Scene Graph 생성에서 환각 완화를 위한 Negative Data Augmentation 기법

백종우, 소진현*
대구경북과학기술원

jaden9420@dgist.ac.kr, *jinhyun@dgist.ac.kr

Negative Data Augmentation for Mitigating Hallucination in Scene Graph Generation

Jongwoo Baek, Jinhyun So*
DGIST

요 약

Scene Graph는 이미지 내 객체와 그들 사이의 관계 및 속성을 구조적으로 표현함으로써 정밀한 시각적 이해를 가능케 한다. 그러나 현존하는 Scene Graph 데이터셋은 클래스 편향과 관계의 다양성 부족 문제를 내포하고 있어, 학습된 모델이 존재하지 않는 정보를 생성하는 환각(hallucination) 문제를 유발한다. 본 연구에서는 이러한 문제를 해결하기 위해 부정 데이터 증강(Negative Data Augmentation, NDA) 기법을 제안한다. NDA는 모델이 빈번히 혼동하는 관계 클래스를 기반으로 부정 후보를 생성하고 이를 학습에 활용하는 전략이다. 본 기법은 추가적인 데이터 수집 없이도 학습 데이터의 분포적 한계를 보완하여 모델 일반화 성능을 향상시킨다. 실험 결과, NDA를 Scene Graph 생성 모델에 적용하였을 때 환각 현상을 효과적으로 줄이고 다양한 관계 인식 성능을 개선함을 확인하였다.

Ⅰ. 서 론

Scene Graph는 이미지 내 객체(object)뿐만 아니라 객체 간의 관계(relation)를 구조화된 그래프 형태로 표현하는 자료구조이다. 이 구조는 객체 간의 상호작용을 명시적으로 기술할 수 있어, 이미지 캡셔닝, 비디오 이해, 질의응답 등 다양한 시각-언어 응용 분야에서 평가 벤치마크 등으로 많이 사용된다.

Scene Graph 생성 모델은 이미지로부터 객체, 관계 클래스, 속성을 예측하여 Scene Graph 를 구성하는 모델이다. 이러한 모델의 학습을 위해서는 이미지-그래프 쌍으로 구성된 데이터셋이 필요하다. 그러나 Public Scene Graph 데이터셋들은 수량과 품질 모두에서 한계를 지니고 있으며, 그림 1,2 에서 볼 수 있듯, 관계 클래스의 편향이 심각하다. 특히 그림 2 의 Open Image[1], GQA[2] 데이터셋의 경우 상위 1~2 개의 관계 클래스가 전체 관계의 90% 이상을 차지한다. 이러한 데이터 편향은 모델이 학습 중 특정 패턴에 과적합되어 존재하지 않는 객체나 관계를 생성하는 환각(hallucination) 문제를 야기한다[3].

본 연구는 이러한 환각 문제를 완화하기 위한 부정 데이터 증강(Negative Data Augmentation, NDA) 기법을 제안한다. NDA 는 모델이 자주 혼동하거나 오답을 생성하는 관계 클래스의 특징을 학습에 반영하여, 학습 중 잘못된 일반화를 방지하고 모델의 강건성을 높인다. 본 기법은 추가적인 데이터 수집 없이도 데이터의 표현력을 향상시키는 데이터 중심 해결책으로 작용한다[4].

Ⅱ. 배경 지식

Scene Graph 데이터셋 D 는 이미지-그래프 쌍의 집합으로 정의된다.

$$\mathrm{D} = (I_n, G_n) | n = 1, \ldots, N$$

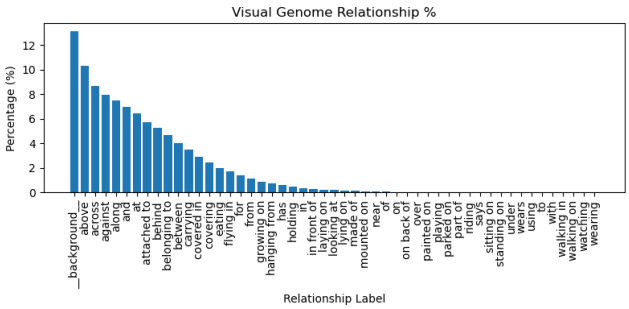

그림 1 Visual Genome 데이터셋의 관계 클래스 분포

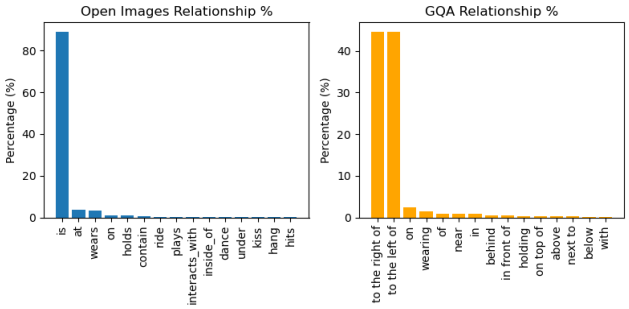

그림 2 Open Image, GQA 데이터셋의 관계 클래스 분포

여기서 I_n 은 이미지, G_n 은 이에 상응하는 그래프를 의미한다. 각 그래프 G_n은 삼중항(triplet) T의 집합이며, 각 triplet $\mathrm{T_i}$은 다음과 같이 정의된다:

