BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation

ICML 2022

Junnan Li, Dongxu Li, Caiming Xiong,
Steven Hoi
Salesforce Research

발제자:
윤예준

https://arxiv.org/abs/2201.12086

01. 연구배경

기존의 VLP 모델의 한계점

- Model perspective
  - 대부분의 모델들은 encoder-based이거나 encoder-decoder based임
  - encoder-based 모델은 text generation task에 약하고,
    encoder-decoder based 모델은 image-text retrieval task에 약함

- Data perspective
  - 최근 SOTA 방법들은 web에서 수집한 image-text 쌍을 사용. 데이터셋의 scaling up으로 인한 성능 향상에도 불구하고 노이즈가 많은 web 텍스트는 학습에 최적이 아님을 보여줌

결과적으로 이러한 한계를 해결하기위해 BLIP을 제안함

- Multimodal mixture of Encoder-Decoder(MED): 3가지 loss를 통해 학습
- Captioning and Filtering (CapFilt): image-text 쌍의 noise를 줄인 dataset 생성 방법

02. 제안 방법

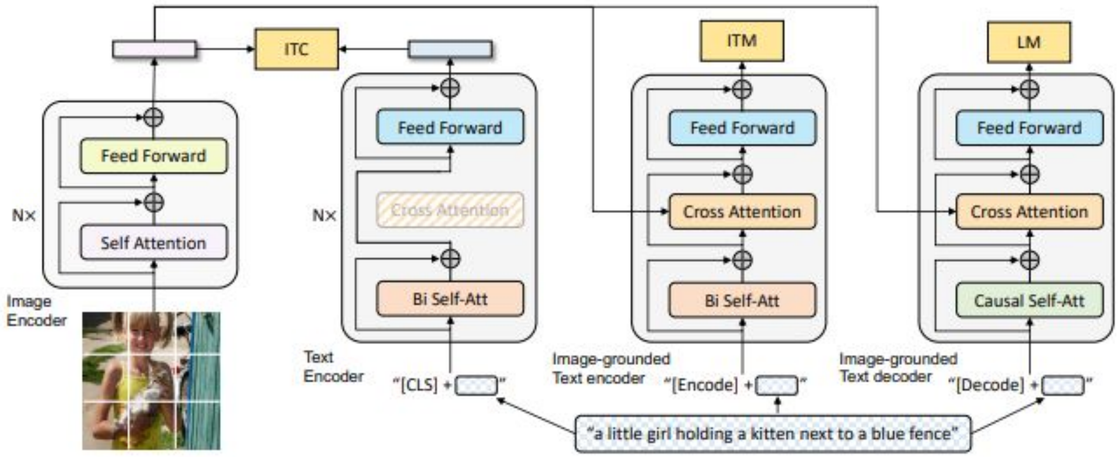
Multimodal mixture of Encoder-Decoder

모델 구조

- Unimodal encoder
    - 이미지와 텍스트를 별도로 인코딩하는 인코더
    - image는 ViT, text는 BERT를 사용
    - ViT, BERT 모두 CLS token을 사용하는 구조를 이용

- Image-grounded text encoder
    - cross attention을 통해 visual information이 추가 됨
    - Encode token이 text에 추가되며 이는 image-text pair의 multimodal representation으로 사용됨

- Image-grounded text decoder
    - Bi Self attention대신 causal self-attention이 사용됨
    - Decode token이 text에 추가되며 이는 문장의 시작을 알리는데 사용됨.



