PatchIt: Self-Supervised Network Weight Initialization for Fine-grained Recognition

Patrick Sudowe
sudowe@vision.rwth-aachen.de
Bastian Leibe
leibe@vision.rwth-aachen.de

Visual Computing Institute
RWTH Aachen University
Germany

ConvNet training is highly sensitive to initialization of the weights. A widespread approach is to initialize the network with weights trained for a different task, an *auxiliary task*. The ImageNet-based ILSVRC classification task is a very popular choice for this, as it has shown to produce powerful feature representations applicable to a wide variety of tasks. However, this creates a significant entry barrier to exploring non-standard architectures. In this paper, we propose a self-supervised pretraining, the *PatchTask*, to obtain weight initializations for fine-grained recognition problems, such as person attribute recognition, pose estimation, or action recognition. Our pretraining allows us to leverage additional unlabeled data from the same source, which is often readily available, such as detection bounding boxes. We experimentally show that our method outperforms a standard random initialization by a considerable margin and closely matches the ImageNet-based initialization.

The *PatchTask* presented in this paper provides a viable alternative to the popular ImageNet-based pretraining. The core idea is to leverage data from the *same domain* as the target task for pretraining. The pretraining is self-supervised, *i.e.*, it solely relies on automatically generated rather than human annotated labels. We target fine-grained recognition tasks that appear in person analysis applications (*e.g.*, pose estimation, re-identification, action and attribute recognition). Their common aspect is that they make predictions for an object that has been located before (*e.g.*, by a detector). So, we will assume such a specific input domain.

The *PatchTask* idea is inspired by the work of Doersch *et al*. [1], who propose an auxiliary task defined by the spatial layout of pairs of patches. In contrast to their work on general images, we focus on fine-grained recognition, where the input images come from a *restricted data domain* (*i.e.*, bounding boxes showing persons). In this restricted setting, it is feasible to directly predict the original location of single patches (Fig. 1).

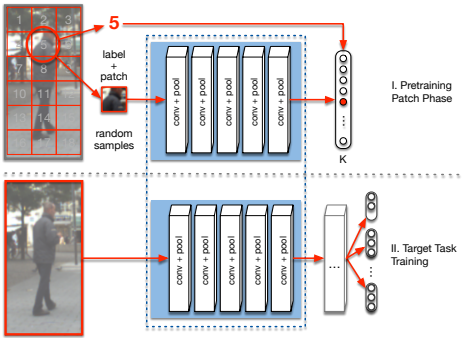

Figure 1: Patch Task: Classify the extraction position given one 32×32 pixel patch. During the pretraining phase, the model needs to encode local patch structure. The parameters are transferred to the target task net. Subsequent fine-tuning benefits from a better initialization.

This paper makes the following contributions: (1) We describe a family of self-supervised *patch tasks* for fine-grained analysis. (2) We demonstrate and evaluate their use for human attribute recognition, where we achieve state-of-the-art performance without using external labels (in particular, without ImageNet). This facilitates further exploration of architectures. (3) We provide data for person analysis pretraining and supporting code that may be used to improve person representations in other ConvNet architectures.

[1] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised Visual Representation Learning by Context Prediction. In *ICCV*, 2015.