

CS657: Information Retrieval

# Cross-lingual Cross-modal Pretraining for Multimodal Retrieval

Hongliang Fei, Tan Yu, Ping Li  
Baidu Research

**Presented By:**

Nitik Jain (170448)

Moksh Shukla (180433)

Aviral Agarwal (180167)

Shubham Gupta (180749)

Archi Gupta (21111014)



# INTRODUCTION

➤ **Premise:** Recent pretrained vision-language models have achieved impressive performance on cross-modal retrieval tasks in English.



➤ Their **success** depends on: Availability of annotated image-caption datasets for pre-training.



➤ **Limitation:** Perform poorly where the texts are not necessarily in English



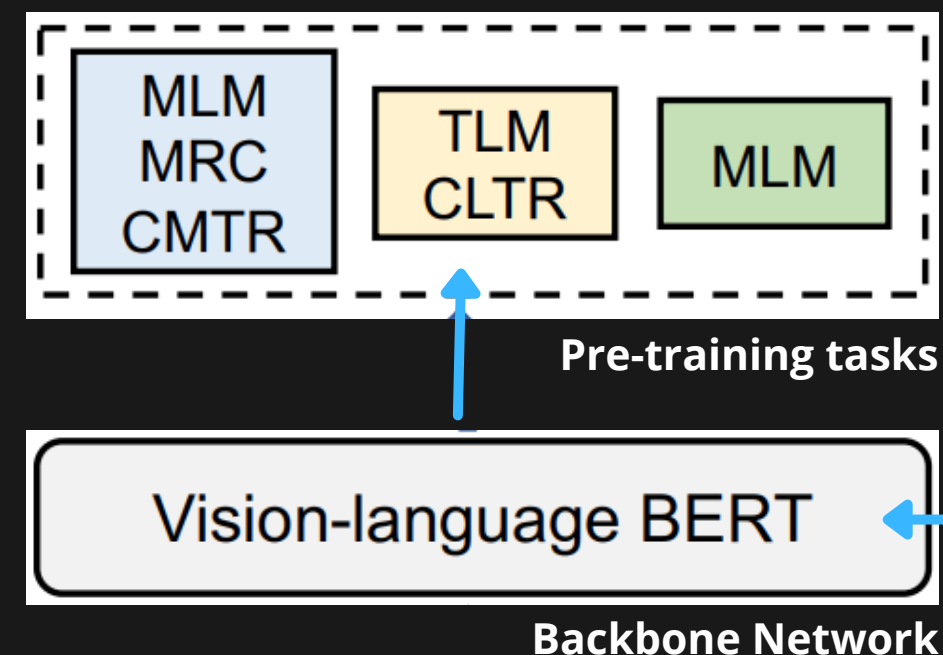
➤ **Alternative:** Utilize Machine Translation (MT) tools to translate non-english text to english



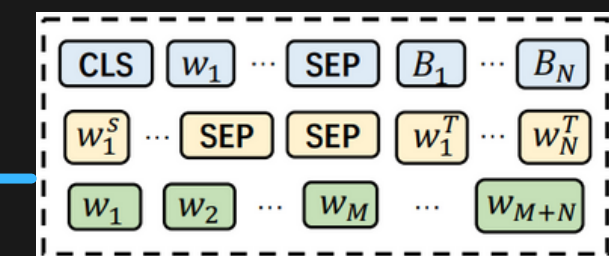
**The alternative largely relies on MTs quality and suffers from latency in real world deployment**

This paper proposes a cross-lingual cross-modal pretraining framework to learn a language invariant representation across image and text modalities.