02. 제안 방법

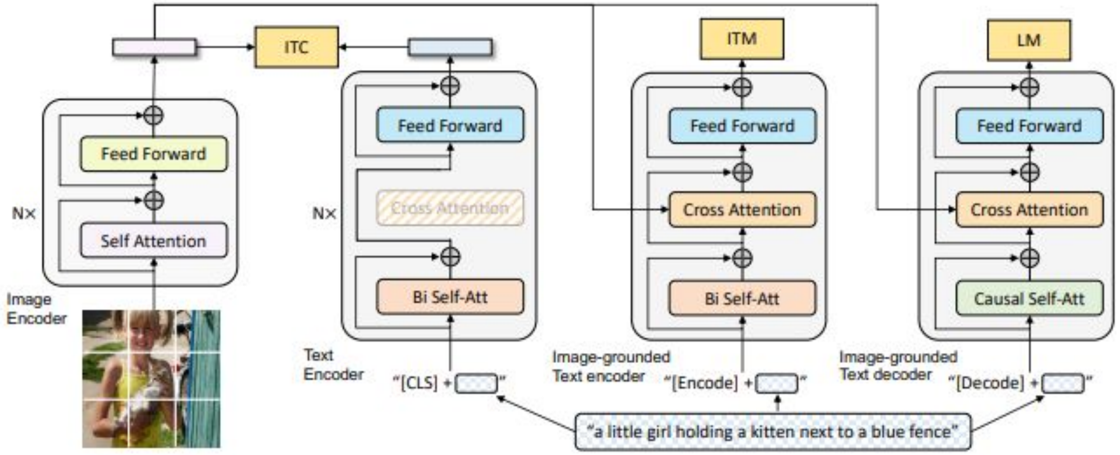
Multimodal mixture of Encoder-Decoder

Pre-training Objectives

- Image-Text Contrastive Loss (ITC)
    - visual, text transformer의 feature space를 align하기 위함
    - image2text, text2image 평균을 loss 사용하여 학습

- Image-Text Matching Loss (ITM)
    - image-text multimodal representation을 잘 포착하도록 학습
    - ITM(linear layer)를 사용하여 이진 분류 작업을 통해 image-text pair가 positive인지 negative인지 예측

- Language Modeling Loss (LM)
    - image가 주어졌을 때 text description을 생성하기 위함
    - autoregressive manner이며 crossentropy loss을 최적화하도록 학습



Align before Fuse: Vision and Language Representation Learning with Momentum Distillation

