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

Junnan Li, Dongxu Li, Caiming Xiong, Steven Hoi

ICML 2022

Presenter: Shivang Chopra, Brandon Ho, Jiachun Zhang
Outline

• Problem Statement
• Related Works
• Approach
• Experiments & Results
• Limitations, Societal Implications
• Summary of Strengths, Weaknesses, Relationship to Other Papers
Problems With Current Vision-Language Pretraining

• No unified architecture for multi-task vision-language pre-training
  o Encoder only models
    ▪ CLIP, ALBEF
    ▪ Not directly applicable to text generation tasks
  o Encoder-Decoder models
    ▪ VL-T5, SimVLM
    ▪ Can't perform image-text retrieval
• Noisy image captions are suboptimal for vision-language pretraining
• High computational costs during pre-training
• Pre-trained encoders experience catastrophic forgetting
Prior Methods

• Frozen
  - Unified Encoders
    - Text-Grounded Image Encoder
  - Catastrophic Forgetting
    - Frozen LLMs
  - High computational cost during pre-training
    - Few-shot learning + Fast concept Binding
  - Model performance very far from SOTA (Underfitting)

• Flamingo
  - Unified Encoders
    - Vision-Grounded Text Encoder
      - Perceiver Sampler + Gated Cross Attention Dense Layer
  - Catastrophic Forgetting
    - Gated Cross Attention Dense Layers
  - Non-trivial model performance without images
    - Text encoder is not grounded well into images (possibly due to simplicity of Gated Cross Attention Layers)
Proposed Solutions

• BLIP: Bootstrapping Language Image Pretraining
  o Multimodal mixture of Encoder-Decoder (MED)
    ▪ Unimodal encoder
    ▪ Image-grounded text encoder
    ▪ Image-grounded text decoder
  o Captioning and Filtering (CapFilt)

• BLIP 2: BLIP with Frozen Unimodal Models
  o Modality bridge with Q-Former
  o Frozen Unimodal Encoders
    ▪ Compute Efficient
    ▪ Very less forgetting
Unified Vision-Language Pre-Training for Image Captioning and VQA

- Prior attempts at unified architecture
  - VideoBERT and CBT pre-trained only the encoder and not the decoder which leads to cross-modal discrepancy during finetuning.

- Different Self-Attention Masks for each task
  - Bidirectional Prediction: All image regions + Words to left and right
  - Seq2seq: All image regions + Words to left

| Type                        | Method                          | Domain | Architecture                                      | Downstream Tasks                                           |
|-----------------------------|---------------------------------|--------|--------------------------------------------------|----------------------------------------------------------|
| Understanding-based only    | LXMERT (Tan and Bansal 2019),   | Image  | Single-stream or two stream Transformer         | Visual question answering, Visual commonsense reasoning, Image retrieval, Grounding referring expressions |
|                             | ViLBERT (Lu et al. 2019),       |        |                                                  |                                                          |
|                             | UNITER (Chen et al. 2019),      |        |                                                  |                                                          |
|                             | VisualBERT (Li et al. 2019b),    |        |                                                  |                                                          |
|                             | B2T2 (Alberti et al. 2019),     |        |                                                  |                                                          |
|                             | Unicoder-VL (Li et al. 2019a),  |        |                                                  |                                                          |
|                             | VL-BERT (Su et al. 2019)        |        |                                                  |                                                          |
| Generation-based and        | VideoBERT (Sun et al. 2019b)    | Video  | Single-stream Transformer+ Masked Transformer   | Zero-shot action classification, Video captioning         |
| understanding-based         | (Zhou et al. 2018)              |        | (Zhou et al. 2018)                              |                                                          |
| Generation-based and        | CBT (Sun et al. 2019a)          | Video  | Two-stream Transformer encoder+ Transformer decoder | Action anticipation, Video captioning                     |
| understanding-based         | Our VLP                         | Image  | Single unified encoder-decoder                  | Visual question answering, Image captioning               |
|                             |                                 |        |                                                  |                                                          |
Unifying Vision-and-Language Tasks via Text Generation

