Masked Auto-Encoders as Scalable Vision Learners

Xinlei Chen

ECCV 2022 tutorial on self-supervised representation learning in computer vision
$f(x) \quad \checkmark \quad \rightarrow \quad f(\lambda) \quad \times$
$f(x)$ → $f(x)$ → $f(x)$
Self-supervised learning
Self-Supervised Learning

• Pre-train representations without labels for downstream tasks
Self-Supervised Learning

• Pre-train representations without labels for downstream tasks
Self-Supervised Learning

• Pre-train representations without labels for downstream tasks

unsupervised pre-training

image classification

object detection and segmentation

human understanding
Self-Supervised Representation Learning

- Pre-train representations without labels for downstream tasks

Unsupervised pre-training

- Image classification
- Object detection and segmentation
- Human understanding
Self-Supervised Representation Learning

• **Scalable**: use unlimited data to train unlimited-sized models
Self-Supervised Representation Learning

- **Scalable**: use unlimited data to train unlimited-sized models
- Tremendously successful in NLP

[Devlin et al, NAACL 2019] [Brown et al, NeurIPS 2020]
Self-Supervised Representation Learning

- **Scalable**: use unlimited data to train unlimited-sized models
- Tremendously successful in NLP

[Devlin et al, NAACL 2019] [Brown et al, NeurIPS 2020]
Self-Supervised Paradigms Covered

- Contrastive / Siamese

→ Tutorial from Ting Chen
   SimCLR 1\textsuperscript{st} author, Google
   5:30 pm – 6:15pm

[Chen et al, ICML 2020] [He et al, CVPR 2022]
Self-Supervised Paradigms Covered

• Contrastive / Siamese

\[ x \xrightarrow{x'} x'' \]

→ Tutorial from Ting Chen

SimCLR 1st author, Google

5:30 pm – 6:15pm

• Reconstructive / Auto-Encoding
Self-Supervised Paradigms Covered

• Contrastive / Siamese

→ Tutorial from Ting Chen
  SimCLR 1st author, Google
  5:30 pm – 6:15pm

• Reconstructive / Auto-Encoding

Masked Auto-Encoders Are Scalable Vision Learners:
  Kaiming, Xinlei, Saining, Yanghao, Piotr, Ross
  CVPR 2022
What is MAE?

• Very simple method, but highly effective
What is MAE?

• Very simple method, but highly effective

• BERT-like algorithm, but with crucial design changes for vision

[Devlin et al, NAACL 2019] [He et al, CVPR 2022]
What is MAE?

• Very simple method, but highly effective

• BERT-like algorithm, but with crucial design changes for vision

• Intriguing properties – better scalability and more from analysis

[Devlin et al, NAACL 2019] [He et al, CVPR 2022]
How MAE Works?

Random masking
How MAE Works?

Encode visible patches
How MAE Works?

Add mask tokens
How MAE Works?

Reconstruct
MAE Reconstruction Example

Masked input: 80%  You guess?
MAE Reconstruction Example

Masked input: 80%  MAE’s guess
MAE Reconstruction Example

Masked input: 80%  MAE’s guess  Ground truth
MAE Can Generalize
MAE Can Generalize
MAE Can Generalize
MAE Can Generalize
MAE Can Generalize
MAE Can Generalize

original

75% mask

85% mask

95% mask
BERT-like: Transformers

• Vision Transformer (ViT)
  • Less inductive bias
  • Non-overlapping tokenization
    • Easier for masked auto-encoding

[ Dosovitskiy et al, ICLR 2021 ]
BERT-like: Transformers

• Vision Transformer (ViT)
  • Less inductive bias
  • Non-overlapping tokenization
    • Easier for masked auto-encoding

• **Scalable**
  • with larger models
  • on larger datasets

[Dosovitskiy et al, ICLR 2021]
BERT-like: Transformers

- Vision Transformer (ViT)
  - Less inductive bias
  - Non-overlapping tokenization
    - Easier for masked auto-encoding
- Scalable
  - with larger models
  - on larger datasets

[Dosovitskiy et al, ICLR 2021]
BERT-unlike: Mask Ratio

• BERT: 15% is enough to create a challenging task
• MAE: a high ratio of 75% - 80% is about optimal

[Devlin et al, NAACL 2019]
BERT-unlike: Encoder-Decoder

• BERT: encoder-only pre-training

[Devlin et al, NAACL 2019]
BERT-unlike: Encoder-Decoder

- **MAE:**
  - *Large* encoder on *visible* tokens
  - Small decoder on *all* tokens
  - *Projection* layer to connect the two

[Devlin et al, NAACL 2019]
BERT-unlike: Encoder-Decoder

- **MAE:**
  - Large encoder on *visible* tokens
  - Small decoder on *all* tokens
  - *Projection* layer to connect the two

- Very efficient when coupled with high mask ratio (75%)
MAE for Downstream Tasks: *Encoder Only*

- After MAE pre-training, just *throw away* the decoder
- Encoder is used for representations with *full-sequence* input
Experimental Protocols

• Pre-training dataset: ImageNet-1K

• Architecture: ViT-\textit{Large} encoder, 512-dim decoder
Experimental Protocols

