MVP: Multimodality-guided Visual Pre-training

Longhui Wei[1,2], Lingxi Xie[2], Wengang Zhou[1], Houqiang Li[1], Qi Tian[2]
[1]University of Science and Technology of China, [2]Huawei Cloud.

Abstract

- Recently, masked image modeling (MIM) has become a promising direction for visual pre-training. In the context of vision transformers, MIM learns effective visual representation by aligning the token-level features with a pre-defined space.
- In this paper, we go one step further by introducing guidance from other modalities and validating that such additional knowledge leads to impressive gains for visual pre-training
- We demonstrate the effectiveness of the proposed method, e.g., our approach reports a 52.4% mIoU on ADE20K, surpassing BEIT with an impressive margin of 6.8%.

Contribution

- ✓ We analyze the recent masked image modeling based pre-training methods lack of semantics knowledge, and then firstly point out they can be enhanced with the guidance of other modalities.
- ✓ We design a simple yet effective algorithm to improve the transfer performance of MIM-based visual pre-training. By resorting to a tokenizer pre-trained with multimodal data, MVP learns richer semantic knowledge for each image.
- ✓ We evaluate the effectiveness of MVP with extensive experiments, and the results clearly demonstrate the advantages of MVP over the recently proposed visual pre-training methods.

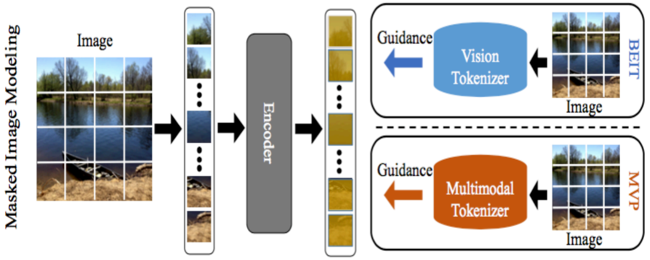
Motivation

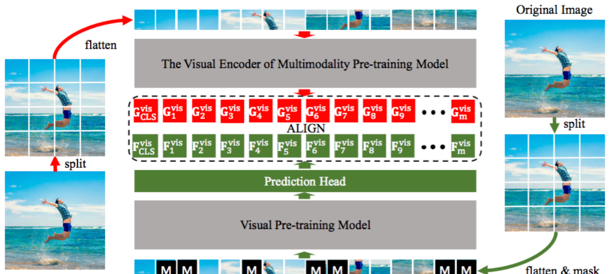
- MIM pre-training methods learn relatively weak semantic feature for visual representation.
- Multimodal data can provide more semantic knowledge. Therefore, how to investigate the use of multimodal pre-training on MIM framework is good direction to improve the semantics of pre-trained models.

Politician and people filling with donated bottles of water | Laughing family of four in a park | The skies above river serve as a canvas during this beautiful sunset

Proposed Approach (MVP)

- Instead of using a tokenizer that was pre-trained with pure image data, MVP replaces it with a tokenizer that is pre-trained with image-text pairs.





- Optimization Goal:

$$\mathcal{L}_{\mathrm{MVP}} \doteq -\frac{\langle \mathbf{F}_{\mathrm{CLS}}^{\mathrm{vis}}, \mathbf{G}_{\mathrm{CLS}}^{\mathrm{vis}} \rangle + \sum_{m=1}^{M} \langle \mathbf{F}_{m}^{\mathrm{vis}}, \mathbf{G}_{m}^{\mathrm{vis}} \rangle}{M+1},$$

where

G^{vis} denotes the extracted feature by CLIP: $\{\mathbf{G}_{\mathrm{CLS}}^{\mathrm{vis}}, \mathbf{G}_{1}^{\mathrm{vis}}, \ldots, \mathbf{G}_{M}^{\mathrm{vis}}\} = g^{\mathrm{vis}}(\{\mathbf{t}_{\mathrm{CLS}}, \mathbf{t}_1, \ldots, \mathbf{t}_M\}),$

F^{vis} denotes the predicted multimodal feature: $\{\mathbf{F}_{\mathrm{CLS}}^{\mathrm{vis}}, \mathbf{F}_{1}^{\mathrm{vis}}, \ldots, \mathbf{F}_{m}^{\mathrm{vis}}\} = f^{\mathrm{head}}(f^{\mathrm{vis}}(\{\mathbf{t}_{\mathrm{CLS}}, \mathbf{t}_1 \ldots, \hat{\mathbf{t}}_m, \ldots, \mathbf{t}_M\})),$

Experimental Results

- ✓ MVP enjoys advantages on image classification while fine-tuning different backbones.

Method	Model	Pre-training Epochs	Top-1 (%)
DINO [3]	ViT-B/16	300	82.8
BEIT [2]	ViT-B/16	800	83.2
MAE [16]	ViT-B/16	1600	83.6
PeCo [12]	ViT-B/16	300	84.1
MaskFeat [34]	ViT-B/16	1600	84.0
MVP (ours)	**ViT-B/16**	**300**	**84.4**
BEIT [2]	ViT-L/16	800	85.2
MAE [16]	ViT-L/16	1600	85.9
MaskFeat [34]	ViT-L/16	1600	85.7
MVP (ours)	**ViT-L/16**	**300**	**86.3**

- ✓ MVP achieves much better transfer performance on dense visual task, e.g., semantic segmentation task on ADE20K

Method	Model	Pre-training Epochs	mIoU (%)
DINO [3]	ViT-B/16	300	44.1
BEIT [2]	ViT-B/16	800	45.6
MAE [16]	ViT-B/16	1600	48.1
CAE [5]	ViT-B/16	800	48.8
PeCo [12]	ViT-B/16	300	46.7
MVP (ours)	**ViT-B/16**	**300**	**52.4**

- ✓ Beyond knowledge distillation

Guidance	Model	Epochs	ImageNet-1K(Top-1)	ADE20K(mIoU)
DINO	ViT-B/16	300	83.6	47.0
CLIP	ViT-B/16	300	84.4	52.4