Multimodal Embedding

(https://medium.com/@raj.pulapakura/multimodal-models-and-fusion-a-complete-guide-225ca91f6861)

1. Mutimodal fusion의 필요성

- 인간은 학습할 때 여러 형태의 data를 효율적으로 사용한다.
- 타인과 상호작용할 때도 Multi-modality(voice, pitch, tone, body language)를 사용한다.
- Multimodal model의 성능이 단일 모달 모델보다 우수하다.

==> to gain **a comprehensive understanding** of the subject matter

2. Unimodal Image Model의 문제와 Multimodality의 필요성

◉ 학습용 데이터셋 구축비용이 큰 반면, 할 수 있는 일이 제한적이다

예: Imagenet

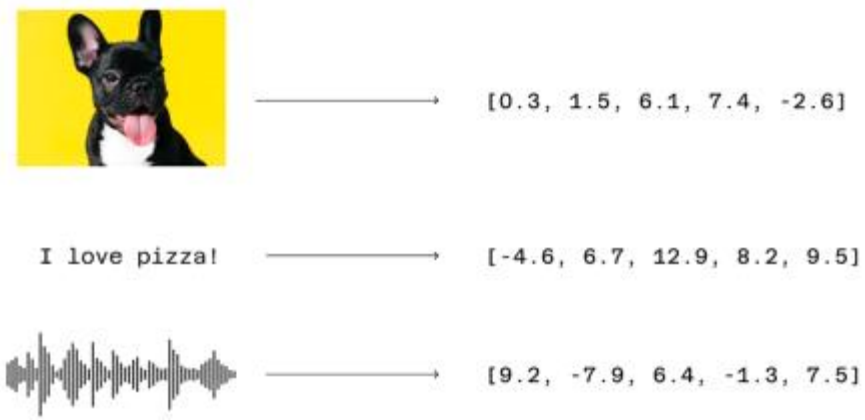
- 25,000명이 참가 , 천4백만장의 Image 구축 (22,0000 object categories)
- But, 1,000개의 카테고리 분류 정도 외의 일은 하기 어려움

◉ “benchmark performance” 와 “real performance” 간의 간극이 크다

◉ Multimodal의 예로 CLIP(OpenAI) 의 경우,

- 사람에 의한 별도의 annotation 없이 인터넷 상의 image-text 쌍을 이용해서 학습
- 다양한 visual task 처리가 가능

※ 어떤 Modality라도 결국 숫자들의 시퀀스로 임베딩 되어야 함



3. 서로 다른 Modality들의 결합 방법 4가지

서로 다른 modality들의 **결합 시점**에 따라 구분

◉ Early Fusion

-입력 레벨에서 한 가지 표현으로 결합

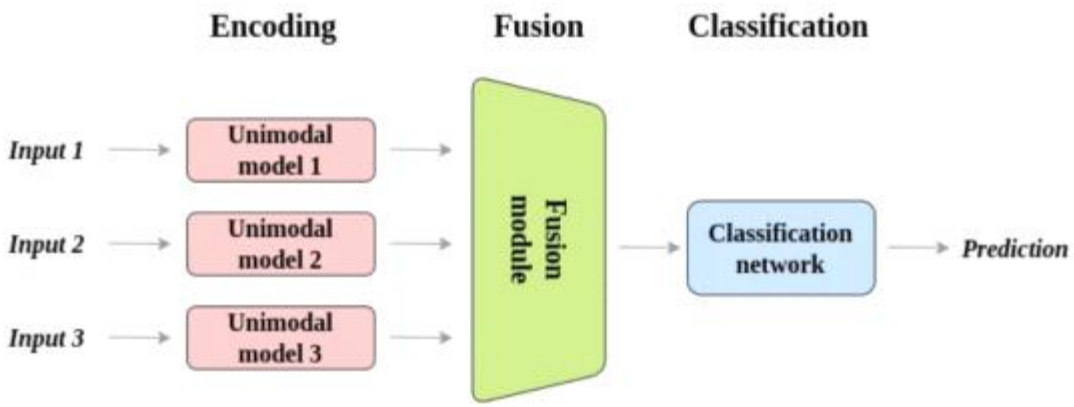
-각 모델리티에 대한 입력 표현의 **단순한 결합**으로도 구현 가능

-입력 모델리티 간의 복잡한의미관계를 잘 나타내지 못할 수 있다

예: Amazon 책 서평 출력(Positive, Negative)의 경우
"book review"와 "책 가격, 저자, 장르"를 결합

◉ intermediate Fusion
-각 입력 모델리티를 잠재적인(latent) 표현으로 변경한 후 결합
-잠재적 표현의 구현 예
: concatenation, element-wise addition, attention, ...
-잠재적 표현이 좋은 경우, **모달리티간의 의미관계를 잘 학습할 수 있는 장점**이 있음
-잠재적 표현으로의 변환 과정에 따른 latency 발생도 가능할 수 있음
예: 자율주행에서 Cameras, LiDar, radar, GPS 입력들의 결합

