KM-BART

Knowledge Enhanced Multimodal BART

ACL-IJCNLP 2021 main conference

집현전 최신반 8조
이나연(발표자), 민지웅, 임정환

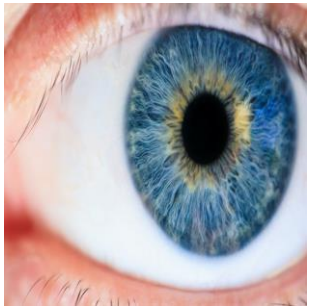
목차

1. Introduction
   a. 논문 선정 이유
   b. Background
   c. Abstract
2. Methodology
   a. Model Architecture
      i. Visual Feature Extractor
      ii. Cross-Modal Encoder
      iii. Decoder
   b. Pre-training Tasks
      i. Knowledge-Based Commonsense Generation
      ii. Attribute prediction and relation
      iii. Masked Language modeling
      iv. masked region modeling
      v. Combining loss
3. Experiments
   a. 설정 및 평가 지표
   b. 결과
4. Conclusion
   a. 결론
   b. 향후 과제

INTRODUCTION

a. 논문 선정 이유
b. Background
c. Abstract

Introduction > 논문 선정 이유

Multimodal Model :

A model processing a number of our senses — visual, auditory, kinesthetic

Introduction > Background

기존 Vision-Language Model:
그림 및 영상을 이해하고 언어로 된 문제를 맞추는 것에 집중

Visual Question Answering 등의 understanding task 수행 가능
But **multimodal generation task 수행 불가**

그림 및 영상을 통해 생성된 Commonsense를
문장으로 표현할 수는 없을까?

Introduction > Background

다양한 언어모델의 활용

- Visual Question Answering (VQA), Image-Text Matching와 같은 Visual-Language task는 visual, text inputs에 대한 복합적인 처리와 이해를 필요로 함
- BERT, GPT-2와 같은 거대한 사전학습 언어모델을 통해 다양한 모델들이 제안

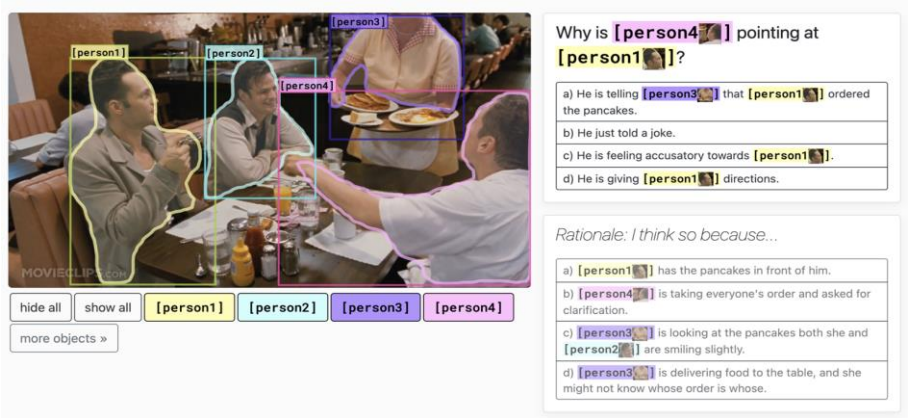
  => Transformer를 backbone으로 image-text pairs를 통해 학습된 denoising autoencoder 사용
- **모두 understanding task에 맞추어 사전학습**

Introduction > Background

Vcr: Visual commonsense reasoning. VCR: Visual Commonsense Reasoning. (n.d.). https://visualcommonsense.com/.

