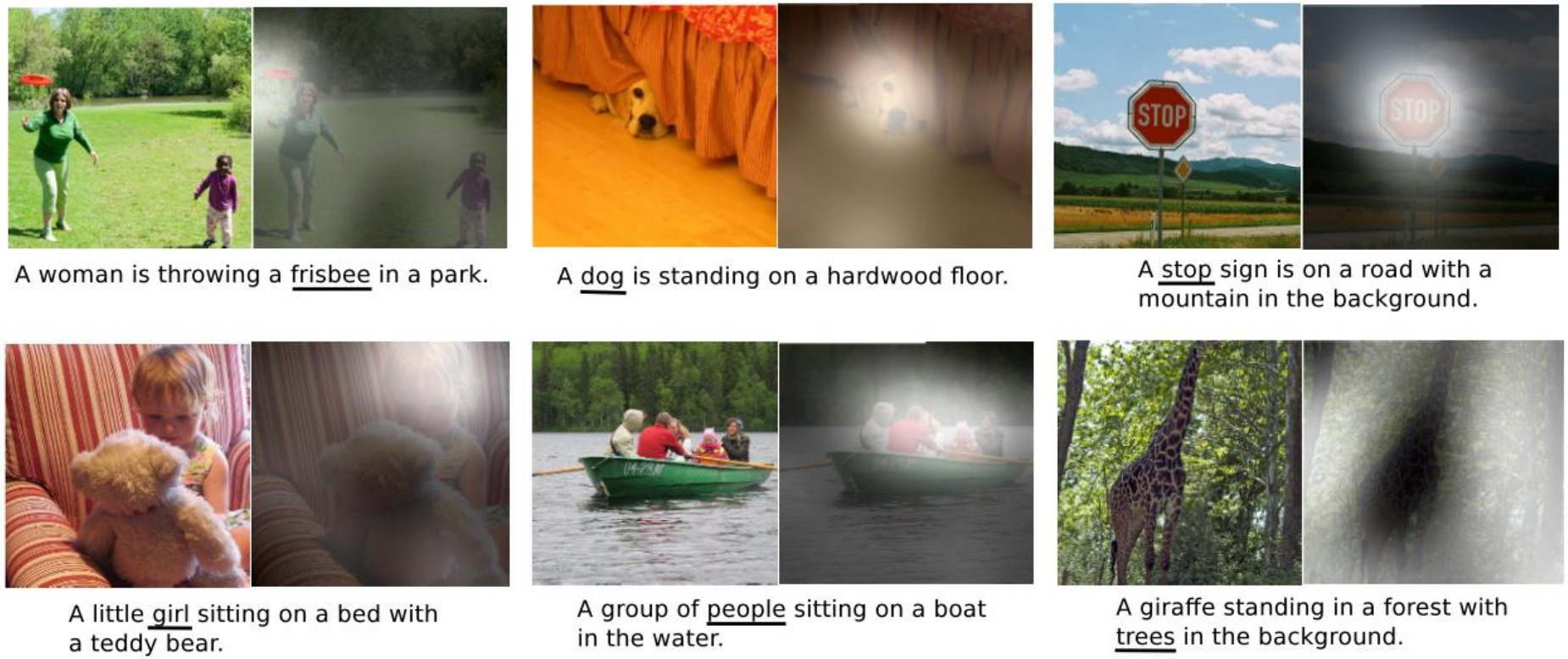
Attention Mechanism
① Hard Attention, ② Soft Attention

: 입력 이미지(Input Image)