◉ Late Fusion
- **앙상블 모델**에 가까움
- 각 입력은 각각의 모달리티를 구현하는 모델들로 처리 후 출력 결과들을 결합함
- **단순한 결합 방법과 모달리티 이용방법간의 단순한 구분**이 장점
-모달리티 간의 보완적일 수 있는 상호관계를 이용하지 못하는 단점이 있음
예: 비디오(YOUTUBE)로부터 장르 구분의 경우
비디오 입력과 자막 입력 각각을 처리하는 모델들로부터 얻은 분류 결과를 결합



◉ Hybrid Fusion
- Early, Intermediate, Late Fusion 방법들을 적절히 결합

4. Modality Dropout

-학습중에, 임의적으로 특정한 modality를 생략하거나 부정확하게 하여 특정 modality가 없어도 동작하도록 함.

-prediction시에 입력 modality들 중에 한 개의 modality만 있어도 동작 가능하게 함

◉ Early Fusion

5. Two Modalities Embedding

5.1 CLIP (Contrastive Language-Image Pretraining) by OpenAI (2021)

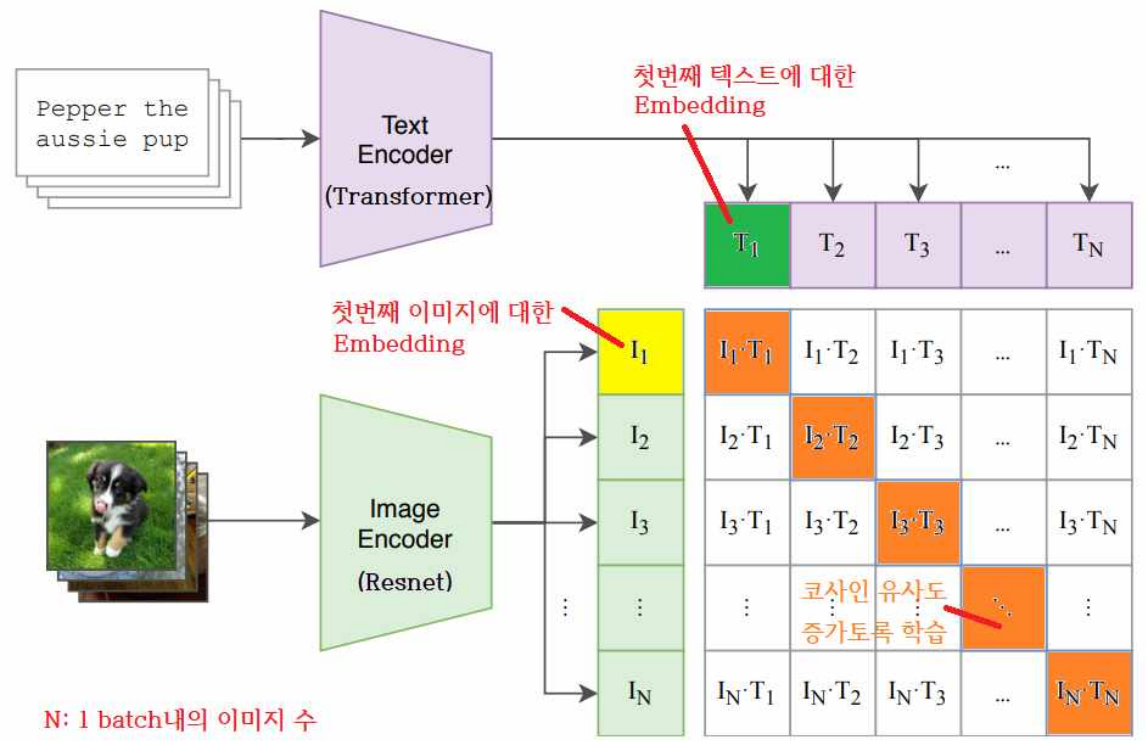
: Zero-shot recognition task

(https://openai.com/index/clip/)