VCR (Visual Commonsense Reasoning)

- From Recognition to Cognition: Visual Commonsense Reasoning (Zellers et al., 2019)
- 이미지 또는 동영상을 표현하는 문장을 맞추는 문제를 해결한 논문
- 주어진 visual data에 대한 understanding에 초점



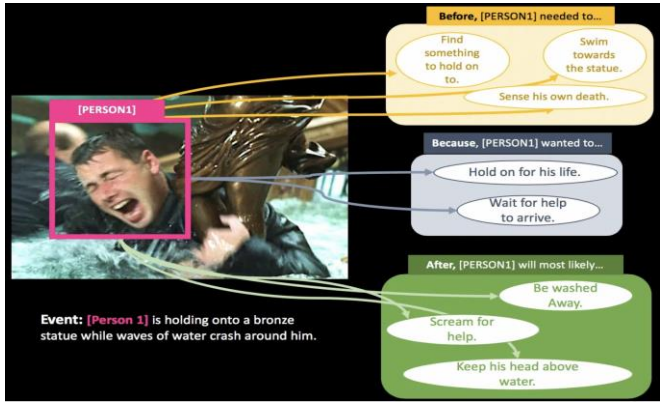
Introduction > Background

VisualComet. (n.d.). *Visual comet: Visual commonsense reasoning in time*. VisualComet. https://visualcomet.xyz/.

VCG (Visual Commonsense Generation)

- Visual-comet: Reasoning about the dynamic context of a still image (Park et al., 2020)에서 소개
- commonsense inferencing

  : 사건 전후로 무슨 일이 일어났는지 & 캐릭터의 현재 의도에 대한 추론을 하는 task



Introduction > Background

Commonsense Knowledge

- 필수적인 실용 지식이나 일반적인 상황이나 사건에 대한 추론을 다룸
- graph-structured representation of knowledge
  - ConceptNet: 일반적인 개념을 나타내는 node와 개념간의 관계를 나타내는 edges를 사용한 지식 그래프
  - ATOMIC: node를 자연어 단락으로, edge를 intent, attribution, effect 등의 관계로 확대
  - 단점: 사람의 개입이 필수적, scalable x, 부정확한 knowledge matching은 모델 성능을 크게 악화
- 문제 극복을 위해 commonsense knowledge graph로 사전학습된 Transformer-based, generative model인 COMET으로 추론한 supervision signal을 이용한 external knowledge를 활용

Introduction > Abstract

“KM-BART”

