**Can We Scale Transformers to Predict Parameters of Diverse ImageNet Models?**

Pretraining a neural network on a large dataset is becoming a cornerstone in machine learning

that is within the reach of only a few communities

with large-resources. We aim at an ambitious goal of democratising pretraining. Towards that goal, we train and release a single neural network that can predict high quality ImageNet parameters of other neural networks. By using predicted parameters for initialization, we can boost training of diverse ImageNet models available in PyTorch.

The parameters predicted by our GHNs show better transferability compared to GHN-2 in all but one of the transfer learning experiments (Tables 7 and 8). We also significantly improve on RANDINIT. By fine-tuning the predicted parameters on ImageNet for 1 epoch before transferring them to the CIFAR datasets we achieve further boosts

and outperform RANDINIT+SGD-1ep. While the gap with full ImageNet pretraining is still noticeable (e.g., 77.8% vs 88.7% for ResNet-50 on CIFAR-10), we narrow this gap compared to GHN-2 (68.4%) and RANDINIT (74.0%). When transferred to other datasets, models initialized with predicted parameters also converge faster and reach competitive final performance.

We improve Graph Hyper Networks by considerably scaling them up. By evaluating on realistic and challenging

ImageNet architectures, we found that scaling up gradually increases the overall performance. This is encouraging as further scaling GHNs can make them a powerful tool.

Our GHN-3 improves random-based and advanced initializations in vision experiments and makes a big step in the quality of predicted parameters compared to the prior GHNs.