- They hypothesize that introducing pretraining tasks involving different languages and modalities and modeling the interaction among them leads to powerful joint representation and generalizes well
- They also introduce monolingual and parallel corpus involving other languages to refine the shared latent space previously based only on English Caption Data






EN-caption & visual features  
Parallel corpus  
Monolingual sentences



# RELATED WORK

## CROSS MODAL

### Visual-language pretrained models

VL-BERT: Pre-training of Generic Visual-Linguistic Representations   




Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, Jifeng Dai

- Emerged recently, pre-training typically consists
  - **Masked Language Modelling**
  - **Masked Region Modelling**
  - **Text-Image matching**
- Cross-modal BERT methods achieved state-of-the-art performance in many text-vision understanding tasks.
- These models exploit cross-modal attention and are pretrained on large datasets.
- **Nevertheless, all the above models deal with a single language English and image or video domain.**

**Hence the paper proposes integrating cross-lingual pretraining tasks with vision-language pretraining to obtain a universal multilingual multimodal representation.**

## CROSS LINGUAL

### Cross-lingual pretrained models

Cross-lingual Language Model Pretraining   

Guillaume Lample, Alexis Conneau

- Cross-lingual pretrained models are capable of simultaneously encoding texts from **multiple languages**
  - For eg. multilingual BERT is similar to BERT (in structure and training) but was pretrained on 100 languages on Wikipedia
- XLM model (Conneau and Lample) was pre-trained on MLM and TLM to take advantage of parallel sentence resources.
- Experiments on a series of cross-lingual transfer tasks have shown that these cross-lingual language models have significant utilities for transferring knowledge between languages.