: Multimodal input으로 부터 commonsense knowledge에 대한 추론을 할 수 있는 Transformer

Based sequence-to-sequence model

Introduction > Abstract

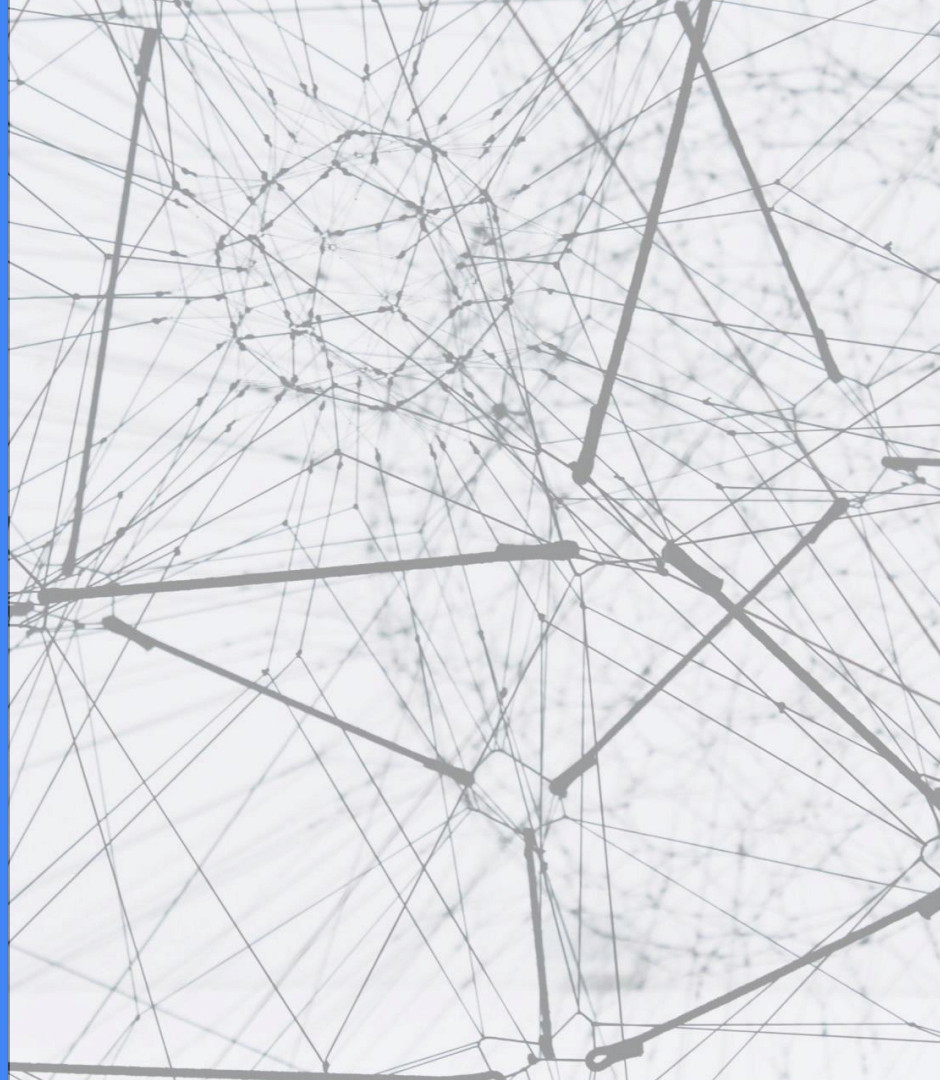
KM-BART

- BART model을 multimodal data에 대해 확장 -> multimodal reasoning
- VCG에 대한 성능 향상을 위해, 외부의 지식 그래프로부터 commonsense knowledge를 통합
- Knowledge-based Commonsense Generation (KCG)라는 새로운 pre-training task 도입 + 기존의 standard한 pretraining 기법 사용
- VCG에 대한 SOTA 달성

Methodology

a. Model Architecture
   i. Visual Feature Extractor
   ii. Cross-Modal Encoder
   iii. Decoder

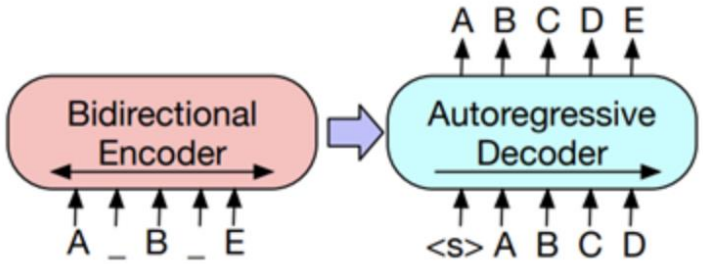
b. Pre-training Task



Methodology > Model Architecture

cross modality input을 위해 BART의 형태를 수정

1. Visual Feature Extractor
2. Cross-Modal Encoder
3. Decoder



Methodology > Model Architecture

1. Visual Feature Extractor

- 입력된 이미지의 특징 임베딩을 얻기 위해 필요
- 사전학습된 Masked R-CNN을 통해 감지된 object에 대한 bounding box 제공 (Region of Interest, RoI)
- RoI의 intermediate representation bounding box의 fixed embedding구함
- 각각의 RoI에 대한 Masked R-CNN의 class distribution은 Masked Region Modeling에 활용

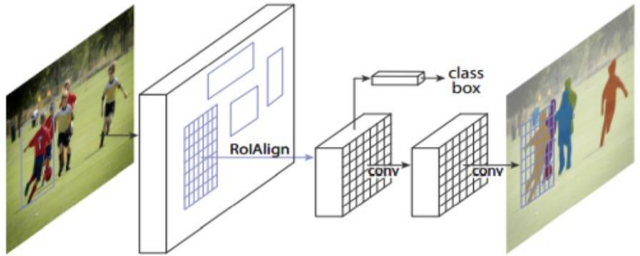


Figure 1. The **Mask R-CNN** framework for instance segmentation.

