

cases	doc_1		doc_2		decision	id
	authors	<ul style="list-style-type: none">Xiaotong LiYixiao GeKun YiZixuan HuYing ShanLing-Yu Duan	authors	<ul style="list-style-type: none">Duan, Ling-YuGe, YixiaoHu, ZixuanLi, XiaotongShan, YingYi, Kun	DUPLICATES	129
	title	mc-BEiT: Multi-choice Discretization for Image BERT Pre-training	title	mc-BEiT: Multi-choice Discretization for Image BERT Pre-training		
	publication_date	2022-03-29 09:08:18+00:00	publication_date	2022-04-09 00:00:00		
	source	SupportedSources.ARXIV	source	SupportedSources.CORE		
	journal	None	journal			
	volume		volume			
	doi		doi	None		
	urls	<ul style="list-style-type: none">http://arxiv.org/pdf/2203.15371v4http://arxiv.org/abs/2203.15371v4http://arxiv.org/pdf/2203.15371v4	urls	<ul style="list-style-type: none">http://arxiv.org/abs/2203.15371		
	id	id-3007661444976185232	id	id-1434096295655901912		
	abstract	Image BERT pre-training with masked image modeling (MIM) becomes a popular practice to cope with self-supervised representation learning. A seminal work, BEiT, casts MIM as a classification task with a visual vocabulary, tokenizing the continuous visual signals into discrete vision tokens using a pre-learned dVAE. Despite a feasible solution, the improper discretization hinders further improvements of image pre-training. Since image discretization has no ground-truth answers, we believe that the masked patch should not be assigned with a unique token id even if a better tokenizer can be obtained. In this work, we introduce an improved BERT-style image pre-training method, namely mc-BEiT, which performs MIM proxy tasks towards eased and refined multi-choice training objectives. Specifically, the multi-choice supervision for the masked image patches is formed by the soft probability vectors of the discrete token ids, which are predicted by the off-the-shelf image tokenizer and further refined by high-level inter-patch perceptions resorting to the observation that similar patches should share their choices. Extensive experiments on classification, segmentation, and detection tasks demonstrate the superiority of our method, e.g., the pre-trained ViT-B achieves 84.1% top-1 fine-tuning accuracy on ImageNet-1K classification, 49.2% AP^b and 44.0% AP^m of object detection and instance segmentation on COCO, 50.8% mIOU on ADE20K semantic segmentation, outperforming the competitive counterparts. The code will be available at https://github.com/lixiaotong97/mc-BEiT.	abstract	Image BERT pre-training with masked image modeling (MIM) becomes a popular practice to cope with self-supervised representation learning. A seminal work, BEiT, casts MIM as a classification task with a visual vocabulary, tokenizing the continuous visual signals into discrete vision tokens using a pre-learned dVAE. Despite a feasible solution, the improper discretization hinders further improvements of image pre-training. Since image discretization has no ground-truth answers, we believe that the masked patch should not be assigned with a unique token id even if a better tokenizer can be obtained. In this work, we introduce an improved BERT-style image pre-training method, namely mc-BEiT, which performs MIM proxy tasks towards eased and refined multi-choice training objectives. Specifically, the multi-choice supervision for the masked image patches is formed by the soft probability vectors of the discrete token ids, which are predicted by the off-the-shelf image tokenizer and further refined by high-level inter-patch perceptions resorting to the observation that similar patches should share their choices. Extensive experiments on classification, segmentation, and detection tasks demonstrate the superiority of our method, e.g., the pre-trained ViT-B achieves 84.1% top-1 fine-tuning accuracy on ImageNet-1K classification, 50.8% mIOU on ADE20K semantic segmentation, 51.2% AP^b and 44.3% AP^m of object detection and instance segmentation on COCO, outperforming the competitive counterparts		
	versions		versions			