ImageBind

Introduction

A major obstacle in learning a true joint embedding is the absence of large quantities of multimodal data where all modalities are present together.

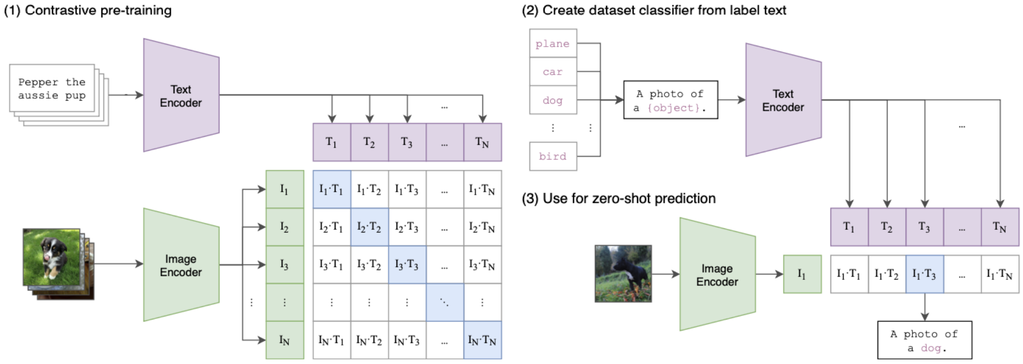
→ **does not need datasets where all modalities co-occur with each other**.

(데이터 부족의 문제점을 보완, 완벽한 pair data 필요로하지 않는다.)

Related Work

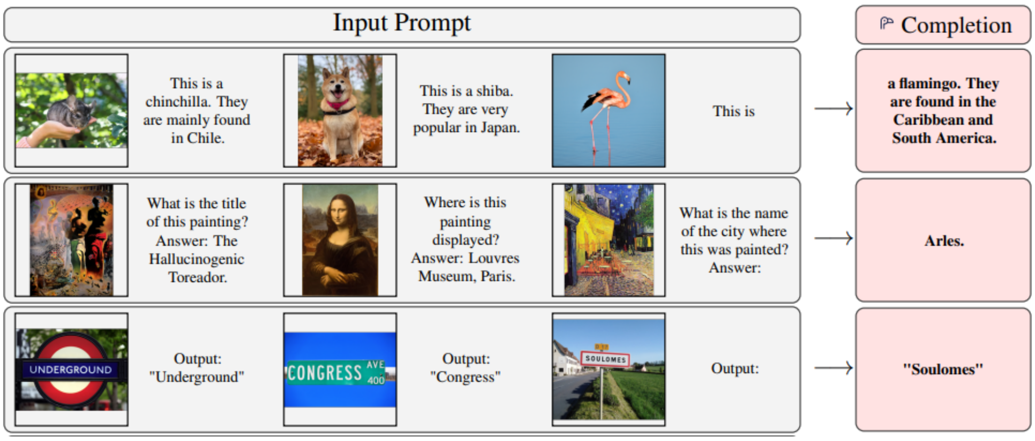
Large Image Pre-training

- Collect large collections of image and text pairs and train models to embed image and language inputs in a joint space using contrastive learning, exhibiting impressive zero-shot performance (ex. CLIP)


(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T1 T2 T3 ... TN
I1 I2 I3 ... IN
I1·T1 I1·T2 I1·T3 ... I1·TN
I2·T1 I2·T2 I2·T3 ... I2·TN
I3·T1 I3·T2 I3·T3 ... I3·TN
IN·T1 IN·T2 IN·T3 ... IN·TN
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
I1
T1 T2 T3 ... TN
I1·T1 I1·T2 I1·T3 ... I1·TN
A photo of a dog.