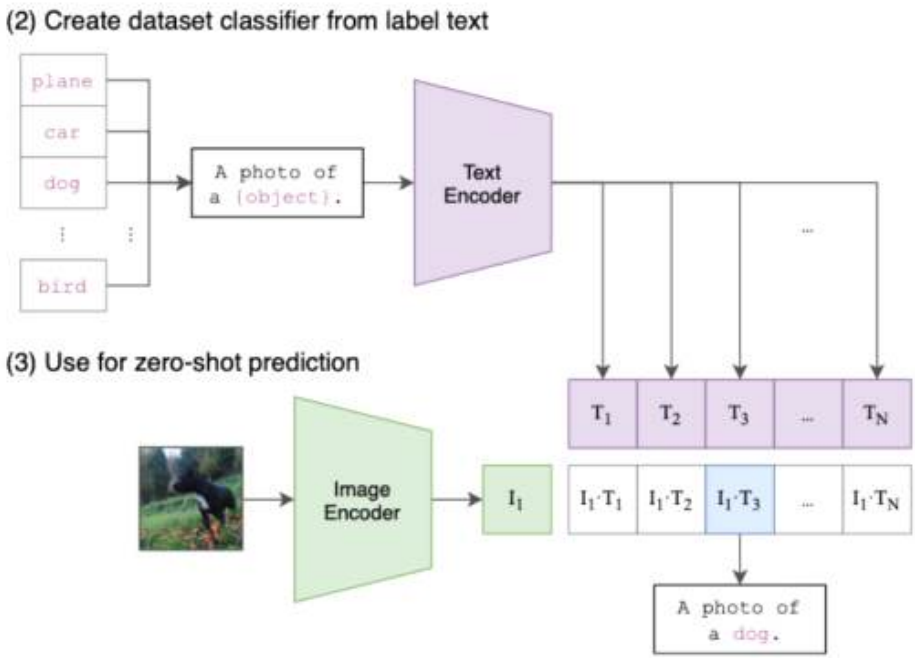
◉ CLIP의 특징

- 학습데이터용 이미지에 대한 인위적인 Labeling 대신 인터넷상에서 스크롤 한 데이터 중 이미지-자연어Text 쌍으로 학습(contrastive training, 1천5백만장) - **annotation free**
- 이미지에 대한 정확한 단어를 predict하는 것은 학습시간도 많이(3배) 덜리고 효율도 낮다

(1) Contrastive Pretraining



Architecture of CLIP model



(https://towardsdatascience.com/clip-model-and-the-importance-of-multimodal-embeddings-1c8f6b13bf72)

Feature Extraction:

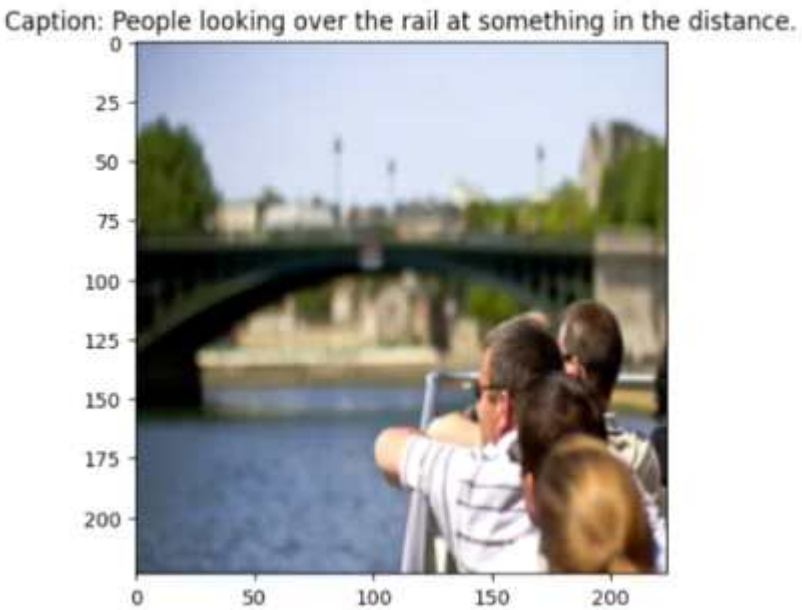
- `I_f = image_encoder(I)` : Extracts feature representations (`I_f`) from the image encoder. The shape of `I_f` is `[n, d_i]` , where `d_i` is the dimensionality of the image features.
- `T_f = text_encoder(T)` : Extracts feature representations (`T_f`) from the text encoder. The shape of `T_f` is `[n, d_t]` , where `d_t` is the dimensionality of the text features.

```
I_f = models.resnet34(pretrained=True)       # for encoding images
T_f= AutoModel.from_pretrained("distilbert-base-multilingual-cased") # for encod
```

