<table>
<tr><th></th><th>doc_1</th><th>doc_2</th><th>decision</th><th>id</th></tr>
<tr><th>cases</th>
<td>
<table>
<tr><th>authors</th><td>

- Xiaoyi Dong
- Jianmin Bao
- Ting Zhang
- Dongdong Chen
- Weiming Zhang
- Lu Yuan
- Dong Chen
- Fang Wen
- Nenghai Yu

</td></tr>
<tr><th>title</th><td>PeCo: Perceptual Codebook for BERT Pre-training of Vision Transformers</td></tr>
<tr><th>publication_date</th><td>2021-11-24 00:00:00</td></tr>
<tr><th>source</th><td>SupportedSources.SEMANTIC_SCHOLAR</td></tr>
<tr><th>journal</th><td>ArXiv</td></tr>
<tr><th>volume</th><td>abs/2111.12710</td></tr>
<tr><th>doi</th><td></td></tr>
<tr><th>urls</th><td>

- https://www.semanticscholar.org/paper/3e38f4b4055abecbac2e618df2ecb33554073e08

</td></tr>
<tr><th>id</th><td>id-2127133131246859522</td></tr>
<tr><th>abstract</th><td>This paper explores a better prediction target for BERT pre- training of vision transformers. We observe that current prediction targets disagree with human perception judgment. This contradiction motivates us to learn a perceptual prediction target. We argue that perceptually similar images should stay close to each other in the prediction target space. We surprisingly ï¬nd one simple yet effective idea: enforcing percep- tual similarity during the dVAE training. Moreover, we adopt a self-supervised transformer model for deep feature extrac- tion and show that it works well for calculating perceptual similarity. We demonstrate that such learned visual tokens in- deed exhibit better semantic meanings, and help pre-training achieve superior transfer performance in various downstream tasks. For example, we achieve 84.5% Top-1 accuracy on ImageNet-1K with ViT-B backbone, outperforming the com- petitive method BEiT by +1.3% under the same pre-training epochs. Our approach also gets signiï¬cant improvement on object detection and segmentation on COCO and semantic segmentation on ADE20K. Equipped with a larger backbone ViT-H, we achieve the state-of-the-art ImageNet accuracy (88.3%) among methods using only ImageNet-1K data.</td></tr>
<tr><th>versions</th><td></td></tr>
</table>
</td>
<td>
<table>
<tr><th>authors</th><td>

- Xiaoyi Dong
- Jianmin Bao
- Ting Zhang
- Dongdong Chen
- Weiming Zhang
- Lu Yuan
- Dong Chen
- Fang Wen
- Nenghai Yu
- Baining Guo

</td></tr>
<tr><th>title</th><td>PeCo: Perceptual Codebook for BERT Pre-training of Vision Transformers</td></tr>
<tr><th>publication_date</th><td>2021-11-24 18:59:58+00:00</td></tr>
<tr><th>source</th><td>SupportedSources.ARXIV</td></tr>
<tr><th>journal</th><td>None</td></tr>
<tr><th>volume</th><td></td></tr>
<tr><th>doi</th><td></td></tr>
<tr><th>urls</th><td>

- http://arxiv.org/pdf/2111.12710v3
- http://arxiv.org/abs/2111.12710v3
- http://arxiv.org/pdf/2111.12710v3

</td></tr>
<tr><th>id</th><td>id-1611205866241915138</td></tr>
<tr><th>abstract</th><td>This paper explores a better prediction target for BERT pre-training of vision transformers. We observe that current prediction targets disagree with human perception judgment.This contradiction motivates us to learn a perceptual prediction target. We argue that perceptually similar images should stay close to each other in the prediction target space. We surprisingly find one simple yet effective idea: enforcing perceptual similarity during the dVAE training. Moreover, we adopt a self-supervised transformer model for deep feature extraction and show that it works well for calculating perceptual similarity.We demonstrate that such learned visual tokens indeed exhibit better semantic meanings, and help pre-training achieve superior transfer performance in various downstream tasks. For example, we achieve $\textbf{84.5\%}$ Top-1 accuracy on ImageNet-1K with ViT-B backbone, outperforming the competitive method BEiT by $\textbf{+1.3\%}$ under the same pre-training epochs. Our approach also gets significant improvement on object detection and segmentation on COCO and semantic segmentation on ADE20K. Equipped with a larger backbone ViT-H, we achieve the state-of-the-art ImageNet accuracy (\textbf{88.3\%}) among methods using only ImageNet-1K data.</td></tr>
<tr><th>versions</th><td></td></tr>
</table>
</td>
<td>DUPLICATES</td>
<td>240</td>
</tr>
</table>