$$\mathrm{T_i} = (o_i, r_i, s_i, b_{o_i}, b_{s_i})$$

여기서 $\mathrm{o_i}$ 는 주체(subject), $\mathrm{r_i}$ 는 관계(relation), $\mathrm{s_i}$ 는 객체(object), b_{o_i} 와 b_{s_i} 는 주체와 객체의 바운딩 박스(bounding box)이다. 바운딩 박스는 객체의 위치를 명확히 정의하기 위해 사용되며, 최소 및 최대 좌표 $(x_{min}, y_{min}, x_{max}, y_{max})$ 형태로 표현된다.

Ⅲ. 제안 기법: Negative Data Augmentation

본 연구에서는 사전 학습된 모델(Pre-trained Model)을 기반으로 오분류(class confusion)에 민감하게 반응하는 NDA 기법을 제안한다. 본 기법은 fine-tuning 단계에서 클래스 간 혼동(confusion) 정보를 활용하여, 모델의 일반화 성능을 향상시키고 드문 클래스의 예측 성능을 보완하는 것을 목표로 한다.

우선, 사전 학습된 모델을 이용하여 validation set 전체에 대한 예측 결과를 수집하고, 이로부터 정답 관계 클래스 c 에 대하여 모델이 오분류한 클래스들 간의 confusion matrix $\boldsymbol{M} \in \mathrm{R}^{C\times C}$ 를 구축한다. 여기서 C 는 전체 클래스 수를 나타낸다. 각 행 $\boldsymbol{M}_{c,:}$는 정답 c에 대해 오답으로 예측된 클래스들의 수를 나타내며, 다음과 같이 정규화되어 확률 분포로 사용된다.

$$\widetilde{\boldsymbol{M}}_{(c,j)} = \frac{\boldsymbol{M}_{(c,j)}}{\sum_{j'\neq c}\boldsymbol{M}_{(c,j')}} \ (for\ j \neq c)$$

이를 기반으로 데이터셋의 각 정답 클래스 $c^{(i)}$ 에 대하여, 정규화된 confusion 분포 $\widetilde{\boldsymbol{M}}_{c^{(i)},:}$ 로부터 샘플링을 수행한다. 이를 통해 $c^{(i)}$ 마다 크기 k 의 negative label 집합 $N_{c^{(i)}} \subset \{1,\dots,C\}$ 가 정의된다.

fine-tuning 과정에서는 기존 손실 함수에 샘플링된 negative label 의 예측을 억제하는 negative loss 항을 추가하여 진행된다. 각 입력에 대해 예측된 softmax 확률 벡터 $\mathrm{p} \in R^{C}$에서 negative label $\mathrm{j} \in N_c$에 해당하는 항만을 추출하여, 다음과 같이 손실을 정의한다:

$$\mathrm{L}_{\mathrm{neg}}^{(\mathrm{i})} = \gamma \cdot \frac{1}{\left|\mathrm{N}_{\mathrm{c}}^{(\mathrm{i})}\right|} \sum_{\mathrm{j}\in \mathrm{N}_{c(i)}} -\log\left(1-\mathrm{p}_{\mathrm{j}}^{(\mathrm{i})}\right)$$

여기서 γ는 negative loss 항의 기여 정도를 조절하는 하이퍼 파라미터이며, p_j는 클래스 j에 대한 예측 확률을 의미한다.

최종적으로 사용되는 loss 함수는 모델이 예측한 N_t 개의 클래스에 대해여 기존 모델의 loss 함수 L_{main}에 L_{neg}를 추가하여 다음과 같이 정의된다.

$$L_{total} = L_{main} + L_{\mathrm{neg}}$$

$$= L_{main} + \sum_{i=1}^{N_t} \gamma \cdot \frac{1}{\left|N_c^{(i)}\right|} \sum_{j\in N_{c(i)}} -\log\left(1-p_j^{(i)}\right)$$

이와 같이 손실 함수를 설계하여 모델이 자주 혼동하던 클래스에 대한 예측 확률이 억제되도록 한다.

Ⅳ. 실 험

본 연구는 제안하는 NDA 기법의 유효성을 정량적으로 검증하기 위해 SOTA Scene Graph 생성 모델 RelTR[5]을 기반으로 Visual Genome[6], Open Images[1] 데이터셋에서 실험을 수행하였다.