- **Methods**: VL-T5, VL-BART
- **Task**: Multimodal conditional text generation
- **Reasoning skills required by multiple tasks overlap**
- **To learn a new task:**
  - Rewrite its input and output in text
- **All tasks are modelled as generative tasks**
VisualGPT: Data-Efficient Adaptation of Pretrained Language Models for Image Captioning

- Encoder: Randomly initialized
- Decoder: Initialized with LLM weights
- Self-Resurrecting Encoder-Decoder Attention
  - Self-Resurrecting Activation Unit (SRAU)
- Balances input from visual encoder and linguistic input from previous decoder layer
Plug-and-Play VQA: Zero-Shot VQA by Conjoining Large Pretrained Models with Zero Training

- **Image-Question Matching Module**
  - Use GradCam to identify image patches relevant to the question

- **Image Captioning Module**
  - Generate diverse captions

- **Question Answering Module**
  - Generate answers using question and captions
BLIP Model Architecture

1. Unimodal Encoder (Image & Text Encoder)
2. Image-Grounded Text Encoder
3. Image-Grounded Text Decoder

ViT-B/16 or ViT-L/16
BLIP Text Encoder/Decoder Parameter Sharing

Sharing parameters reduces model size
Pre-training Loss Function (ITC & ITM)

Image-Text Contrastive Loss (ITC) [1]
Align the feature space of the visual transformer and text transformer
Encourage positive image-text pairs, having similar representation

Image-Text Matching Loss (ITM) [1]
Capture the fine-grained alignment b/w vision & language
Binary classify of whether an image-text pair matched or unmatched

Bi-directional self-attention block -> build representations of current tokens

[1] Li, et al. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation, NeurIPS21
Pre-training Loss Function (LM)

Language Modeling Loss (LM)
Generate text captions given an image
Maximize the likelihood of the caption in an autoregressive manner
Enables generalization capability of converting vision into coherent caption
Causal self-attention block -> predict next tokens
Captioning and Filtering (CapFilt)

Create synthetic captions & filter out noisy captions

Gap: Image-text pairs from web data are noisy -> suboptimal performance in VLM

Captioner: Decoder Finetuned with LM Loss.
Filter: Encoder Finetuned with ITC and ITM Loss. Filter text (both from web and synthetic data) if ITM is 0.
Diversity is The Key

Nucleus Sampling (Stochastic Search) for Caption Selection

Each token is sampled from a set of tokens whose cumulative probability mass exceeds a threshold $p (0.9)$.

Noisier data, but better performance.

More diverse and surprise data for better robustness?

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

*Table 2. Comparison between beam search and nucleus sampling for synthetic caption generation. Models are pre-trained on 14M images.*
Pre-training & Finetuning Procedure

**Pre-training & Finetuning Procedure**

- **web-data & human-annotated data**
  - Pretrained Unimodal Encoder
  - Multimodal Mixture of Encoder-Decoder
  - Model Pretraining
    - $D = \{(i_w, T_w)\} + \{(i_h, T_h)\}$
  - Pre-train
  - ITC&ITM finetune
  - LM finetune

**Finetuned CapFilt**

- Filter (Image-grounded Text Encoder)
- Captioner (Image-grounded Text Decoder)
- Dataset Bootstrapping
  - Filtering
  - $\{(i_w, T_w)\}$
  - $\{(i_h, T_h)\}$

- $D = \{(i_w, T_w)\} + \{(i_h, T_h)\}$
- $\{(i_w, T_w)\}$
- $\{(i_h, T_h)\}$
- $\{(i_w, T_w)\}$
- $\{(i_h, T_h)\}$

- Filtered web-data & synthetic data

---

$I_w$: web images
$I_h$: human-annotated images
$T_w$: web texts
$T_w$: filtered web texts
$T_s$: synthetic texts
$T_s$: filtered synthetic texts
$T_h$: human-annotated texts
CapFilt Demonstration