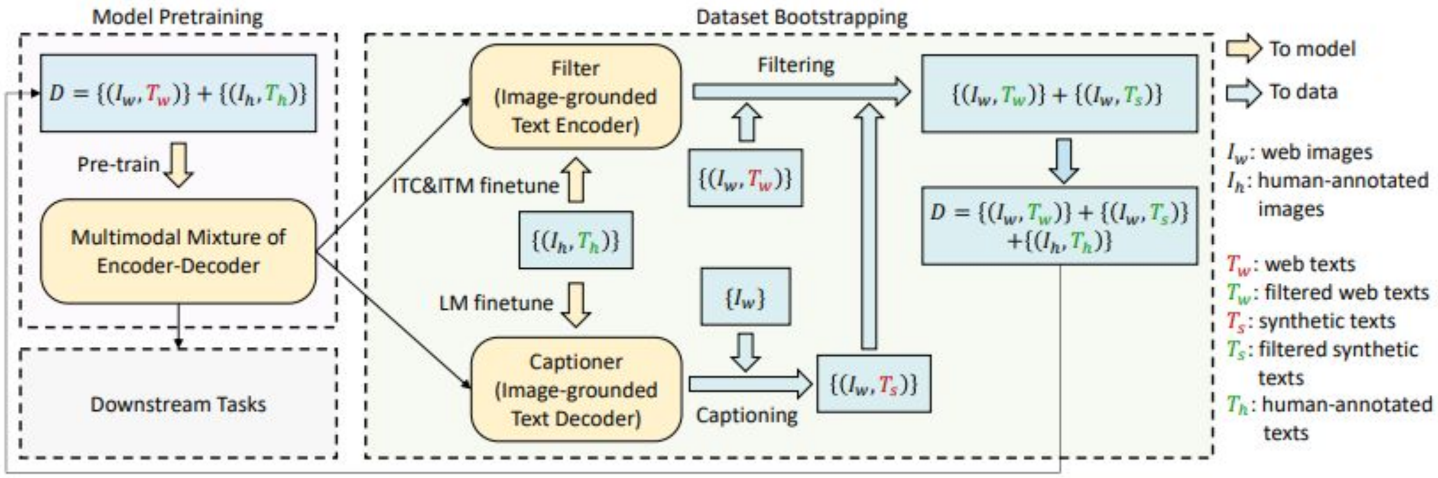
02. 제안 방법

CapFilt

text corpus의 질을 향상시키기 위한 방법

Filter: noisy한 image-text pair 제거
Captioner: web image가 주어지면 caption 생성

- Filter와 Captioner는 모두 MED로 initialized
- human annotated dataset으로 Filter와 Captioner finetune



03. 실험 결과

Effect of CapFilt

- 모두 Bootstrap이 존재하는 경우의 성능이 좋은 것을 알 수 있음

Pre-train dataset	Bootstrap C	Bootstrap F	Vision backbone	Retrieval-FT (COCO) TR@1	Retrieval-FT (COCO) IR@1	Retrieval-ZS (Flickr) TR@1	Retrieval-ZS (Flickr) IR@1	Caption-FT (COCO) B@4	Caption-FT (COCO) CIDEr	Caption-ZS (NoCaps) CIDEr	Caption-ZS (NoCaps) SPICE
COCO+VG +CC+SBU (14M imgs)	✗	✗	ViT-B/16	78.4	60.7	93.9	82.1	38.0	127.8	102.2	13.9
	✗	✓$_B$		79.1	61.5	94.1	82.8	38.1	128.2	102.7	14.0
	✓$_B$	✗		79.7	62.0	94.4	83.6	38.4	128.9	103.4	14.2
	✓$_B$	✓$_B$		80.6	63.1	94.8	84.9	38.6	129.7	105.1	14.4
COCO+VG +CC+SBU +LAION (129M imgs)	✗	✗	ViT-B/16	79.6	62.0	94.3	83.6	38.8	130.1	105.4	14.2
	✓$_B$	✓$_B$		81.9	64.3	96.0	85.0	39.4	131.4	106.3	14.3
	✓$_L$	✓$_L$		81.2	64.1	96.0	85.5	39.7	133.3	109.6	14.7
	✗	✗	ViT-L/16	80.6	64.1	95.1	85.5	40.3	135.5	112.5	14.7
	✓$_L$	✓$_L$		82.4	65.1	96.7	86.7	40.4	136.7	113.2	14.8

Table 1. Evaluation of the effect of the captioner (C) and filter (F) for dataset bootstrapping. Downstream tasks include image-text retrieval and image captioning with finetuning (FT) and zero-shot (ZS) settings. TR / IR@1: recall@1 for text retrieval / image retrieval. ✓$_{B/L}$: captioner or filter uses ViT-B / ViT-L as vision backbone.

03. 실험 결과

Diversity is Key for Synthetic Captions

- Synthetic caption generation method 방식 3가지 비교
  - Beam: 해당 시점에서 가장 유망한 빔 개수만큼(이하 k) 골라서 탐색하여 sampling하는 방법
  - Nucleus: 누적 확률이 top-p가 될 때까지의 단어만 sampling 하는 확률적 디코딩 방법
    threshold p(p=0.9)

  - Nucleus의 noise 비율이 가장 크나 성능이 제일 좋은 것을 볼 수 있음
    => Nucleus sampling이 더 다양한 캡션을 새성하여 모델이 활용할 수 있는 새로운 정보를 더 많이 포함하기 있기 때문 (가설)
  - Beam은 데이터셋에서 흔히 볼 수 있는 안전한 캡션을 생성하는 경향이 있으므로 추가 지식이 적음

Generation method	Noise ratio	Retrieval-FT (COCO) TR@1	Retrieval-FT (COCO) IR@1	Retrieval-ZS (Flickr) TR@1	Retrieval-ZS (Flickr) IR@1	Caption-FT (COCO) B@4	Caption-FT (COCO) CIDEr	Caption-ZS (NoCaps) CIDEr	Caption-ZS (NoCaps) SPICE
None	N.A.	78.4	60.7	93.9	82.1	38.0	127.8	102.2	13.9
Beam	19%	79.6	61.9	94.1	83.1	38.4	128.9	103.5	14.2
Nucleus	25%	80.6	63.1	94.8	84.9	38.6	129.7	105.1	14.4

Table 2. Comparison between beam search and nucleus sampling for synthetic caption generation. Models are pre-trained on 14M images.