Methodology > Model Architecture

2. Cross-Modal Encoder

- Starting special token 도입
  - Knowledge-Based Commonsense Generation: <before>, <after>, <intent>
  - Attribution Prediction, Relation Prediction: <region caption>
  - Masked Language Modeling, Masked Region Modeling: <caption>
- multimodal input을 위한 token
  - image: <img>, </img>
  - text: 2개의 textual input set -> events, captions
    - events: 과거, 미래 event, 현재 캐릭터의 의도에 대한 추론을 위해 사용하는 image description. <event>, </event>를 사용.
    - caption: MLM을 위해 <mlm>,</mlm>

Methodology > Model Architecture

3. Decoder

- 디코더는 task generating을 위해 unidirectional visual embedding를 input으로 사용하지 않음
- <img_feat>이라는 실제 visual embedding을 대체하기 위한 token 사용
- MRM, MLM의 masking 위해 <cls> 사용

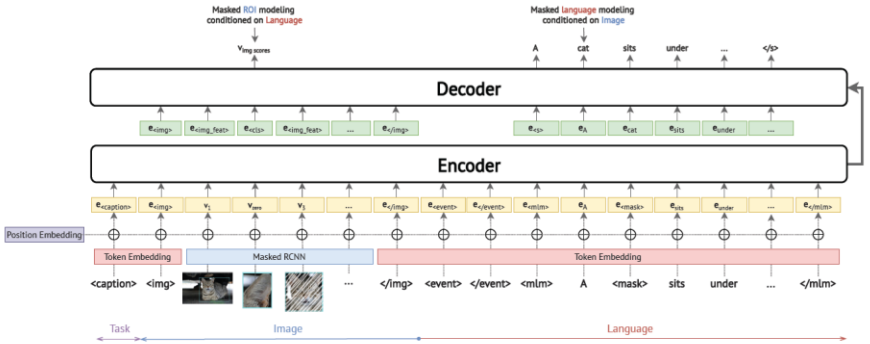


Figure 1: Model architecture. Our model is based on BART. Conditioned on prompts that indicate the task type, such as `<caption>` in the figure, our model can generate texts based on visual and textual inputs from the encoder. Our model uses different special tokens to indicate task types and inform the model of different modalities of input.

Methodology

a. Model Architecture

a. Pre-training Tasks

i. Knowledge-Based Commonsense Generation
ii. Attribute prediction and relation
iii. Marked Language model
iv. masked region modeling
v. Combining loss

Methodology > Pre-training Tasks

Pretraining Datasets

- Pretraining task에 4개의 image-text dataset 사용

1. Conceptual Captions Dataset
2. SBU Dataset
3. Microsoft COCO Dataset
4. Visual Genome

	#images	#sentences
Conceptual Captions (Sharma et al., 2018)	2,683,686	2,683,686
SBU (Ordonez et al., 2011)	780,750	780,750
COCO (Lin et al., 2014)	82,783	414,113
Visual Genome (Krishna et al., 2017)	86,461	4,322,358
Total	3,633,680	8,200,907

Table 1: Statistics of pretraining datasets.

Methodology > Pre-training Tasks

Conceptual Captions

"trees in a winter snowstorm"

"a cartoon illustration of a bear waving and smiling"

"the scenic route through mountain range includes these unbelievably coloured mountains"

"facade of an old shop"