$T_w$: “from bridge near my house”

$T_s$: “a flock of birds flying over a lake at sunset”

$T_w$: “in front of a house door in Reichenfels, Austria”

$T_s$: “a potted plant sitting on top of a pile of rocks”

$T_w$: “the current castle was built in 1180, replacing a 9th century wooden castle”

$T_s$: “a large building with a lot of windows on it”

$T_w$ - web text

$T_s$ - synthetic text

Green: Accepted by filter

Red: Rejected by the filter

Store the filtered caption and image pair in the dataset
Experiments and Results

• Image Transformer
  ViT-B/16 or ViT-L/16 architecture pre-trained on ImageNet

• Text Transformer
  BERT\text{\textsubscript{base}}

• Pre-training Dataset (14M, 129M Images)
  2 human-annotated datasets: COCO & Visual Genome
  3 web datasets: Conceptual Captions, Conceptual 12M, SBU captions

• Image Resolution (384 x 384)
The use of captioning or/and filtering improves the performance across all tasks. Performance scales with more data (14M -> 129M) and more parameters (ViT-B/16 -> ViT-L/16).
Nucleus Sampling / Top-\(p\) Sampling

- Deterministic Methods, such as beam search, use maximum likelihood -> bland, non-diverse, and incoherent set of captions
- Nucleus Sampling (Stochastic Decoding Method) [1]
  
  Generate the smallest vocabulary set such that the cumulative probability of the next tokens is greater than a threshold \(p\)

\[
\sum_{x \in V(p)} P(x | x_{1:i-1}) \geq p.
\]

Sample from that set.

For BLIP, \(p = 0.9\)

Example of Nucleus Search [2]
# Beam Search vs Nucleus Search

## Performance: Nucleus Search > Beam Search

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

## Noise Ratio (filtered caption ratio): Nucleus Search > Beam Search

**Reasoning:** Despite of the higher noise ratio, nucleus search generates diverse set of captions, containing rich information, while beam search outputs safe but less-informative captions.
Sharing Parameters Results

| Layers shared | #parameters | Retrieval-FT (COCO) | Retrieval-ZS (Flickr) | Caption-FT (COCO) | Caption-ZS (NoCaps) |
|---------------|-------------|---------------------|-----------------------|-------------------|---------------------|
|               |             | TR@1 | IR@1 | TR@1 | IR@1 | B@4 | CIDEr | CIDEr | 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  |

Sharing all parameters except for self-attention (SA) in the text encoder and decoder -> best performance & reduce model size

**Reasoning:** Sharing SA layer would degrade performance due to conflicting objective between the encoder and decoder.
Comparison Tests

Tasks
• Image Text Retrieval*
• Image Captioning*
• Visual Question Answering*
• Natural language Visual Reasoning
• Visual Dialog
• Zero-shot Transfer to Video-Language Tasks

Dataset
• COCO and Flickr30K*
• NoCaps and COCO*
• Not Specified (Self-Made)
• Not Specified
• Not Specified
• Not Specified

*Also done in the BLIP2 paper.
Image Text Retrieval

- Evaluate BLIP for both image-to-text retrieval (TR) and text-to-image retrieval (IR), fine tuned on image-text contrastive loss (ITC) and image text matching loss (ITM).
- Select k candidates based on the image-text feature similarity, and then rank the selected candidates based on their pairwise ITM scores.

| Method              | Pre-train # Images | COCO (5K test set) | Flickr30K (1K test set) |
|---------------------|--------------------|--------------------|-------------------------|
|                     |                    | TR @1  | TR @5  | TR @10 | IR @1  | IR @5  | IR @10 | TR @1  | TR @5  | TR @10 | IR @1  | IR @5  | IR @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<sub>CapFilt-L</sub> | 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<sub>ViT-L</sub>    | 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<sub>CapFilt-L</sub> pre-trains a model with ViT-B backbone using a dataset bootstrapped by captioner and filter with ViT-L.
**Image Captioning**