Learned Projections:

- `W_i[d_i, d_e]` : Represents the learned projection matrix for mapping image features (`I_f`) to an embedding space (`I_e`). The shape of `W_i` is `[d_i, d_e]` , where `d_e` is the desired dimensionality of the joint embedding space.
- `W_t[d_t, d_e]` : Represents the learned projection matrix for mapping text features (`T_f`) to the same embedding space (`T_e`). The shape of `W_t` is `[d_t, d_e]` .

※ 학습용 데이터의 예



SUN397

television studio (90.2%) Ranked 1 out of 397 labels

✓ a photo of a **television studio**.

× a photo of a **podium indoor**.

× a photo of a **conference room**.

× a photo of a **lecture room**.

× a photo of a **control room**.

Food101

guacamole (90.1%) Ranked 1 out of 101 labels

✓ a photo of **guacamole**, a type of food.

× a photo of **ceviche**, a type of food.

× a photo of **edamame**, a type of food.

× a photo of **tuna tartare**, a type of food.

× a photo of **hummus**, a type of food.

5.2 Text Embedding을 CNN의 성능향상에 사용한 경우

(https://medium.com/@mgupta70/multimodal-learning-using-text-embeddings-to-fine-tune-cnns-40bdc32ec409)

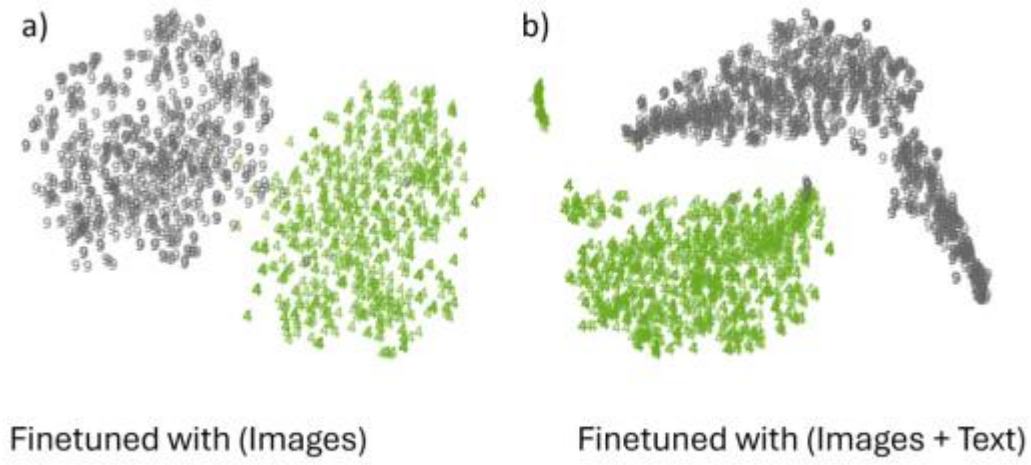


Figure 1: t-SNE plots of Visual Features for MNIST digits 4 & 9 Fine-tuned with ResNet18 using a) only Images and b) Text and Images