MS COCO

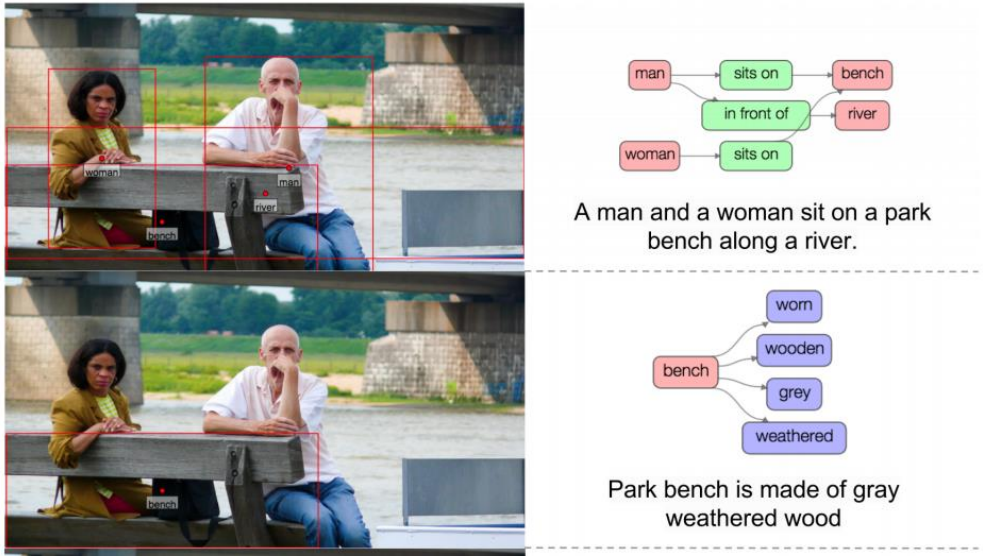
- a cat staring out a window at a bird.
- a car sitting at a window staring at a bird.
- a cat is sitting by a window and watching a bird.
- a cat watching a small bird through a window.
- a cat looking at a bird that is on the other side of a window.

SBU

- Murray in cat bed, Neko not in cat bed

Methodology > Pre-training Tasks

Visual Genome



Methodology > Pre-training Tasks

Knowledge-Based Commonsense Generation

- 사전학습된 COMET으로 external knowledge를 활용
- SBU, COCO 데이터셋에도 COMET을 사용하여 commonsense description 생성
- COMET 데이터셋에서 의도와 이전 행동, 이후 행동을 구분하기 위해 아래와 같이 변경

  xIntent, xWant -> intent

  xNeed -> before

  xReact, xEffect -> after

- 생성된 3.6M개 중 Self-Training Based Data Filtering의 과정을 거쳐 1.46M개의 데이터 최종적으로 사용

Methodology > Pre-training Tasks

Self-Training Based Data Filtering

- VCG dataset과 유사한 예제를 만들기 위해 사용하는 기법
- BART의 parameter들로 초기화한 KM-BART를 VCG dataset에 대하여 fine-tuning
- VCG dataset 내의 모든 commonsense description과 COMET이 생성한 새로운 dataset에 대하여 cross-entropy loss 계산 (Fig. 2)
- CE loss가 3.5 이하인 예제들만 필터링 -> 40%의 example 사용

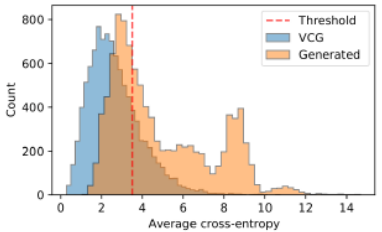


Figure 2: The distribution of the average cross-entropy on 10000 samples in the VCG dataset and our enhanced dataset. For the generated dataset, we can keep the examples of which cross entropy loss is below 3.5.

Methodology > Pre-training Tasks

Self-Training Based Data Filtering

- Filtering된 40%의 example로 생성된 commonsense knowledge dataset으로 KM-BART의 pre-training 진행
- Knowledge Based Commonsense Generation의 loss 식

$$\mathcal{L}_{KCG}(\theta) = -\mathbb{E}_{(W,V)\sim D}\sum_{l=1}^{L}\log(P_\theta(w_l|w_{<l},W,V))$$