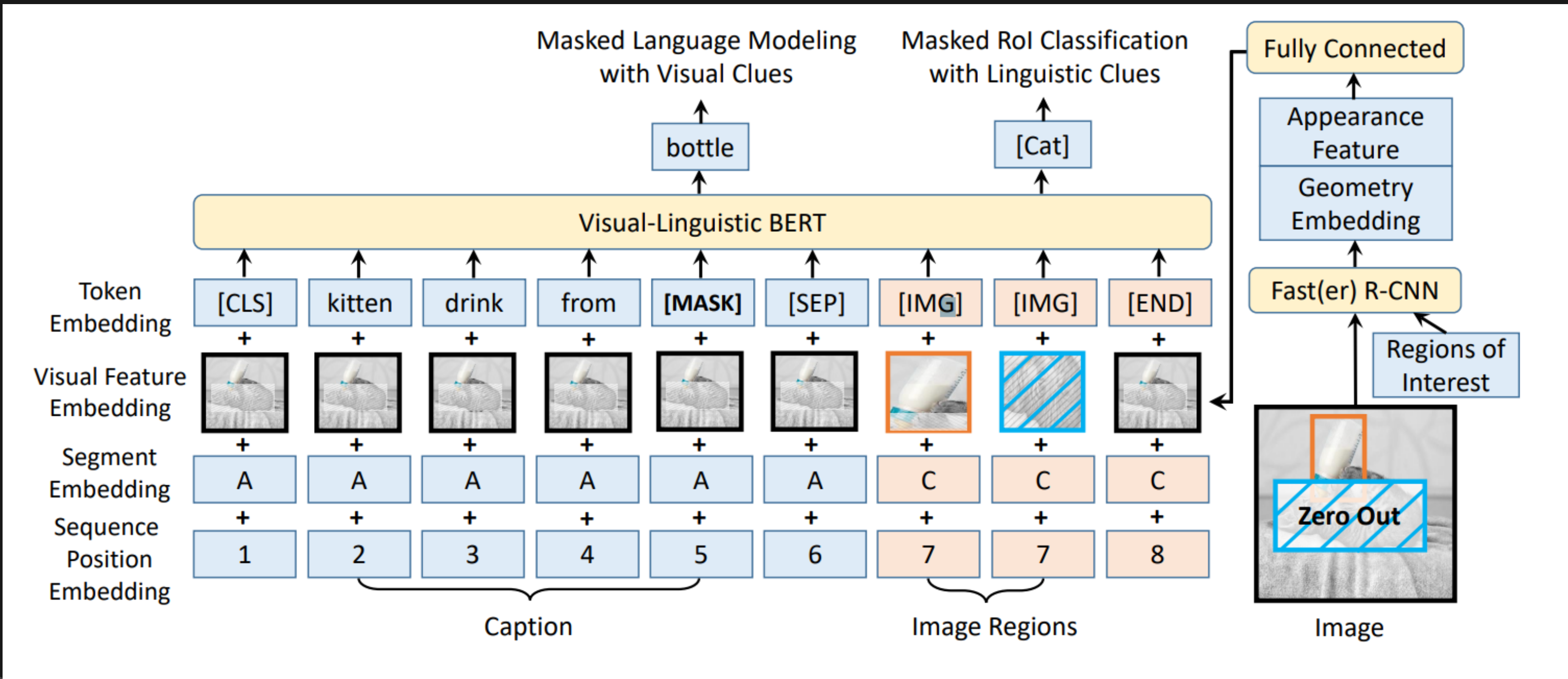
# METHODOLOGY

- The framework adopts the network structure of **VL-BERT** : transformer based single stream cross-modal vision-language model
- The backbone is of multi-layer bidirectional **Transformer encoder**, enabling **dependency modeling** among all the input elements
- Input - concatenated word features from the text and bounding box features from the image (Fast R-CNN to detect features defined on regions-of-interest (Rols))
- For disambiguating different input formats <Caption, img> , < Ques, Ans, img> : every input element has four types of embedding : **token embedding, visual feature embedding, segment embedding, and sequence position embedding**
- Pre-train on both visual-linguistic (Conceptual Captions dataset) and text only datasets (BooksCorpus, English Wikipedia datasets)

- Task #1: Masked Language Modeling with Visual Clues
- Task #2: Masked RoI Classification with Linguistic Clues