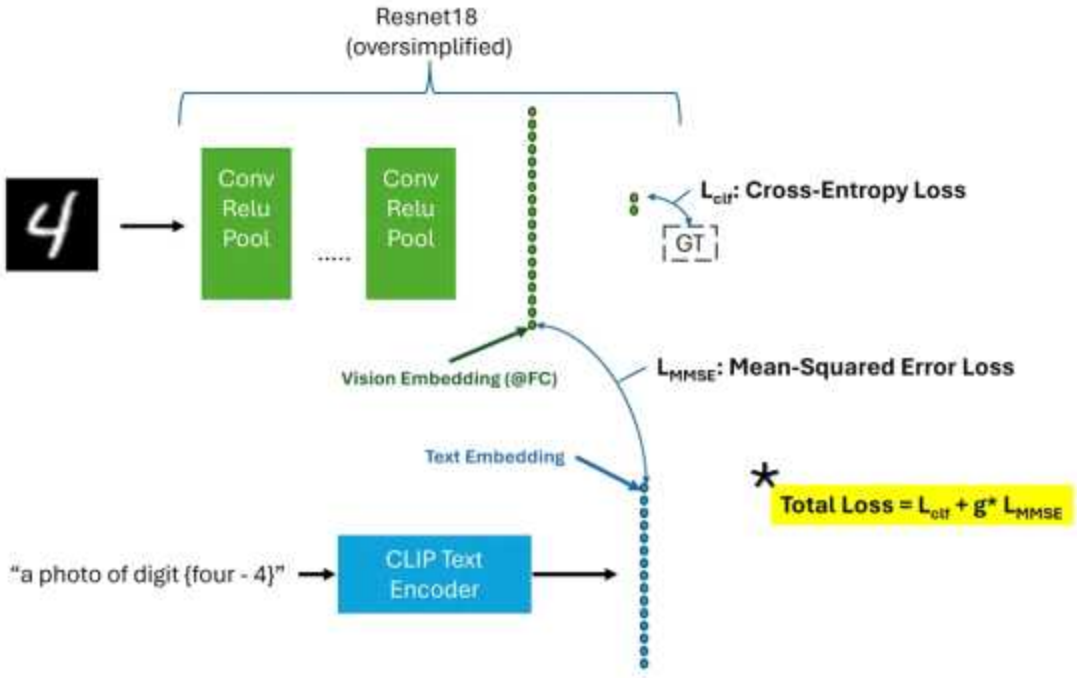


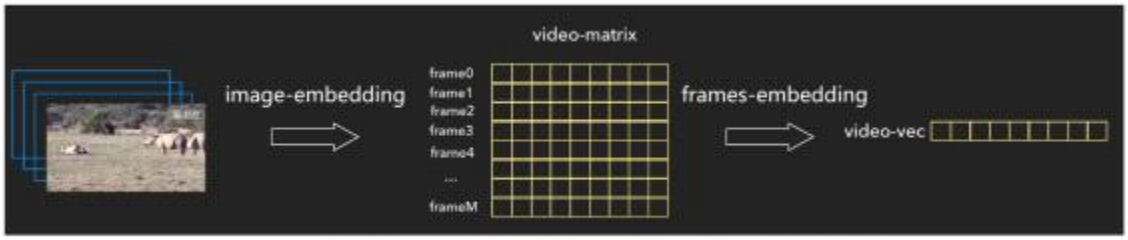
Figure 2 Overview of the proposed framework

5.3 Other Multimodalities

(https://www.twelvelabs.io/blog/multimodal-embeddings)

Pano-AVQA(the first large-scale spatial and audio-visual question-answering dataset on 360-degree videos),YouTube-360 (YT-360)(360-degree videos),AIST++(a new multimodal dataset of 3D dance motion and music), andArtEmis(affective language for visual arts). Astoundingly,MultiBenchis a dataset including ten modalities.