03. 실험 결과

Parameter Sharing and Decoupling

- SA를 제외하고 parameter를 공유하는 것이 성능 가장 좋은 것을 알 수 있음

Layers shared	#parameters	Retrieval-FT (COCO) TR@1	Retrieval-FT (COCO) IR@1	Retrieval-ZS (Flickr) TR@1	Retrieval-ZS (Flickr) IR@1	Caption-FT (COCO) B@4	Caption-FT (COCO) CIDEr	Caption-ZS (NoCaps) CIDEr	Caption-ZS (NoCaps) SPICE
All	224M	77.3	59.5	93.1	81.0	37.2	125.9	100.9	13.1
All except CA	252M	77.5	59.9	93.1	81.3	37.4	126.1	101.2	13.1
All except SA	252M	78.4	60.7	93.9	82.1	38.0	127.8	102.2	13.9
None	361M	78.3	60.5	93.6	81.9	37.8	127.4	101.8	13.9

Table 3. Comparison between different parameter sharing strategies for the text encoder and decoder during pre-training.

03. 실험 결과

Parameter Sharing and Decoupling

- Captioner와 filter가 parameter 공유 시 성능이 저하되는 것을 볼 수 있음
- Confirmation bias로 인해 저하가 됨
- Parameter 공유로 인해 captioner가 생성한 노이즈 캡션이 필터에 걸러질 가능성이 낮아지기 때문

Captioner & Filter	Noise ratio	Retrieval-FT (COCO) TR@1	Retrieval-FT (COCO) IR@1	Retrieval-ZS (Flickr) TR@1	Retrieval-ZS (Flickr) IR@1	Caption-FT (COCO) B@4	Caption-FT (COCO) CIDEr	Caption-ZS (NoCaps) CIDEr	Caption-ZS (NoCaps) SPICE
Share parameters	8%	79.8	62.2	94.3	83.7	38.4	129.0	103.5	14.2
Decoupled	25%	80.6	63.1	94.8	84.9	38.6	129.7	105.1	14.4

Table 4. Effect of sharing parameters between the captioner and filter. Models are pre-trained on 14M images.

03. 실험 결과

Comparison with State-of-the-arts

Image-text retrieval

Method	Pre-train # Images	COCO (5K test set) TR R@1	R@5	R@10	IR R@1	R@5	R@10	Flickr30K (1K test set) TR R@1	R@5	R@10	IR R@1	R@5	R@10
UNITER (Chen et al., 2020)	4M	65.7	88.6	93.8	52.9	79.9	88.0	87.3	98.0	99.2	75.6	94.1	96.8
VILLA (Gan et al., 2020)	4M	-	-	-	-	-	-	87.9	97.5	98.8	76.3	94.2	96.8
OSCAR (Li et al., 2020)	4M	70.0	91.1	95.5	54.0	80.8	88.5	-	-	-	-	-	-
UNIMO (Li et al., 2021b)	5.7M	-	-	-	-	-	-	89.4	98.9	99.8	78.0	94.2	97.1
ALIGN (Jia et al., 2021)	1.8B	77.0	93.5	96.9	59.9	83.3	89.8	95.3	99.8	100.0	84.9	97.4	98.6
ALBEF (Li et al., 2021a)	14M	77.6	94.3	97.2	60.7	84.3	90.5	95.9	99.8	100.0	85.6	97.5	98.9
BLIP	14M	80.6	95.2	97.6	63.1	85.3	91.1	96.6	99.8	**100.0**	87.2	97.5	98.8
BLIP	129M	**81.9**	95.4	97.8	**64.3**	85.7	91.5	**97.3**	**99.9**	**100.0**	87.3	97.6	**98.9**
$BLIP_{CapFilt-L}$	129M	81.2	**95.7**	**97.9**	64.1	**85.8**	**91.6**	97.2	**99.9**	**100.0**	**87.5**	**97.7**	**98.9**
$BLIP_{ViT-L}$	129M	82.4	95.4	97.9	65.1	86.3	91.8	97.4	99.8	99.9	87.6	97.7	99.0