# VL-BERT

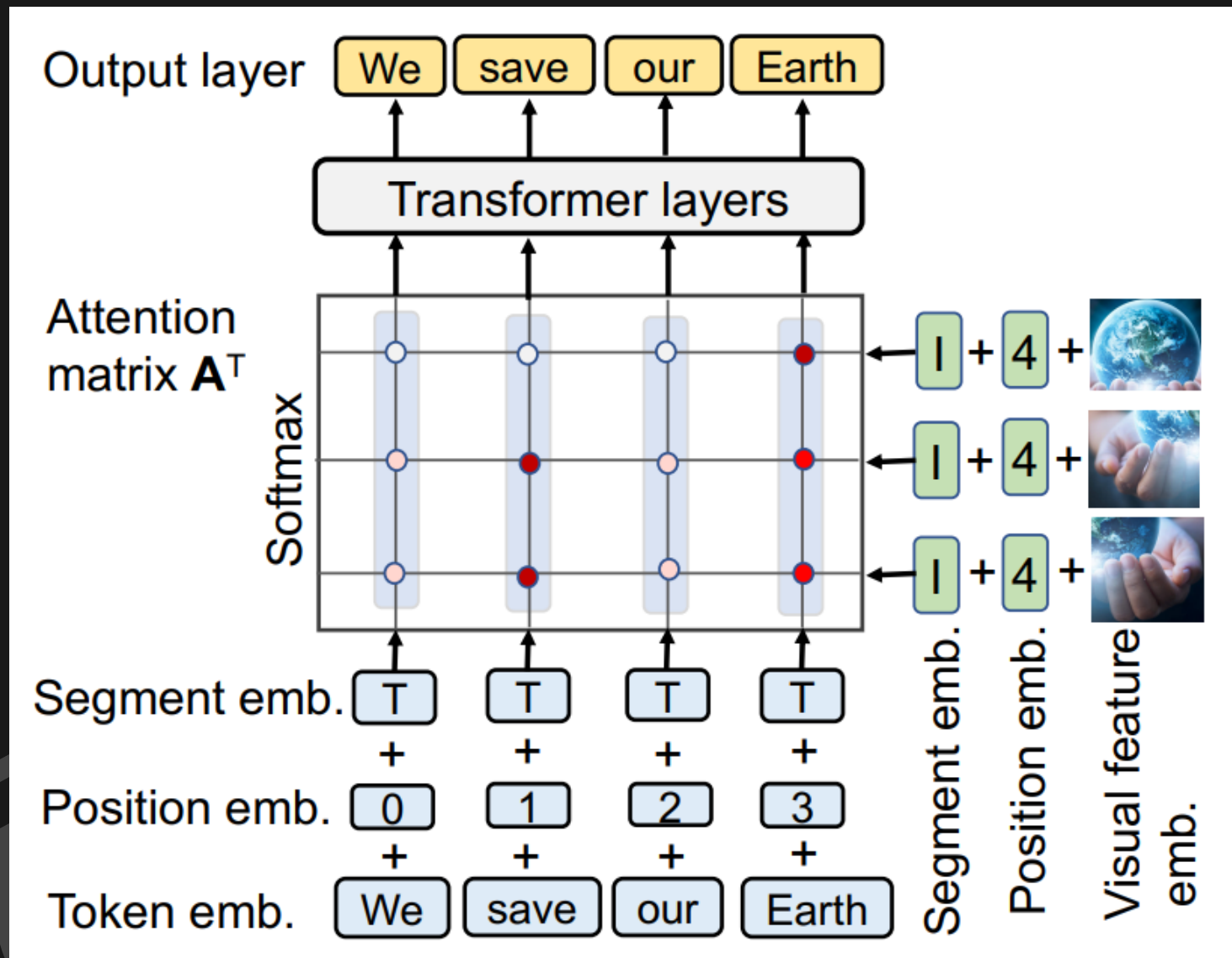


- Task #1: Masked Language Modeling with Visual Clues
- Task #2: Masked RoI Classification with Linguistic Clues

# ADDITIONAL PRE-TRAINING TASKS

## Building on top of VL-BERT

They introduce two additional cross-lingual pretraining tasks and one cross-modal task for improving the performance



## Task #3 : Cross-model Text Recovery

It computes an alignment between word features and bounding box features and uses attended image features to recover all input words simultaneously

## Task #4 : Cross-lingual Text Recovery

Learns word alignments between a pair of parallel sentences (X, Y) (bi-linear attention) and uses attended representation of X (source language) with its parallel sentence Y to recover X.

## Task #5 : Translation Language Model

Takes a pair of parallel sentences with randomly masked tokens in different languages to predict the masked tokens by attending to local contexts and distant contexts in another language

# FINE TUNING

↘ For Cross Modal Retrieval

- Minimize the Triplet Ranking Loss
- To boost the performance, hard negative mining strategy is used
- For every text query, there is only one +ive image sample
- The hardest negative image in the mini-batch is penalized

$$\mathcal{L}(I_i) = \max_{j \neq i} [R(q_j, I_i) - R(q_i, I_i) + m]_+.$$

- For each image, the hardest negative query is penalized
- Final loss function is calculated by taking average of the above losses over all the samples in the mini batch

# EXPERIMENTAL SETUP and DATASET DESCRIPTION

## For Pretraining:

- Image Caption Data (3.7M text-image pairs):
  - SBU Captions (Ordonez et al., 2011)
  - Conceptual Captions (Sharma et al., 2018)
- For Monolingual (English (en), German (de) and Japanese (ja):
  - Randomly sampled 20M sentences from Wikipedia
- Parallel Corpus
  - en-de - 9M Parallel Sentences from MultiUN Corpus
  - en-ja - 2.8M Parallel Sentences from Pryzant et al.