CLIP

- Handles arbitrarily interleaved images and texts, and achieves state of the art on many few-shot learning benchmarks (ex. Flamingo)


Input Prompt
Completion
This is a chinchilla. They are mainly found in Chile.
This is a shiba. They are very popular in Japan.
This is
a flamingo. They are found in the Caribbean and South America.
What is the title of this painting? Answer: The Hallucinogenic Toreador.
Where is this painting displayed? Answer: Louvres Museum, Paris.
What is the name of the city where this was painted? Answer:
Arles.
UNDERGROUND
Output: "Underground"
CONGRESS AVE 400
Output: "Congress"
SOULOMES
Output:
"Soulomes"

Flamingo

(CLIP은 text와 image 사이의 similarity score만 제공하기에 classification 같은 문제에서만 효과적이었지 텍스트를 생성해내야 하는 open-ended tasks에선 취약 → 이를 보완하여 광범위하게 적용가능하게 만듦)

- Adopts contrastive training for fine-tuning and observes freezing image encoders works the best. (ex. LiT)

✅ Background Knowledge: Zero-shot Transfer

Transfer Learning:
(1) Pre-training
(2) Fine-tuning

Zero-shot Transfer: Transfer Learning **without** Fine-tuning!
(→ 어떤 데이터이더라도 적용 가능)

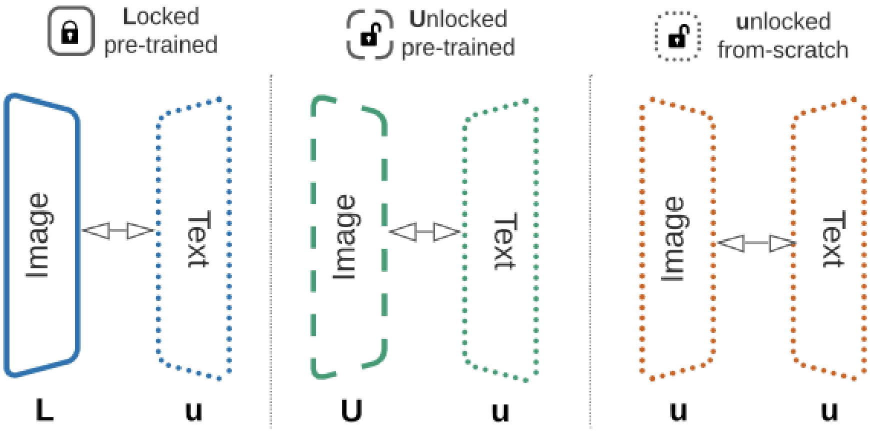


Figure 2. Design choices for contrastive-tuning on image-text data. Two letters are introduced to represent the image tower and text tower setups. L stands for locked variables and initialized from a pre-trained model, U stands for unlocked and initialized from a pre-trained model, u stands for unlocked and randomly initialized. Lu is named as Locked-image Text tuning (LiT-tuning).

LiT 내용: 어떤 zero-shot transfer setting이 가장 좋은가

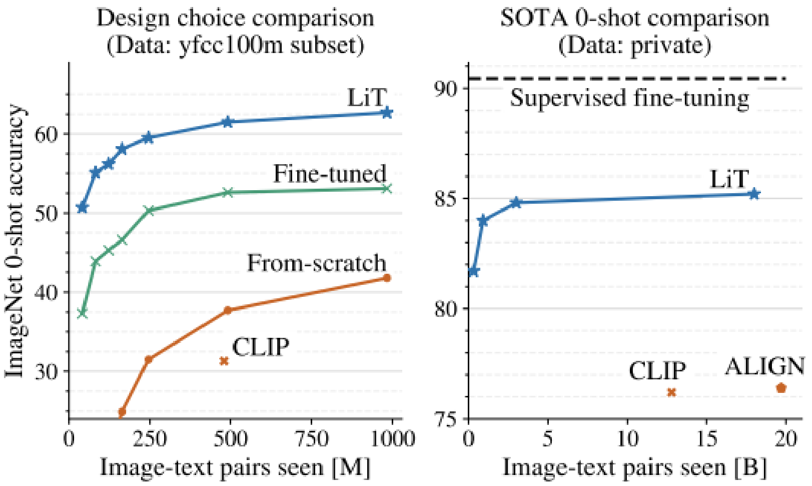


Figure 1. Comparison to the previous SOTA methods. **Left**: results on public YFCC100m subset, with from-scratch, fine-tuned from a pre-trained image model, and LiT with a pre-trained image model. The proposed LiT improves over 30% ImageNet zero-shot transfer accuracy on YFCC100m subset. **Right**: results on privately gathered data, LiT halves the gap between previous from-scratch methods CLIP [46], ALIGN [31] and supervised fine-tuning [13, 69].