Show, Attend and Tell

Attention Mechanism Visualization

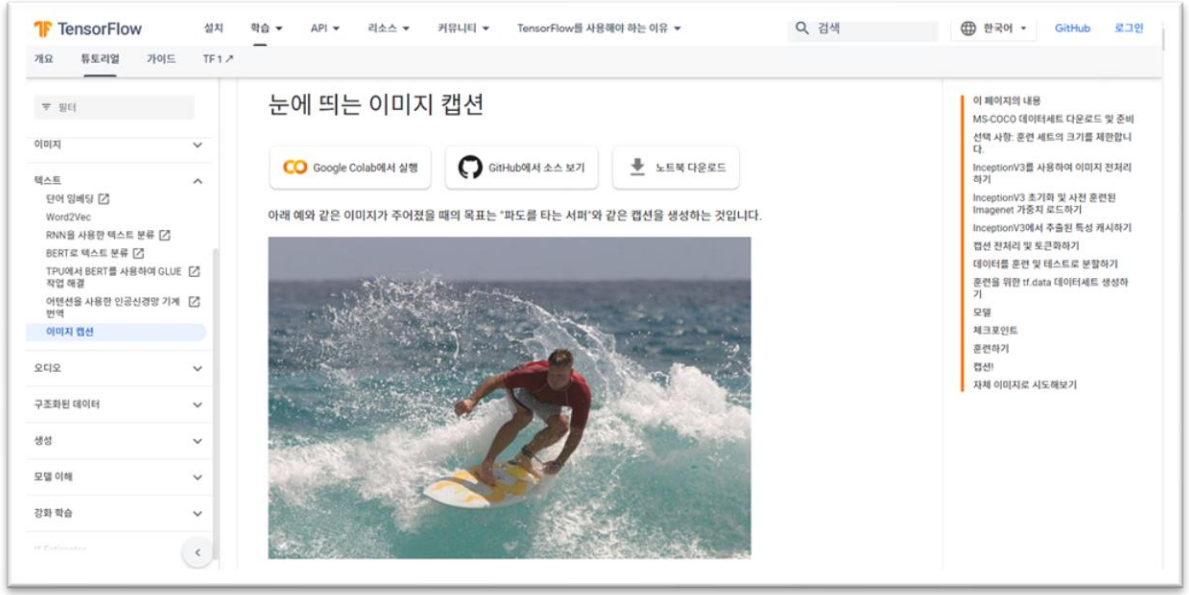


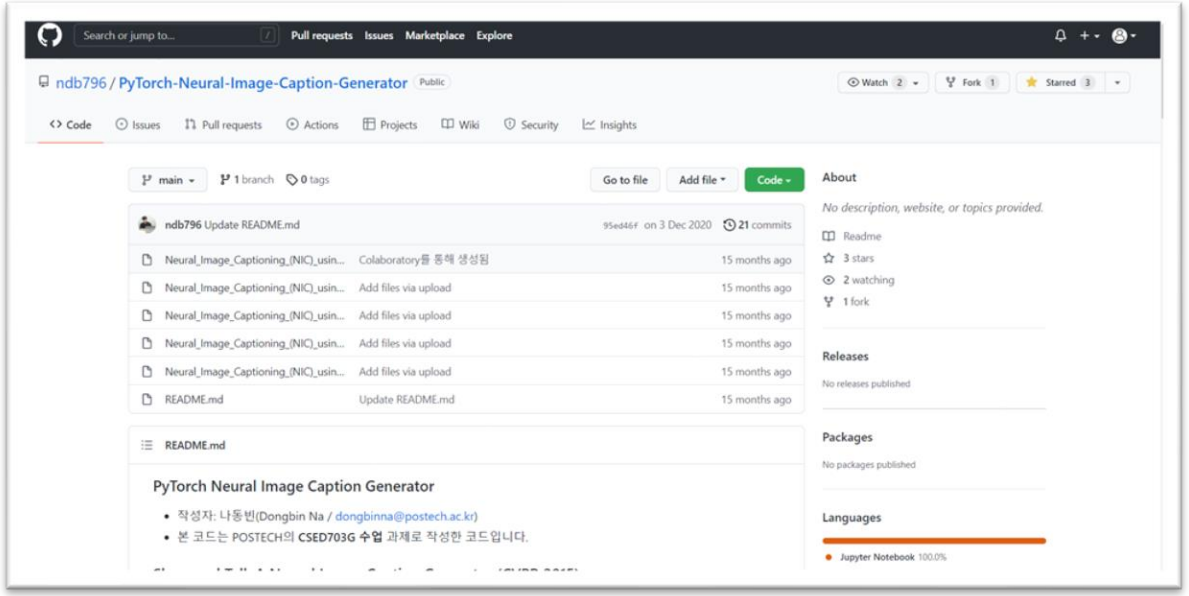
Practice

Source code

https://www.tensorflow.org/tutorials/text/image_captioning

https://github.com/ndb796/PyTorch-Neural-Image-Caption-Generator#pytorch-neural-image-caption-generator