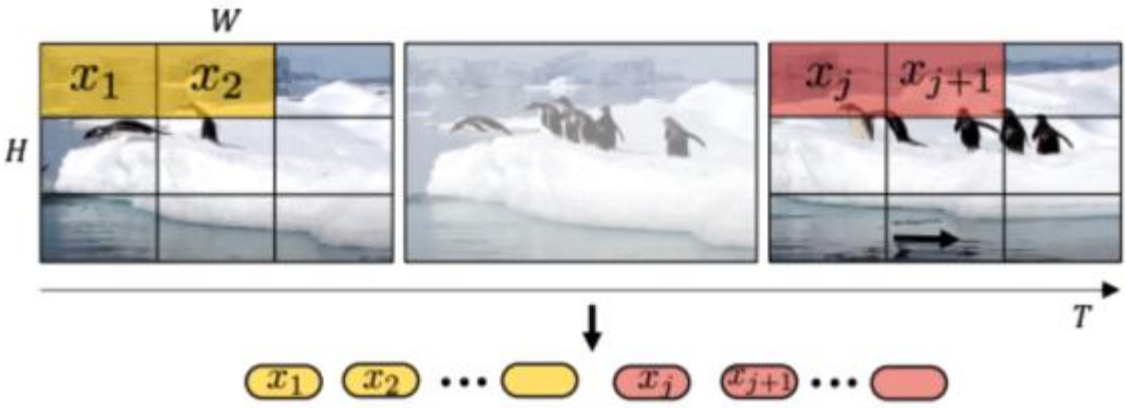
◉ Video(Image+Audio+Language) Embedding



Source: https://github.com/fengyoung/video_embedding

◉ Video Tokenization

- **Frame-level** tokenization : 한 개 frmae을 대상으로 여러개의 token 생성
- **Clip-level** tokenization : 연속적인 몇 개의 클립당 하나의 token 생성, 긴 비디오에 적합
- **Object-level** tokenization: frmae 내의 object나 region을 탐지 후, 이들을 각각 토큰화 함
- **Co**-tokenization : 비디오와 자막을 같이 토큰화 함. 비디오 Q-A 시스템에 적합
  object detection이나 action recognition task에 적합



Source: https://arxiv.org/abs/2103.15691

6. ImageBind (by Meta, May 2023. 논문:https://arxiv.org/pdf/2305.05665)

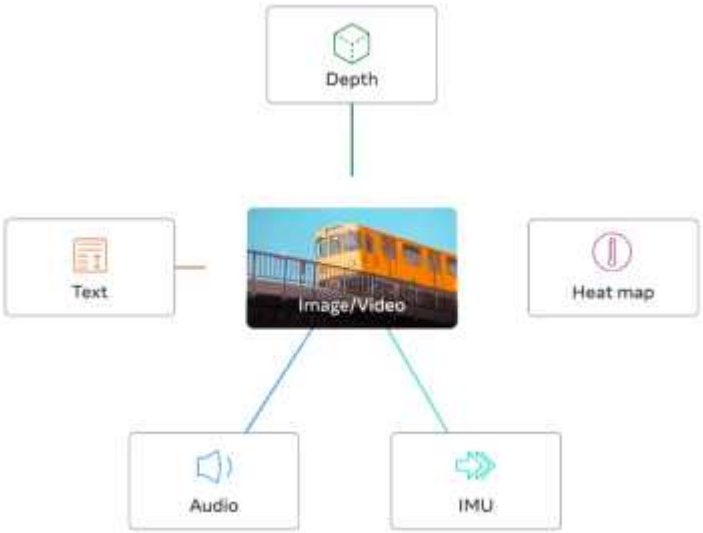
(https://github.com/facebookresearch/ImageBind)

◉ 6 Modalities :

image, text, audio, depth, thermal (infrared), and IMU (inertial measuring unit) data

◉ ImageBind Training Data :

image-paired data 사용. 즉 image + (1 of 5 modal data)

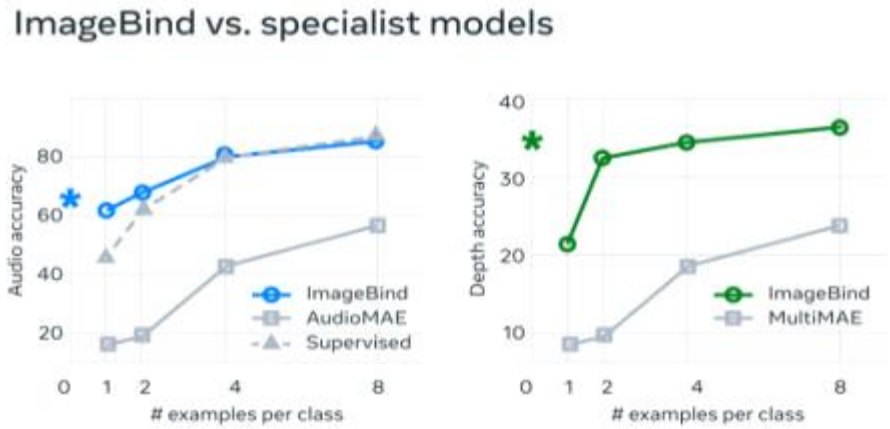


(https://imagebind.metademolab.com/)

◉ ImageBind Training Data

- **OpenCLIP ViT-H Encoder**: This encoder is utilized for initializing and freezing the image and text encoders. The ViT-H encoder is a part of the OpenCLIP model, a robust vision-language model that provides rich image and text representations.
- **Audio Embeddings**: ImageBind uses the Audioset dataset for training audio embeddings. Audioset is a comprehensive collection of audio event annotations and recordings, offering the model a wide array of sounds to learn from.
- **Depth Embeddings**: The SUN RGB-D datasetis used to train depth embeddings. This dataset includes images annotated with depth information, allowing the model to understand spatial relationships within the image.
- **IMU Data:**The Ego4D dataset is used for IMU data. This dataset provides IMU readings, which are instrumental in understanding movement and orientation related to the images.
- **Thermal Embeddings**: The LLVIP dataset is used for training thermal embeddings. This dataset provides thermal imaging data, adding another layer of information to the model's understanding of the images.

