<table>
<tr><th rowspan="2">cases</th><th>doc_1</th><th>doc_2</th><th>decision</th><th>id</th></tr>
<tr>
<td>
<table>
<tr><th>authors</th><td>Xiaotong Li
Yixiao Ge
Kun Yi
Zixuan Hu
Ying Shan
Ling-Yu Duan</td></tr>
<tr><th>title</th><td>mc-BEiT: Multi-choice Discretization for Image BERT Pre-training</td></tr>
<tr><th>publication_date</th><td>2022-07-28 00:00:00</td></tr>
<tr><th>source</th><td>SupportedSources.INTERNET_ARCHIVE</td></tr>
<tr><th>journal</th><td></td></tr>
<tr><th>volume</th><td></td></tr>
<tr><th>doi</th><td></td></tr>
<tr><th>urls</th><td>https://web.archive.org/web/20220729034617/https://arxiv.org/pdf/2203.15371v4.pdf</td></tr>
<tr><th>id</th><td>id9120167953335425519</td></tr>
<tr><th>abstract</th><td>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.</td></tr>
<tr><th>versions</th><td></td></tr>
</table>
</td>
<td>
<table>
<tr><th>authors</th><td>Duan, Ling-Yu
Ge, Yixiao
Hu, Zixuan
Li, Xiaotong
Shan, Ying
Yi, Kun</td></tr>
<tr><th>title</th><td>mc-BEiT: Multi-choice Discretization for Image BERT Pre-training</td></tr>
<tr><th>publication_date</th><td>2022-04-09 00:00:00</td></tr>
<tr><th>source</th><td>SupportedSources.CORE</td></tr>
<tr><th>journal</th><td></td></tr>
<tr><th>volume</th><td></td></tr>
<tr><th>doi</th><td>None</td></tr>
<tr><th>urls</th><td>http://arxiv.org/abs/2203.15371</td></tr>
<tr><th>id</th><td>id-1434096295655901912</td></tr>
<tr><th>abstract</th><td>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</td></tr>
<tr><th>versions</th><td></td></tr>
</table>
</td>
<td>DUPLICATES</td>
<td>130</td>
</tr>
</table>