Practice

Dataset : Flickr8k
Encoder : VGG16, ResNet 등
Decoder : LSTM, Attention

모델 학습 결과

Flickr8k test image

정답 캡션

the two small dog run through the grass
two small dog run through the grass

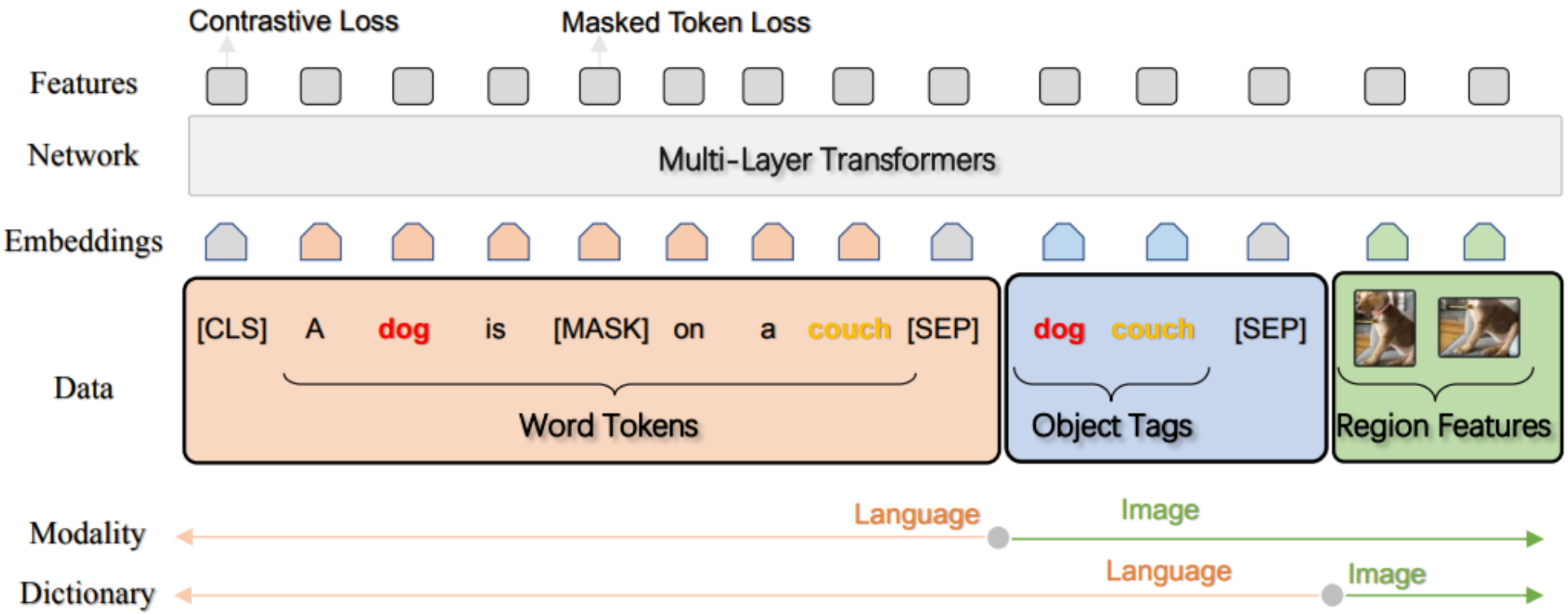
예측 캡션

VGG16 + LSTM : the little dog is playing in the snow
ResNet101 + LSTM : two dogs are running on beach
ResNet101 + Attention : the two brown dogs are playing in grass

OSCAR

Rank	Model	BLEU-4	CIDER	METEOR	SPICE	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	CIDEr-D	Paper	Code	Result	Year	Tags
1	**Oscar**	41.7	140	30.6	24.5						Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks			2020	
2	**SimVLM**	40.6	143.3	33.4	25.4						SimVLM: Simple Visual Language Model Pretraining with Weak Supervision			2021	
3	**X-Transformer**	39.7	132.8	29.5	23.4	59.1	80.9	65.8	51.5	132.8	X-Linear Attention Networks for Image Captioning			2020	

OSCAR


Contrastive Loss
Masked Token Loss
Features
Network
Multi-Layer Transformers
Embeddings
Data
[CLS] A dog is [MASK] on a couch [SEP]
Word Tokens
dog couch [SEP]
Object Tags
Region Features
Modality
Language
Image
Dictionary
Language
Image

Reference Paper.

1. Vinyals, Oriol, et al, “Show and Tell: A Neural Image Caption Generator”, In CVPR, 2015.

2. Xu, Kelvin et al, “Show, Attend and Tell: Neural image caption generation with visual attention”, In ICML, 2015.

3. Li, Xiujun, et al. "Oscar: Object-semantics aligned pre-training for vision-language tasks." In ECCV, 2020.

QnA?