- Similar as Simvlm (Wang et al. (2021)), add a prompt “a picture of” at the beginning of each caption, which leads to slightly better results.

| Method                          | Pre-train #Images | Pre-train | NoCaps validation | COCO Caption Karpathy test |
|--------------------------------|-------------------|-----------|-------------------|----------------------------|
|                                |                   | in-domain | near-domain | out-domain | overall | B@1 | C     |
| Enc-Dec (Changpinyo et al., 2021) | 15M               |           |             |           |         | -    | 110.9 |
| VinVL (Zhang et al., 2021)     | 5.7M              | 92.6      | 12.5        | 88.3      | 12.1    | 94.5  | 11.9  | 90.2  | 12.1  |
| LEMON$_{base}$ (Hu et al., 2021)| 12M               | 103.1     | 14.2        | 96.1      | 13.8    | 88.3  | 12.1  | 95.5  | 13.5  |
| LEMON$_{base}$ (Hu et al., 2021)| 200M              | 104.5     | 14.6        | 100.7     | 14.0    | 96.7  | 12.4  | 100.4 | 13.8  |
| BLIP                           | 14M               | 110.7     | 14.7        | 96.2      | 14.3    | 110.4 | 13.1  | 106.9 | 14.1  |
| BLIP                           | 129M              | 111.1     | 15.1        | 110.3     | 14.4    | 105.1 | 13.7  | 105.6 | 14.7  |
| BLIP$_{CapFilt-L}$              | 129M              | 111.3     | 14.9        | 108.6     | 14.8    | 111.5 | 14.2  | 109.6 | 14.7  |
| LEMON$_{large}$ (Hu et al., 2021)| 200M             |           |             |           |         | -    | 40.3  | 133.3 |
| SimVLM$_{huge}$ (Wang et al., 2021)| 1.8B             | 116.9     | 15.8        | 113.3     | 15.1    | 111.3 | 14.0  | 113.4 | 15.0  |
| BLIP$_{ViT-L}$                  | 129M              | 114.9     | 15.2        | 112.1     | 14.9    | 115.3 | 14.4  | 113.2 | 14.8  |
Visual Question Answering

• An answer generation task, which enables open-ended VQA.
• During finetuning, they rearrange the pre-trained model, where an image-question is first encoded into multimodal embeddings and then given to an answer decoder. The VQA model is finetuned with the LM loss using ground-truth answers as targets.

| Method      | Pre-train #Images | VQA test-dev | VQA test-std | NLVR² dev | NLVR² 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        |
| SimVLMbase† | 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**    |
BLIPcapFilt-L| 129M              | **78.25**    | **78.32**    | 82.15     | 82.24        |
• NLVR (Suhr et al., 2019) asks the model to predict whether a sentence describes a pair of images.

• Make a "simple" modification to our pre-trained model -> a more computational-efficient architecture.

| Method        | Pre-train #Images | VQA test-dev | VQA test-std | NLVR² dev | NLVR² 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\textsubscript{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\textsubscript{CapFilt-L} | 129M            | 78.25        | 78.32        | 82.15     | 82.24        |
Visual Dialog

• VisDial (Das et al., 2017) extends VQA in a natural conversational setting, where the model needs to predict an answer not only based on the image-question pair, but also considering the dialog history and the image’s caption.

• Follow the discriminative setting where the model ranks a pool of answer candidates.

| 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 |

(c) VisDial

```
Image Encoder
```

```
Caption Encoder
```

```
Dialog Encoder
```

```
"[Encode] + C " "[Encode] + QA + Dialog History"
```

true/false
Zero-Shot Transfer to Video Language

- Perform zero-shot transfer to text-to-video retrieval and video question answering, where we directly evaluate the models trained on COCO-retrieval and VQA, respectively.