Table 5. Comparison with state-of-the-art image-text retrieval methods, finetuned on COCO and Flickr30K datasets. $BLIP_{CapFilt-L}$ pre-trains a model with ViT-B backbone using a dataset bootstrapped by captioner and filter with ViT-L.

Method	Pre-train # Images	Flickr30K (1K test set) TR R@1	R@5	R@10	IR R@1	R@5	R@10
CLIP	400M	88.0	98.7	99.4	68.7	90.6	95.2
ALIGN	1.8B	88.6	98.7	99.7	75.7	93.8	96.8
ALBEF	14M	94.1	99.5	99.7	82.8	96.3	98.1
BLIP	14M	94.8	99.7	**100.0**	84.9	96.7	98.3
BLIP	129M	**96.0**	**99.9**	**100.0**	85.0	**96.8**	98.6
$BLIP_{CapFilt-L}$	129M	**96.0**	**99.9**	**100.0**	**85.5**	**96.8**	**98.7**
$BLIP_{ViT-L}$	129M	96.7	100.0	100.0	86.7	97.3	98.7

Table 6. Zero-shot image-text retrieval results on Flickr30K.

03. 실험 결과

Comparison with State-of-the-arts

Image Captioning

Method	Pre-train #Images	NoCaps validation in-domain C	in-domain S	near-domain C	near-domain S	out-domain C	out-domain S	overall C	overall S	COCO Caption Karpathy test B@4	C
Enc-Dec (Changpinyo et al., 2021)	15M	92.6	12.5	88.3	12.1	94.5	11.9	90.2	12.1	-	110.9
VinVL† (Zhang et al., 2021)	5.7M	103.1	14.2	96.1	13.8	88.3	12.1	95.5	13.5	38.2	129.3
$\text{LEMON}_{\text{base}}$† (Hu et al., 2021)	12M	104.5	14.6	100.7	14.0	96.7	12.4	100.4	13.8	-	-
$\text{LEMON}_{\text{base}}$† (Hu et al., 2021)	200M	107.7	14.7	106.2	14.3	107.9	13.1	106.8	14.1	**40.3**	**133.3**
BLIP	14M	111.3	15.1	104.5	14.4	102.4	13.7	105.1	14.4	38.6	129.7
BLIP	129M	109.1	14.8	105.8	14.4	105.7	13.7	106.3	14.3	39.4	131.4
$\text{BLIP}_{\text{CapFilt-L}}$	129M	**111.8**	**14.9**	**108.6**	**14.8**	**111.5**	**14.2**	**109.6**	**14.7**	39.7	**133.3**
$\text{LEMON}_{\text{large}}$† (Hu et al., 2021)	200M	116.9	15.8	113.3	15.1	111.3	14.0	113.4	15.0	40.6	135.7
$\text{SimVLM}_{\text{huge}}$ (Wang et al., 2021)	1.8B	113.7	-	110.9	-	115.2	-	112.2	-	40.6	143.3
$\text{BLIP}_{\text{ViT-L}}$	129M	114.9	15.2	112.1	14.9	115.3	14.4	113.2	14.8	40.4	136.7

Table 7. Comparison with state-of-the-art image captioning methods on NoCaps and COCO Caption. All methods optimize the cross-entropy loss during finetuning. C: CIDEr, S: SPICE, B@4: BLEU@4. $\text{BLIP}_{\text{CapFilt-L}}$ is pre-trained on a dataset bootstrapped by captioner and filter with ViT-L. VinVL† and LEMON† require an object detector pre-trained on 2.5M images with human-annotated bounding boxes and high resolution (800×1333) input images. $\text{SimVLM}_{\text{huge}}$ uses 13× more training data and a larger vision backbone than ViT-L.