## For Fine-tuning:

- Two multilingual multimodal benchmarks
  - MSCOCO (en, ja) (Lin et al., 2014):
    - MSCOCO contains 123, 287 images, and each image contains five captions.
  - Multi30k (en, de) (Elliott et al., 2016):
    - Multi30K contains 31, 783 images, with each having five captions as well

## Experiment Setting

- Model Architecture -
  - Multilingual BERT uncased version (Devlin et al., 2019)
    - 12 Layers of Transformer Blocks
    - Each block has
      - 768 hidden hidden units
      - 12 self-attention heads
  - Vocabulary size is 105,879, max\_sequence\_len = 64
- Evaluation metrics for validation set - R@K (K=1,5,10)
  - R@K is the percentage of ground-truth matchings appearing in the top K-ranked results.
- Optimizer - Adam
- Pretrained for 50 epochs
- Finetuned based on average of R@{1,5,10} on validation set



# EXPERIMENTAL RESULTS

## Cross-Modal Retrieval on English Tasks

Method	MSCOCO (en)						Multi30K (en)					
	img2txt Recall@			txt2img Recall@			img2txt Recall@			txt2img Recall@		
	1	5	10	1	5	10	1	5	10	1	5	10
SCAN	72.7	94.8	98.4	58.8	88.4	94.8	67.4	90.3	95.8	48.6	77.7	85.2
Unicoder-VL	<b>84.3</b>	<b>97.3</b>	99.3	<b>69.7</b>	<b>93.5</b>	<b>97.2</b>	<b>86.2</b>	<b>96.3</b>	<b>99.0</b>	<b>71.5</b>	<b>90.9</b>	<b>94.9</b>
VL-BERT	76.4	96.8	99.2	64.1	90.9	96.3	79.8	94.9	96.8	61.8	86.4	92.1
Ours	80.5	97.1	<b>99.5</b>	65.1	91.7	96.5	80.6	94.9	97.9	63.3	87.6	92.4

**Table:** Cross-modal retrieval results (in percentage %) for English.

- Compared with Unicoder-VL (Li et al., 2020), proposed model performs slightly worse but obtains better results than VL-BERT.
- A possible reason is that Unicoder-VL is initialized with English BERT, which is specifically optimized for English.

## Cross-Modal Retrieval on non-English Tasks

Method	MSCOCO (ja)						Multi30K (de)					
	img2txt Recall@			txt2img Recall@			img2txt Recall@			txt2img Recall@		
	1	5	10	1	5	10	1	5	10	1	5	10
SCAN	56.5	85.7	93.0	42.5	73.6	83.4	51.8	82.0	91.0	35.7	60.9	71.0
AME	55.5	87.9	95.2	44.9	80.7	89.3	40.5	74.3	83.4	31.0	60.5	70.6
LIWE	56.9	86.1	94.1	45.1	78.0	88.2	59.9	87.5	93.7	42.3	71.1	79.8
Translate-test	66.2	88.8	94.8	52.1	82.5	90.6	69.8	90.2	94.8	51.2	77.9	86.6
VL-BERT	60.3	85.9	94.5	48.4	81.7	90.5	65.7	88.0	94.0	47.4	77.0	85.4
Ours	<b>67.4</b>	<b>90.6</b>	<b>96.2</b>	<b>54.4</b>	<b>84.4</b>	<b>92.2</b>	<b>71.1</b>	<b>91.2</b>	<b>95.7</b>	<b>53.7</b>	<b>80.5</b>	<b>87.6</b>

**Table:** Cross-modal retrieval results for Japanese (MSCOCO) and German (Multi30K).

- “Translate-test” baseline achieves better results than VL-BERT pretrained with MLM objective only on multilingual corpus and finetuned in the target language.
- Recall of “Translate-test” is 1-2% lower than proposed method indicating the effectiveness of pretraining with additional cross-lingual objectives than translating the target language into English
- Proposed method also performs better compared with VL-BERT (Su et al., 2020) that is only pretrained with MLM task on multilingual corpus

# CONCLUSION

In this work, the author introduce multilingual corpus and three pretraining objectives to improve transformer based vision-language models for retrieval tasks. Extensive experiments demonstrate the effectiveness of on cross-modal retrieval tasks.

Future work as mentioned by the author is to explore the zero-shot transferring capability of the proposed framework.





*Thank!  
You!*