Text Encoder은 freeze하고 Image쪽 모델만 훈련시키는 게 성능에 가장 좋다

→ This prior line of works mostly considers image and text, while our work enables zero-shot recognition on multiple modalities.

Multi-modal Learning

- Most related to our method, Nagrani et al.

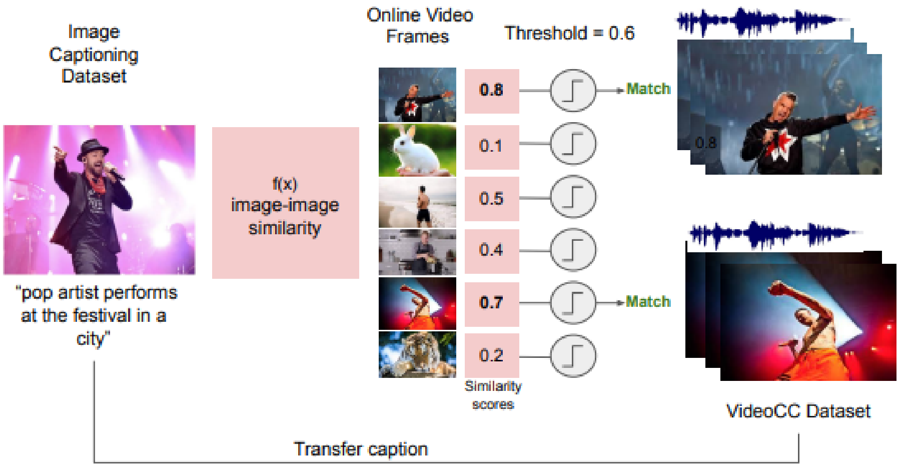


Figure 1. **Mining Audio-video clips automatically.** We use the images in image captioning datasets as 'seed' frames to mine related audio-visual clips. For each seed image-caption pair in a dataset, we find frames in videos with high similarity scores to the seed image. We then extract short video clips around the matching frames and transfer the caption to those clips. This gives us free captioning supervision for video and audio clips.

Text+Image+Audio (Video)

[논문 일부 발췌] We use images from image captioning datasets as seeds to find similar clips in videos online (Fig. 1). We then transfer the image captions directly to these clips, obtaining weak, albeit free video and audio captioning supervision in the process. This can also provide us with motion and audio supervision – for example, sometimes human-generated captions for images infer other modalities, eg. the caption 'Person throws a pitch during a game against university' from the CC3M dataset [70] was written for a single, still image, but is actually describing motion that would occur in a video. Similarly, the caption 'A person singing a song', is also inferring a potential audio track.

- AudioCLIP: adds audio as an additional modality into a CLIP framework, enabling zero-shot audio classification.