- To process video input, we uniformly sample $n$ frames per video ($n = 8$ for retrieval and $n = 16$ for QA) and concatenate the frame features into a single sequence. Note that this simple approach ignores all temporal information.

| Method                  | MSRVTT-QA | MSVD-QA |
|-------------------------|-----------|---------|
| ActBERT (Zhu & Yang, 2020) | 8.6       | 23.4    |
| SupportSet (Patrick et al., 2021) | 8.7       | 23.0    |
| MIL-NCE (Miech et al., 2020) | 9.9       | 24.0    |
| VideoCLIP (Xu et al., 2021) | 10.4      | 22.2    |
| FiT (Bain et al., 2021)    | 18.7      | 39.5    |
| BLIP                      | 43.3      | 65.6    |

| Method                  | MSRVTT-QA | MSVD-QA |
|-------------------------|-----------|---------|
| VQA-T (Yang et al., 2021) | 2.9       | 7.5     |
| BLIP                     | 19.2      | 35.2    |

| Method                  | MSRVTT-QA | MSVD-QA |
|-------------------------|-----------|---------|
| 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    |

| 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        |

| Method                  | R1↑       | R5↑      | R10↑     | MdR↓     |
|-------------------------|-----------|----------|----------|----------|
| finetuning              |           |          |          |          |
| ClipBERT (Lei et al., 2021) | 22.0      | 46.8     | 59.9     | 6        |
| VideoCLIP (Xu et al., 2021) | 30.9      | 55.4     | 66.8     | -        |
### BLIP2 Test Results

| Models                | #Trainable Params | Open-sourced? | Visual Question Answering | Image Captioning | Image-Text Retrieval |
|-----------------------|-------------------|---------------|---------------------------|------------------|----------------------|
|                       |                   |               | VQA acc. (test-dev)       | NoCaps (val)     | Flickr (test)        |
|                       |                   |               |                           | CIDEr | SPICE     | TR@1 | IR@1  |
| BLIP (Li et al., 2022)| 583M              | ✓             | -                         | 113.2 | 14.8      | 96.7  | 86.7  |
| SimVLM (Wang et al., 2021b) | 1.4B          | x             | -                         | 112.2 | -         | -    | -    |
| BEIT-3 (Wang et al., 2022b) | 1.9B          | x             | -                         | -     | -         | 94.9  | 81.5  |
| Flamingo (Alayrac et al., 2022) | 10.2B         | x             | 56.3                      | -     | -         | -    | -    |
| BLIP-2                | 188M              | ✓             | 65.0                      | 121.6 | 15.8      | 97.6  | 89.7  |

As for now, the presenter did not find same testing set up with same result shown in BLIP paper.
Limitations & Societal Implications

- No specified dataset for most of the downstream task.
- Social and Confirmation bias introduced by CapFilt.
- Bias introduced by sampling method for downstream tasks.
- Various big changes to the model modules in downstream tasks.
- Hard to debug
Summary of Strengths, Weaknesses, Relationships

- New pre-training (fine-tuning) techniques
- Synthetic data (less noise/variance)
- Scalability (In terms of performance)
- Network architecture is modular

- Very specific finetuning procedure
- Introduced new architectures after pre-training
- Very specific testing setup with a lot of model changes
- Synthetic data (more bias)
Question for Discussion

1. Is Nucleus Sampling the best in this field? In general, does maximum likelihood have a place in generalization?

2. What feature of the data impact BLIP the most? Is it the data quality or the data diversity?

3. The authors hypothesize that there may be confirmation bias in their CapFilt as the captioning and filtering were finetuned from the same COCO dataset, and the downstream tasks were similar to the COCO dataset. What methods to generate diverse captions?

4. Is the design of 3 separate modules with 3 different losses (especially the encoder and decoder) necessary for the model to succeed?

5. Is there a better way to generate or augment current annotated data?
Align before Fuse: Vision and Language Representation Learning with Momentum Distillation, NeurIPS’21
ALBEF Continued

- “ALign image and text representations BEFORE Fusing
- Momentum distillation (self-training method) -> learn from pseudo-targets generated by the momentum model
- Gap: challenging for multimodal encoder to learn image-text interactions when visual tokens and word tokens are unaligned