Towards Fine-grained Visual Representations by Combining Contrastive Learning with Image Reconstruction and Attention-weighted Pooling

Jonas Dippel 1  Steffen Vogler 1  Johannes Höhne 1

1Bayer AG

Abstract

- This paper presents Contrastive Reconstruction, ConRec - a self-supervised learning algorithm that obtains image representations by jointly optimizing a contrastive and a self-reconstruction loss.
- State-of-the-art contrastive learning methods (i.e. SimCLR) have shortcomings with regard to fine-grained classification tasks.
- ConRec tackles these shortcomings and extends the SimCLR framework by adding (1) a self-reconstruction task, (2) an attention mechanism within the contrastive learning task.
- This is accomplished by applying a simple encoder-decoder architecture with two heads.
- We show that both extensions contribute towards an improved vector representation for images with fine-grained visual features.

Method

In the training process, the model receives a masked image $\tilde{x}_i$ and outputs the reconstructed image $\tilde{x}_i$ as well as the contrastive vector representation $z_i = p_c(e(x_i))$. The training loss is composed of two parts: the contrastive loss $L_c$ and the reconstruction loss $L_r$.

Attention Pooling

- Global average pooling discards some local features in the encoder output activation map, which may carry relevant fine-grained information.
- We introduce an attention-weighted pooling mechanism that aggregates the spatial content of the final feature map of the encoder in a parametric manner.

Augmentations

<table>
<thead>
<tr>
<th>Input</th>
<th>Prediction</th>
<th>Target</th>
</tr>
</thead>
</table>

Figure 1: Augmented images and respective reconstruction predictions by our ConRec model.

Figure 2: Learning Framework for Contrastive Reconstruction - ConRec. The ConRec model consists of a fully convolutional encoder-decoder architecture with skip connections as well as a projection head comprising fully connected layers. The model receives a masked image $\tilde{x}_i$ and outputs the unmasked reconstruction target $\tilde{x}_i$ as well as the contrastive image representation vector $z_i$.

Conv2D + BatchNorm
Attention-weighted Pooling
Fully Connected

Table 1: Linear evaluation results and respective baselines. ImageNet results in parenthesis indicate flaws in the evaluation as the datasets were included in supervised ImageNet-pretraining.