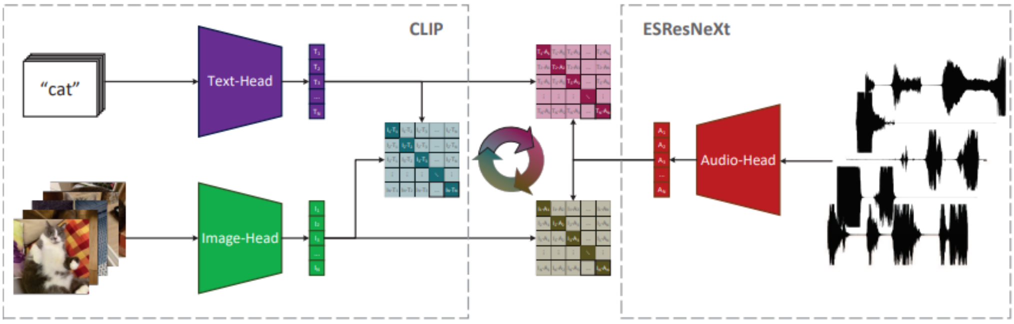


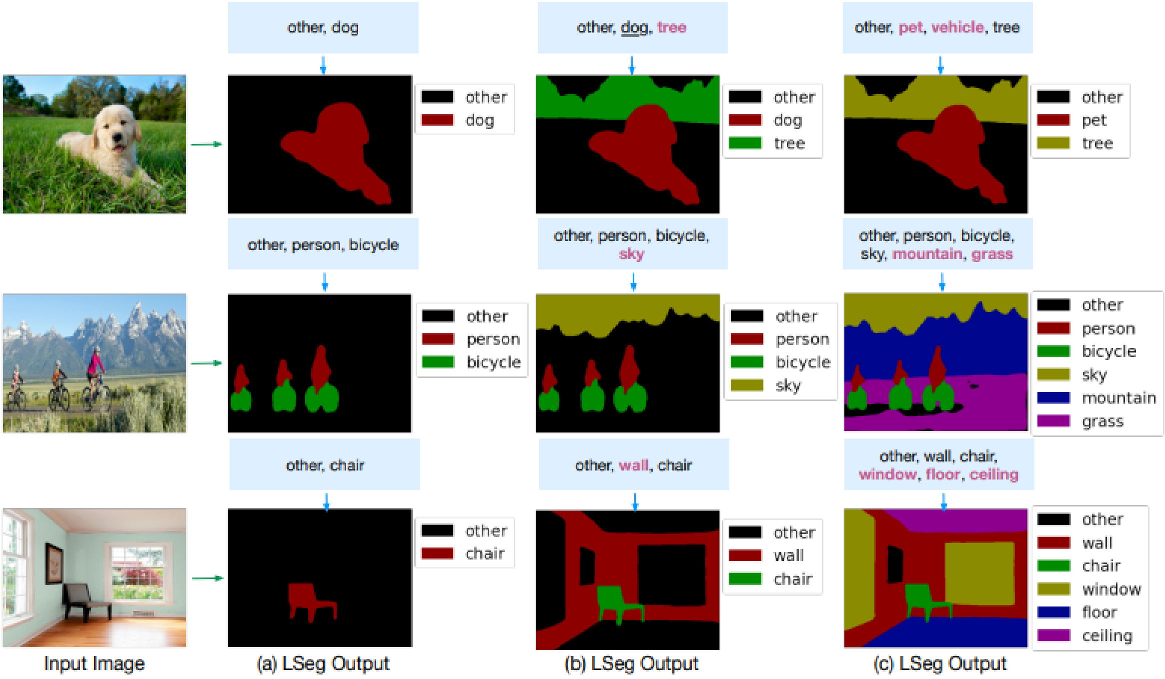
Figure 1. Overview of the proposed AudioCLIP model. On the left, the workflow of the text-image-model CLIP is shown. Performing joint training of the text- and image-heads, CLIP learns to align representations of the same concept in a shared multimodal embedding space. On the right, the audio-model ESResNeXT is shown. Here, the added audible modality interacts with two others, enabling the model to handle 3 modalities simultaneously.

AudioCLIP

→ In contrast, IMAGEBIND does not require explicit paired data between all modalities and instead leverages image as a natural weak supervision for unifying modalities.

Feature Alignment

Pre-trained CLIP models have been utilized as teachers to supervise other models due to the strength of its visual representations. Moreover, CLIP joint image and text embedding space has also been leveraged for a variety of zero-shot tasks like detection, segmentation, mesh animation etc. showing the power of joint embedding spaces.



Language-Driven Semantic Segmentation

이미지에서 원하는 객체만 segmentation 가능하다! (위 논문에선 Text Encoder CLIP을 가져와서 freeze하여 사용)

In multilingual neural machine translation, a similar phenomenon to the emergence behavior of IMAGEBIND is commonly observed and utilized: if languages are trained in the same latent space through learned implicit bridging, translation can be done between language pairs on which no paired data is provided

—> 같은 space에서 multimodal을 학습시킨다면, modal간 translation/transfer이 가능하다!