03. 실험 결과

Comparison with State-of-the-arts

Visual Question Answering (VQA) & Natural Language Visual Reasoning ($NLVR^2$)

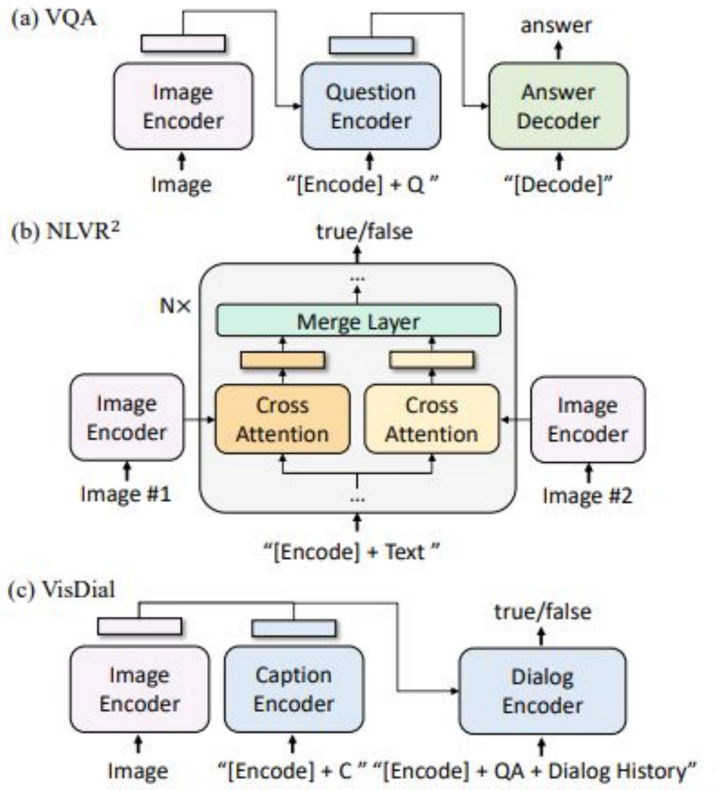


Figure 5. Model architecture for the downstream tasks. Q: question; C: caption; QA: question-answer pair.

Method	Pre-train #Images	VQA test-dev	VQA test-std	$NLVR^2$ dev	$NLVR^2$ test-P
LXMERT	180K	72.42	72.54	74.90	74.50
UNITER	4M	72.70	72.91	77.18	77.85
VL-T5/BART	180K	-	71.3	-	73.6
OSCAR	4M	73.16	73.44	78.07	78.36
SOHO	219K	73.25	73.47	76.37	77.32
VILLA	4M	73.59	73.67	78.39	79.30
UNIMO	5.6M	75.06	75.27	-	-
ALBEF	14M	75.84	76.04	82.55	83.14
$SimVLM_{base}$†	1.8B	77.87	78.14	81.72	81.77
BLIP	14M	77.54	77.62	**82.67**	82.30
BLIP	129M	78.24	78.17	82.48	**83.08**
$BLIP_{CapFilt-L}$	129M	**78.25**	**78.32**	82.15	82.24

Table 8. Comparison with state-of-the-art methods on VQA and $NLVR^2$. ALBEF performs an extra pre-training step for $NLVR^2$. SimVLM† uses 13× more training data and a larger vision backbone (ResNet+ViT) than BLIP.