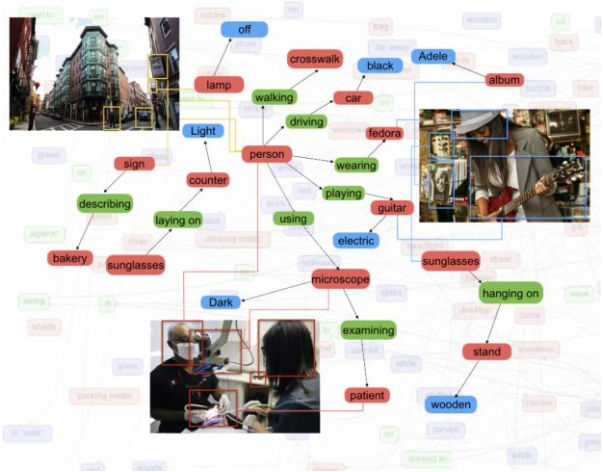
S : {w1, ···, wL}
새롭게 생성된 dataset D의 commonsense description

V, W : 시각 입력, 문자 데이터

Methodology > Pre-training Tasks

Attribute Prediction and Relation

- Visual Genome dataset: 2.3M relationships, 2.8 million attributes
- 이미지에서 다른 object간의 고유한 속성을 학습하기 위해 attribute prediction(AP), relation prediction (RP)를 pre-training task로 활용
- 2개의 task 모두 cross-entropy loss 사용



$$\mathcal{L}_{AP}(\theta) = -\mathbb{E}_{(W,V)\sim D}\sum_{j=1}^{A}\log(P_\theta(L_a(v_j) \mid W, V))$$

$$\mathcal{L}_{RP}(\theta) = -\mathbb{E}_{(W,V)\sim D}\sum_{k=1}^{R}\log(P_\theta(L_r(v_{k_1}, v_{k_2})) \mid W, V))$$

Methodology > Pre-training Tasks

Masked Language Modeling

- 15% token 마스킹
- <mask> token 중 80%: replace

  10%: random replace

  10%: unchanged
- MLM의 loss function: $\mathcal{L}_{MLM}(\theta) = -\mathbb{E}_{(W,V)\sim D} \sum_{m=1}^{M} \log(P_\theta(w_m | w_{\setminus m}, W, V))$

Methodology > Pre-training Tasks

Masked Region Modeling

- 15% 확률로 image region을 sampling해서 feature vector를 masking
- masked vector는 zero vector로 대체
- masked region에 대한 semantic class의 분포를 예측
- Loss function: $\mathcal{L}_{MRM}(\theta) = \mathbb{E}_{(W,V)\sim D} \sum_{n=1}^{N} D_{KL}(p(v_n)||q_\theta(v_n)))$
- output distribution과 Masked R-CNN에 의해 예측된 분포 사이의 KL divergence를 최소화

Methodology > Pre-training Tasks

Combining Loss

- 위에서 언급한 5개의 Loss를 weight와 함께 사용

$$\mathcal{L} = W_{KCG}\mathcal{L}_{KCG} + W_{AP}\mathcal{L}_{AP} + W_{RP}\mathcal{L}_{RP} + W_{MLM}\mathcal{L}_{MLM} + W_{MRM}\mathcal{L}_{MRM}$$

EXPERIMENTS

a. 설정 및 평가 지표
b. 결과

Experiments > 설정 및 평가 지표

설정 및 평가 지표

- VCG task 성능을 알아보기 위해 pretraining task에 대해 ablation study 진행
- 모델 평가를 위해 Visual Commonsense Generation(VCG) dataset 사용

  SIZE - train set: 1174K, valid set: 146k
- 아래의 평가 지표를 사용
  - BLUE-2, METOR, CIDER, Unique, Novel
  - Unique: 고유한 추론 문장 수(생성된 문장 수)를 전체 문장 수로 나눈 값
  - Novel: 학습 데이터에 없는 생성 문장 수 / 총 문장 수