모든 실험은 동일하게 150 epoch 으로 사전 학습(pre-trained) 모델을 사용하였으며, NDA fine-tuning 10 epoch 을 추가로 진행하였다. 클래스 당 부정 샘플 개수 $\mathrm{k}=1$, negative loss 의 하이퍼 파라미터 $\gamma=0.5$ 를 사용하였다.

성능 평가는 이미지 입력만으로 클래스 레이블, 위치 정보를 추론하는 SGDET(Scene Graph Detection) 환경에서 상위 K 개의 클래스 평균에 대한 관계 예측 정밀도(top-K mean Precision, mP@K)와 관계 예측 재현율(top-K mean Recall, mR@K)을 활용하였다.

Metric	mP@20	mP@50	mP@100
Baseline	**5.159**	2.837	1.691
NDA	5.118	**3.116**	**2.066**

Metric	mR@20	mR@50	mR@100
Baseline	5.685	8.393	**9.869**
NDA	**6.042**	**8.548**	9.867

표 1. Visual Genome 데이터셋에서의 SGDET 결과 (mean Precision 및 mean Recall)

Metric	mP@20	mP@50	mP@100
Baseline	3.806	1.962	1.185
NDA	**3.812**	**2.038**	**1.260**

Metric	mR@20	mR@50	mR@100
Baseline	6.673	**9.378**	**10.719**
NDA	**6.675**	9.122	10.372

표 2. Open Image 데이터셋에서의 SGDET 결과 (mean Precision 및 mean Recall)

실험 결과, NDA 를 적용한 모델은 Baseline 대비 mean Precision 지표에서 전반적으로 향상된 성능을 보였다. Visual Genome 의 경우 mR@20 과 mR@50 구간에서 소폭 증가하였으며, mR@100 에서는 거의 동일한 수준을 유지하였다. 반면 mP@20 에서는 다소 감소가 관찰되었으나, mP@50 과 mP@100 에서는 뚜렷한 개선 효과가 확인되었다. Open Images 에서는 mR@20 이 거의 동일하고, mR@50 과 mR@100 에서 소폭 감소하는 양상을 보였으나, mean Precision 은 모든 cutoff 구간에서 일관되게 향상되었다.

Ⅴ. 결 론

본 연구는 Scene Graph 생성 모델에서 학습 데이터의 클래스 편향으로 인해 발생하는 환각 문제를 완화하기 위해 부정 데이터 증강(Negative Data Augmentation, NDA) 기법을 제안하였다. NDA 는 학습 과정에서 빈번히 잘못 예측되는 관계 클래스를 기반으로 negative label 을 생성하고 이를 fine-tuning 에 활용함으로써, 모델이 잘못된 관계를 반복적으로 예측하는 경향을 억제하여 일반화 능력을 확보한다.

실험 결과, NDA 는 mean Precision 지표에서 뚜렷한 개선 효과를 보였으며, 특히 top-K 평가에서 K 값(상위 예측 개수)이 커질수록 개선 폭이 두드러져 모델이 전체 예측 분포 전반에서 성능을 향상시켰음을 확인할 수 있다. mean Recall 은 데이터셋의 특성에 따라 소폭의 증감은 있었으나 전반적으로 baseline 과 유사한 성능을 유지하였다. 이러한 결과는 추가적인 데이터 수집 없이도 Scene Graph 생성 모델의 일반화 능력을 강화할 수 있음을 보여준다.

ACKNOWLEDGMENT

이 논문은 2025 년도 정부(과학기술정보통신부)의 재원으로 2025 년 혁신거점 인공지능 데이터 융합과제 사업의 지원을 받아 수행된 연구임(S2201-24-1002)

This thesis was conducted with the support of the 2025 innovation base artificial intelligence data convergence project with the funding of the 2025 government (Ministry of Science and ICT) (S2201-24-1002)

참 고 문 헌

[1] Open Images V7 Dataset: Krasin, Ivan, et al. OpenImages: A Public Dataset for Large-Scale Multi-Label and Multi-Class Image Classification. Google AI, 2017.

[2] GQA Dataset: Hudson, Drew A., and Christopher D. Manning. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[3] Li, Wei, et al. "Ppdl: Predicate probability distribution based loss for unbiased scene graph generation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

[4] Sinha, Abhishek, et al. "Negative data augmentation." arXiv preprint arXiv:2102.05113 (2021).

[5] Cong, Yuren, Michael Ying Yang, and Bodo Rosenhahn. "Reltr: Relation transformer for scene graph generation." IEEE Transactions on Pattern Analysis and Machine Intelligence 45.9 (2023): 11169-11183.

[6] Krishna, Ranjay, et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations." *International journal of computer vision* 123.1 (2017): 32-73.