03. 실험 결과

Comparison with State-of-the-arts

Visual Dialog (VisDial) & Zero-shot Transfer to Video-Language Tasks

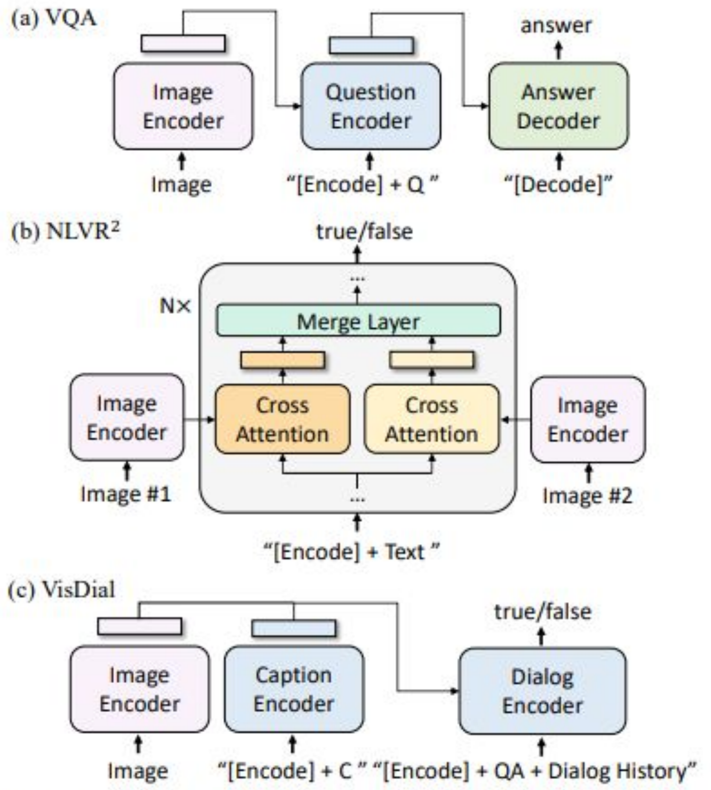


Figure 5. Model architecture for the downstream tasks. Q: question; C: caption; QA: question-answer pair.

Method	MRR↑	R@1↑	R@5↑	R@10↑	MR↓
VD-BERT	67.44	54.02	83.96	92.33	3.53
VD-ViLBERT†	69.10	55.88	85.50	93.29	3.25
BLIP	**69.41**	**56.44**	**85.90**	**93.30**	**3.20**

Table 9. Comparison with state-of-the-art methods on VisDial v1.0 validation set. VD-ViLBERT† (Murahari et al., 2020) pre-trains ViLBERT (Lu et al., 2019) with additional VQA data.

Method	R1↑	R5↑	R10↑	MdR↓
zero-shot				
ActBERT (Zhu & Yang, 2020)	8.6	23.4	33.1	36
SupportSet (Patrick et al., 2021)	8.7	23.0	31.1	31
MIL-NCE (Miech et al., 2020)	9.9	24.0	32.4	29.5
VideoCLIP (Xu et al., 2021)	10.4	22.2	30.0	-
FiT (Bain et al., 2021)	18.7	39.5	51.6	10
BLIP	**43.3**	**65.6**	**74.7**	**2**
finetuning				
ClipBERT (Lei et al., 2021)	22.0	46.8	59.9	6
VideoCLIP (Xu et al., 2021)	30.9	55.4	66.8	-

Table 10. Comparisons with state-of-the-art methods for text-to-**video** retrieval on the 1k test split of the MSRVTT dataset.

Method	MSRVTT-QA	MSVD-QA
zero-shot		
VQA-T (Yang et al., 2021)	2.9	7.5
BLIP	19.2	35.2
finetuning		
HME (Fan et al., 2019)	33.0	33.7
HCRN (Le et al., 2020)	35.6	36.1
VQA-T (Yang et al., 2021)	41.5	46.3

Table 11. Comparisons with state-of-the-art methods for **video** question answering. We report top-1 test accuracy on two datasets.

04. 결론

- 다양한 downstream VL task에서 좋은 성능을 보이는 새로운 프레임워크 BLIP 제안
- BLIP을 통해 기존 VLP 한계점 해결

감사합니
다.