Experiments > 결과

Pretraining Task(s)	Event	BLEU-2	METEOR	CIDER	Unique	Novel
Random init						
w/o pretraining	Y	22.28	14.55	36.49	27.81	29.71
KCG	Y	22.16	14.52	37.06	33.01	31.20
KCG (before filtering)	Y	22.24	14.43	37.08	*33.64*	31.37
AP & RP	Y	22.49	14.64	37.18	28.97	30.28
MLM & MRM	Y	22.44	14.70	37.44	31.16	*31.64*
Full Model	Y	-	-	-	-	-
BART init						
w/o pretraining	Y	22.86	**15.17**	39.13	27.41	28.32
KCG	Y	**23.47**	*15.02*	**39.76**	27.28	27.97
KCG (before filtering)	Y	22.90	14.98	39.01	26.59	27.13
AP & RP	Y	22.93	14.99	39.18	28.06	28.88
MLM & MRM	Y	23.13	14.93	38.75	28.68	28.74
Full Model†	Y	*23.25*	15.01	*39.20*	**35.71**	**32.85**
Random init						
w/o pretraining	N	13.54	10.14	14.87	12.19	24.22
KCG	N	13.64	10.12	15.34	15.95	25.79
KCG (before filtering)	N	13.67	10.13	15.22	16.47	24.97
AP & RP	N	13.83	10.28	15.48	14.60	24.75
MLM & MRM	N	*14.36*	*10.73*	*16.72*	15.86	26.12
Full Model§	N	**14.49**	**10.86**	**17.37**	*16.89*	25.69
BART init						
w/o pretraining	N	8.108	8.673	6.335	4.850	10.55
KCG	N	13.28	10.06	14.17	13.08	25.70
KCG (before filtering)	N	13.29	10.12	13.93	13.51	25.59
AP & RP	N	12.17	9.503	12.49	**20.98**	**29.01**
MLM & MRM	N	13.36	10.22	14.52	15.02	*28.36*
Full Model	N	-	-	-	-	-

	Modalities	Event	BLEU-2	METEOR	CIDER	Unique	Novel
Park et al. (2020)^{a*}	Image+Event+ Place+Person	N	**10.21**	**10.66**	**11.86**	*33.90*	*49.84*
Park et al. (2020)^{b*}	Image	N	6.79	7.13	5.63	26.38	46.80
Ours§	Image	N	*9.04*	*8.33*	*9.12*	**50.75**	**52.92**
Park et al. (2020)^{c*}	Image+Event+ Place+Person	Y	*13.50*	**11.55**	*18.27*	*44.49*	*49.03*
Park et al. (2020)^{d*}	Image+Event	Y	12.52	10.73	16.49	42.83	47.40
Ours†	Image+Event	Y	**14.21**	*11.19*	**21.23**	**57.64**	**58.22**

Models	Event	Before	After	Intent	Total
Park et al. (2020)^{c*}	N	38.7	31.3	30.7	33.3
Ours§	N	**61.3**	**68.7**	**69.3**	**66.7**
Park et al. (2020)^{c*}	Y	48.0	48.0	38.7	44.9
Ours†	Y	**52.0**	**52.0**	**61.3**	**55.1**

Table 6: Human Evaluation results. We compare the inference generated by our best model under the setting of *with event* or *without event*. † and § indicate corresponding models in Table 4. We use Park et al. (2020)^{c*} for both *with event* and *without event* as Park et al. (2020) only release the weights of this model.

CONCLUSION

a. 결론
b. 향후 과제

결론 및 향후 연구

결론

- multimodal input으로부터 commonsense knowledge에 대한 추론을 할 수 있는 Transformer Based sequence-to-sequence model 제안
- Knowledge-Based Commonsense Generation이라는 새로운 pre-training task를 통해서 VCG task의 성능 향상
- 자동적으로 생성한 commonsense description에 대한 self-training 기법 사용
- VCG task에 대한 SOTA 달성, human evaluation으로도 더 좋은 평가

결론 및 향후 연구

향후 연구

- Conceptual Captions Dataset을 사용하여 pre-training set 확장
- 장소와 사람에 대한 정보 활용

감사합니다

Q & A