• Pre-training dataset: ImageNet-1K

• Architecture: ViT-\textit{Large} encoder, 512-dim decoder

• Transfer task: ImageNet-1K classification
  • \textit{"ft"}: end-to-end tuning with MAE as an initialization
  • \textit{"lin"}: linear probing, a single classifier on top of frozen encoder features
Analysis: Decoder Size

- Encoder has 24-blocks, 1024-dimensional

<table>
<thead>
<tr>
<th>blocks</th>
<th>ft</th>
<th>lin</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>84.8</td>
<td>65.5</td>
</tr>
<tr>
<td>2</td>
<td><strong>84.9</strong></td>
<td>70.0</td>
</tr>
<tr>
<td>4</td>
<td><strong>84.9</strong></td>
<td>71.9</td>
</tr>
<tr>
<td>8</td>
<td><strong>84.9</strong></td>
<td><strong>73.5</strong></td>
</tr>
<tr>
<td>12</td>
<td>84.4</td>
<td>73.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>dim</th>
<th>ft</th>
<th>lin</th>
</tr>
</thead>
<tbody>
<tr>
<td>128</td>
<td><strong>84.9</strong></td>
<td>69.1</td>
</tr>
<tr>
<td>256</td>
<td>84.8</td>
<td>71.3</td>
</tr>
<tr>
<td>512</td>
<td><strong>84.9</strong></td>
<td><strong>73.5</strong></td>
</tr>
<tr>
<td>768</td>
<td>84.4</td>
<td>73.1</td>
</tr>
<tr>
<td>1024</td>
<td>84.3</td>
<td>73.1</td>
</tr>
</tbody>
</table>

Decoder depth

Decoder width
Analysis: Mask Ratio

- **fine-tuning**
- **linear probing**

![Graph showing masking ratio (%) vs. accuracy for fine-tuning and linear probing.](image-url)
Analysis: Mask Token $[M]$ in Encoder

- Encoder w/ $[M]$ is default in BERT
- Big domain gap for linear probing
  - Pre-train sees 25% of the images only, while evaluation sees 100%

<table>
<thead>
<tr>
<th>case</th>
<th>ft</th>
<th>lin</th>
<th>FLOPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>encoder w/ $[M]$</td>
<td>84.2</td>
<td>59.6</td>
<td>3.3x</td>
</tr>
<tr>
<td>encoder w/o $[M]$</td>
<td>84.9</td>
<td>73.5</td>
<td>1x</td>
</tr>
</tbody>
</table>
Analysis: Reconstruction Target

- Pixels with normalization: per-patch -- minus mean, divide by std
- PCA: only low-frequency component is retained
- dVAE token: from DALLE, expensive to compute

<table>
<thead>
<tr>
<th>case</th>
<th>ft</th>
<th>lin</th>
</tr>
</thead>
<tbody>
<tr>
<td>pixel (w/o norm)</td>
<td>84.9</td>
<td>73.5</td>
</tr>
<tr>
<td>pixel (w/ norm)</td>
<td>85.4</td>
<td>73.9</td>
</tr>
<tr>
<td>PCA</td>
<td>84.6</td>
<td>72.3</td>
</tr>
<tr>
<td>dVAE token</td>
<td>85.3</td>
<td>71.6</td>
</tr>
</tbody>
</table>

[Ramesh et al, ICML 2021] [Bao et al, ICLR 2022]
Analysis: Augmentations

<table>
<thead>
<tr>
<th>case</th>
<th>ft</th>
<th>lin</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>84.0</td>
<td>65.7</td>
</tr>
<tr>
<td>crop, fixed size</td>
<td>84.7</td>
<td>73.1</td>
</tr>
<tr>
<td>crop, rand size</td>
<td><strong>84.9</strong></td>
<td><strong>73.5</strong></td>
</tr>
<tr>
<td>crop + color jit</td>
<td>84.3</td>
<td>71.9</td>
</tr>
</tbody>
</table>

- MAE can work with minimal data augmentation
Analysis: Augmentations

<table>
<thead>
<tr>
<th>case</th>
<th>ft</th>
<th>lin</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>84.0</td>
<td>65.7</td>
</tr>
<tr>
<td>crop, fixed size</td>
<td>84.7</td>
<td>73.1</td>
</tr>
<tr>
<td>crop, rand size</td>
<td><strong>84.9</strong></td>
<td><strong>73.5</strong></td>
</tr>
<tr>
<td>crop + color jit</td>
<td>84.3</td>
<td>71.9</td>
</tr>
</tbody>
</table>

- MAE can work with minimal data augmentation
- For Siamese learning, augmentation is crucial
Scalability: Longer Training

![Graphs showing scalability over longer training periods for fine-tuning and linear probing, with epochs on a log-scale.](image)
Scalability: Longer Training