◉ ImageBind Performance



audio classification depth classification

(https://ai.meta.com/blog/imagebind-six-modalities-binding-ai/)

◉ Modality-Specific Encoder

- For images and videos, it uses the Vision Transformer (ViT). For video inputs, 2-frame video clips were sampled over a 2-second duration.
- The audio inputs are transformed into 2D mel-spectrograms using the method outlined in AST: Audio Spectrogram Transformer, which involves converting a 2-second audio sample at 26kHz. As the mel-spectrogram is a 2D signal similar to an image, a ViT model is used to process it.
- For texts, a recurrent neural network (RNN) or transformer is used as the encoder. The transformer takes the raw text as input and produces a sequence of hidden states, which are then aggregated to produce the audio embedding vector.
- The thermal and depth inputs are treated as 1-channel images where ViT-B and ViT-S encoders are used, respectively.

◉ Multimodal Capabilities

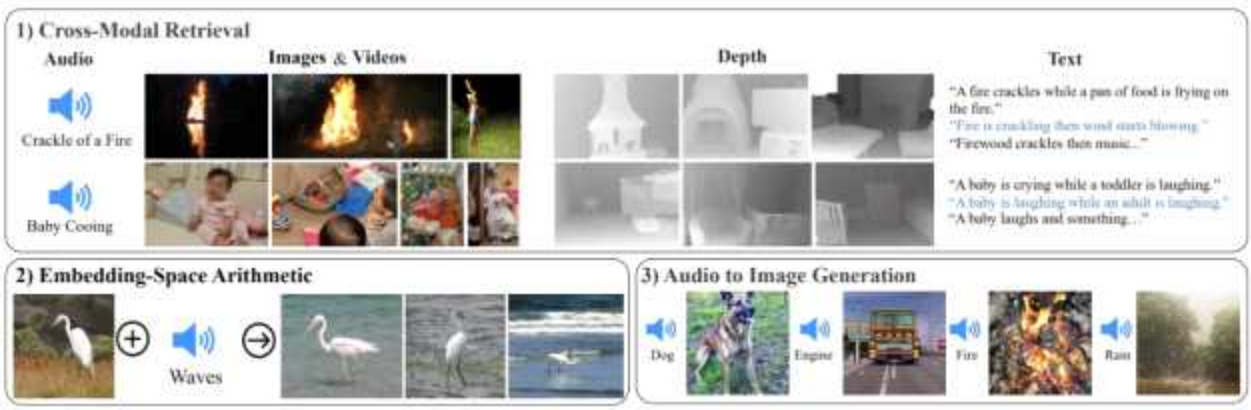


Figure 1. IMAGEBIND's joint embedding space enables novel multimodal capabilities. By aligning six modalities' embedding into a common space, IMAGEBIND enables: **1)** Cross-Modal Retrieval, which shows *emergent* alignment of modalities such as audio, depth or text, that aren't observed together. **2)** Adding embeddings from different modalities naturally composes their semantics. And **3)** Audio-to-Image generation, by using our audio embeddings with a pre-trained DALLE-2 [60] decoder designed to work with CLIP text embeddings.

◉ **Multimodal Embedding/Learning의 미래**

- 보다 다양한 Modality로의 발전
  : touch, smell, brain signals, emotion
- 보다 적은 data의 사용
  : joint embedding space의 사용으로 학습이 가능하기 때문에
- 보다 다양한 분야에의 AI 적용
  : AI + (neuroscience, linguistics, cognitive science, etc.)

7. 앞으로,,,,,,

◉ 경향

- multi-modal embedding : Multimodal chatGPT-4o
- model의 경량화 : Edge AI model Octopus-v2 (Stanford, 2024)

◉ Chameleon: Mixed-Modal Early-Fusion Foundation Models (2024)

.....

◉ AI의 가능성과 발전 속도를 예단하지 말라!!!

◉ Lifelong Learning의 시대