Self-supervised Learning for scaling to more modalities and data

Ishan Misra

GenAI@Meta

BigMAC Workshop, ICCV 2023
The era of multimodal learning

- Get billions of (image, text) pairs
- Learn representations that “align” images with text

A pineapple sitting on the counter

Image source: CLIP - Radford et al., 2021
Aligned image-text features

- Aligned representations are really useful

Image-text retrieval
Open-vocabulary classification[1]

Open-vocabulary detection and segmentation[2]

Text to image generation[3]

[1] CLIP - Radford et al., 2021
[3] GLIDE - Nichol et al., 2022, LAFITE - Zhou et al., 2022
Does SSL Matter?!

• Especially in this era of strong image features from (image, text)?
• Scaling (image, text) data is the way forward?
Standalone SSL is scaling well

DINOv2 - Oquab et al., 2023
SSL vs. Weakly supervised Debate

Image credit - Wikimedia
SSL vs. and Weakly supervised Debate
SSL
Self-Supervised Learning

Ex: Image Reconstruction (MAE)

WSP
Weakly Supervised Pretraining

Ex: Noisy Label Supervision (SwAG)
Great potential on diverse downstream tasks

Great fine-tuning classification performance

Great on dense prediction tasks like detection (ViTDeT)

Basis for SOTA foundational models

SOTA for classification (fine-tuning)

SOTA Zero Shot Capabilities (CLIP, LiT)
The effectiveness of MAE pre-pretraining for billion scale pretraining

Mannat Singh*, Quentin Duval*, Kalyan Vasudev Alwala*, Haoqi Fan, Vaibhav Aggarwal, Aaron Adcock, Armand Joulin, Piotr Dollár, Christoph Feichtenhofer, Ross Girshick, Rohit Girdhar, Ishan Misra
Key idea

• Introduce a “pre” pre-training stage
• Pre-pretraining uses self-supervised learning (no labels)
• Initialize and train as usual
Pre-pretraining

**Step 1:** Pre-pretraining
- Use Masked AutoEncoders (MAE)
- Low FLOPs (75% masking)

**Step 2:** Standard weakly supervised training
- Use image labels
- Multi-target prediction (no contrastive learning!)
- Simple yet SOTA
Pre-pretraining at scale

**Dataset**: Instagram-3B
- 3B unique images
- 5B images after resampling

For weakly-supervised
- 28K unique hashtags

**Architecture**: ViT up to 6.5B params
MAE scales with **both** data and model.

He et al., 2022 showed it scaled only with model size.
Pre-pretraining matters at large scale too!

IN1k linear probe (accuracy)

- Improves performance across all model & data sizes
Pre-pretraining matters at large scale too!

- More efficient! → Better performance at fewer FLOPs
Best of SSL and WSP

**MAE** shines on **dense prediction tasks**

**WSP** shines on **classification tasks**

**MAE→WSP** combines their strengths
## Pushing the state-of-the-art

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Evaluation</th>
<th>Top-1 Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>iNaturalist-18</td>
<td>Fine-tuning</td>
<td>91.3%</td>
</tr>
<tr>
<td>ImageNet1k</td>
<td>1-shot</td>
<td>62.1%</td>
</tr>
<tr>
<td>Food101</td>
<td>0-shot</td>
<td>96.2%</td>
</tr>
<tr>
<td>Object Net</td>
<td>OOD eval</td>
<td>75.8%</td>
</tr>
</tbody>
</table>
Multi-modal != Bi-modal
There are other modalities ...
Aligned data is hard to get

Depth

Thermal

Motion (IMU)

Audio

Image source: Rawpixel, The Rijksmuseum
Solution 1: Single model

Omnivore: A Single Model for Many Visual Modalities

Solution 1: Single model

Omnivore: A Single Model for Many Visual Modalities
Omnivore: Cross-modal alignment emerges!
Images are a universal language

Depth

Thermal

Motion (IMU)

Audio

RGB

RGB

RGB

RGB

Image source (L to R): SUN RGB-D, LLVIP, Isaque Pereira, Ego4D, Wikimedia, Gabriel Peter
Images are a universal language

Naturally co-occurring “Self-supervised”
ImageBind: One Embedding to Rule them All

Rohit Girdhar*, Alaaeldin El-Nouby*, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, Ishan Misra*

https://github.com/facebookresearch/ImageBind
Key Idea

• Images naturally co-occur with different modalities
• Align every modality’s representation with images
• Heavily leverage self-supervised learning
Emergent behavior (Transitive alignment!)

- After training all modalities are aligned
Training setup

• 6 modalities — Image/Video, Text, Audio, Depth, IMU, Thermal
• Train only with image-paired data
• Separate encoder per modality
• Initialize image & text encoder from CLIP/OpenCLIP and keep frozen
Measuring emergent alignment to text

- Train on (Image, X) (Image, Text)
- Test on (X, Text) —> “Emergent” zero-shot classification

<table>
<thead>
<tr>
<th>Image</th>
<th>Video</th>
<th>Depth</th>
<th>Audio</th>
<th>Thermal</th>
<th>IMU</th>
</tr>
</thead>
<tbody>
<tr>
<td>IN1k</td>
<td>P365</td>
<td>K400</td>
<td>MSVTT</td>
<td>NYU</td>
<td>SUN</td>
</tr>
<tr>
<td>Random</td>
<td>0.1</td>
<td>0.27</td>
<td>0.25</td>
<td>0.1</td>
<td>10.0</td>
</tr>
<tr>
<td>ImageBind</td>
<td>77.7</td>
<td>45.4</td>
<td>50.0</td>
<td>36.1</td>
<td>54.0</td>
</tr>
<tr>
<td>Text paired</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>41.9</td>
</tr>
<tr>
<td>Absolute SOTA</td>
<td>91.0</td>
<td>60.7</td>
<td>89.9</td>
<td>57.7</td>
<td>76.7</td>
</tr>
</tbody>
</table>

SOTA: 91.0  60.7  89.9  57.7

IN1k: ImageNet 1k; P365: PASCAL 365; K400: KITTI 400; MSVTT: MSV mastery training; NYU: NYU; SUN: SUN; AudioSet: AudioSet; VGGs: VGGs; ESC: ESC; LLVIP: LLVIP; Ego4D: Ego4D
ImageBind for “upgrading” existing models

Only takes text inputs

Your Favorite Model

Text
ImageBind for “upgrading” existing models

Your Favorite Model

Only takes text inputs

“Upgrade”

No re-training

“Multi” Modal

New modality

Your Favorite Model

Only takes text inputs

“Upgrade”

No re-training

“Multi” Modal

New modality
Audio-based prompting for image generation

Rain

Bark

Fire

Engine
Aligned embeddings can be “added”
Thanks!

**ImageBind**

- Code & Models released
  - [https://imagebind.metademolab.com/](https://imagebind.metademolab.com/)

**Effectiveness of MAE Pre-pretraining**

- Poster session (Wednesday)
  - Code & Models
    - [https://github.com/facebookresearch/maws](https://github.com/facebookresearch/maws)

**MOST: Unsupervised Object Discovery**

- Poster session (Friday)
  - Code & Models
    - [https://github.com/rssaketh/MOST/](https://github.com/rssaketh/MOST/)