Wall-clock speed still efficient thanks to MAE design
Scalability: Larger Models

![Graph showing fine-tuning accuracy vs. params (M) for different models: ViT-B/16, ViT-L/16, ViT-H/14. The graph includes lines for MAE, IN1K, supervised, IN1K, our impl., supervised, IN1K [16], and supervised, JFT300M [16].]
Scalability: Larger Models
Scalability: Larger Models
Scalability: Larger Models
Scalability: Larger Models

new SOTA on ImageNet-1K (no extra data): 87.8%
### Scalability: Larger Models

<table>
<thead>
<tr>
<th>dataset</th>
<th>ViT-B</th>
<th>ViT-L</th>
<th>ViT-H</th>
<th>ViT-H\textsubscript{448}</th>
<th>prev best</th>
</tr>
</thead>
<tbody>
<tr>
<td>iNat 2017</td>
<td>70.5</td>
<td>75.7</td>
<td>79.3</td>
<td>83.4</td>
<td>75.4 [50]</td>
</tr>
<tr>
<td>iNat 2018</td>
<td>75.4</td>
<td>80.1</td>
<td>83.0</td>
<td>86.8</td>
<td>81.2 [49]</td>
</tr>
<tr>
<td>iNat 2019</td>
<td>80.5</td>
<td>83.4</td>
<td>85.7</td>
<td>88.3</td>
<td>84.1 [49]</td>
</tr>
<tr>
<td>Places205</td>
<td>63.9</td>
<td>65.8</td>
<td>65.9</td>
<td>66.8</td>
<td>66.0 [19]\dagger</td>
</tr>
<tr>
<td>Places365</td>
<td>57.9</td>
<td>59.4</td>
<td>59.8</td>
<td>60.3</td>
<td>58.0 [36]\ddagger</td>
</tr>
</tbody>
</table>

**new SOTA on 5 large-scale classification datasets**

<table>
<thead>
<tr>
<th>dataset</th>
<th>ViT-B</th>
<th>ViT-L</th>
<th>ViT-H</th>
<th>ViT-H\textsubscript{448}</th>
<th>prev best</th>
</tr>
</thead>
<tbody>
<tr>
<td>IN-Corruption \dagger [27]</td>
<td>51.7</td>
<td>41.8</td>
<td>33.8</td>
<td>36.8</td>
<td>42.5 [32]</td>
</tr>
<tr>
<td>IN-Adversarial [28]</td>
<td>35.9</td>
<td>57.1</td>
<td>68.2</td>
<td>76.7</td>
<td>35.8 [41]</td>
</tr>
<tr>
<td>IN-Rendition [26]</td>
<td>48.3</td>
<td>59.9</td>
<td>64.4</td>
<td>66.5</td>
<td>48.7 [41]</td>
</tr>
<tr>
<td>IN-Sketch [60]</td>
<td>34.5</td>
<td>45.3</td>
<td>49.6</td>
<td>50.9</td>
<td>36.0 [41]</td>
</tr>
</tbody>
</table>

**new SOTA on 4 ImageNet robust evaluations**
Scalability: Larger Models

<table>
<thead>
<tr>
<th>method</th>
<th>pre-train data</th>
<th>ViT-B</th>
<th>ViT-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>supervised</td>
<td>IN1K w/ labels</td>
<td>47.9</td>
<td>49.3</td>
</tr>
<tr>
<td>MoCo v3</td>
<td>IN1K</td>
<td>47.9</td>
<td>49.3</td>
</tr>
<tr>
<td>BEiT</td>
<td>IN1K+DALLE</td>
<td>49.8</td>
<td>53.3</td>
</tr>
<tr>
<td>MAE</td>
<td>IN1K</td>
<td>50.3</td>
<td>53.3</td>
</tr>
</tbody>
</table>

COCO detection: +4.0%

<table>
<thead>
<tr>
<th>method</th>
<th>pre-train data</th>
<th>ViT-B</th>
<th>ViT-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>supervised</td>
<td>IN1K w/ labels</td>
<td>47.4</td>
<td>49.9</td>
</tr>
<tr>
<td>MoCo v3</td>
<td>IN1K</td>
<td>47.3</td>
<td>49.1</td>
</tr>
<tr>
<td>BEiT</td>
<td>IN1K+DALLE</td>
<td>47.1</td>
<td>53.3</td>
</tr>
<tr>
<td>MAE</td>
<td>IN1K</td>
<td>48.1</td>
<td>53.6</td>
</tr>
</tbody>
</table>

ADE20K segmentation: +3.7%
Scalability: Sequence Length

- Longer sequence length during pre-training, but fixed length during downstream transfers
Scalability: Sequence Length

- Longer sequence length during pre-training, but fixed length during downstream transfers
Is the Journey 99% Done?

- NLP has witnessed amazing progress in scaling since BERT
Is the Journey 99% Done?

• NLP has witnessed amazing progress in scaling since BERT

• It’s just **starting** in vision:
  • Temporal data – Christoph
  • Architectures – ConvNets?
  • Other modalities? 3D?
  • Other downstream tasks?
  • Other axes to scale?
  • [Your exploration] here!
Take-aways

• Self-supervised learning aims at *scalable* representation learning

• Masked auto-encoders can serve as scalable vision learners

• Exciting years ahead in this direction!

code (GPU): https://github.com/facebookresearch/mae

code (TPU): https://github.com/facebookresearch/long_seq_mae