CS482/682 Final Project Report Group 20

Unsupervised Learning for Image Classification using DINO

Vijay Murari Tiyyala/vtiyyal1, Lidia Berhe/lberhe2, Sam Lipschitz/slipsch3, Sean Murray/smurra42*

1 Introduction

Background Unsupervised learning for image classification has gained significant interest in recent years, with the aim to leverage unlabeled data for learning meaningful representations. This project focuses on investigating unsupervised learning techniques for image classification, in particular using DINOv2, a state-of-the-art method for learning robust visual features without supervision.

Related Work Various unsupervised and self-supervised learning methods have been proposed in recent years, such as SimCLR [1], MoCo [2], BYOL [3], and DINO [4]. DINOv2 [5] is a recent advancement in the field, improving upon the original DINO method [4].

2 Methods

Dataset We used the Tiny ImageNet dataset, which consists of 200 classes with 500 training images and 50 validation images per class. Images are downscaled to 64x64 pixels. We also set aside a test dataset for evaluation purposes.

Setup, Training, and Evaluation We started by exploring various unsupervised and self-supervised learning techniques and their associated architectures, such as SimCLR, Clustering and DINO. After preliminary experiments, we chose to focus on DINOv2 for its robust visual feature learning capabilities.

We initially attempted to use pre-trained CNN architectures, such as VGG, ResNet, and Vision Transformer (ViT) models to extract image representations. Our plan was to use these models to extract feature representations, and cluster based on these encodings. One specific approach we tried utilized Local Aggregation Loss [7]. Here, the goal is to move similar encodings togther, and dissimilar encodings apart. We calculate two sets of neighbors: background neighbors (i.e. nearest neighbors), and closest neighbors (i.e. neighbors in the same cluster).

We encountered several issues with this approach:

The image representations obtained from these architectures were not robust enough. The simple clustering method did not work effectively, possibly leading to imbalanced clusters. Tiny ImageNet's inherent diversity and hierarchical nature made it difficult for a single class to adequately represent some images. Due to computational constraints, we could not train our Local Aggregation Loss model for many epochs; however, it displayed very poor performance gain epoch to epoch, and would likely result with a max accuracy of 10%.

Consequently, we turned to DINOv2, a state-of-the-art unsupervised learning approach that leverages teacher-student training and has shown promising results in robust feature learning. We fine-tuned DINOv2 base and small models by adapting the classifier head to the Tiny ImageNet dataset. We also experimented with various data augmentation techniques, which proved crucial for the model's successful training.

We used the Adam optimizer and Cosine Anneal-

*Project code is available at `https://github.com/iMvijay23/Dinov2SSLImageCL/tree/main`

ing learning rate scheduler, the weight decay also follows a cosine schedule from 0.04 to 0.4 as recommended in the DINO paper. Additionally, we employed gradient accumulation and weight decay to enhance the training process further. After fine-tuning and evaluating the model on the Tiny ImageNet train and validation datasets, we achieved an accuracy of 84% on the test dataset.

3 Results

Our approach using the DINOv2 Small(21M) and DINOv2 Base(87M) models with custom data augmentation, linear classifier head, and optimization techniques achieved an accuracy of 84% on the test dataset. This result highlights the potential of DINOv2 for unsupervised learning in image classification tasks.

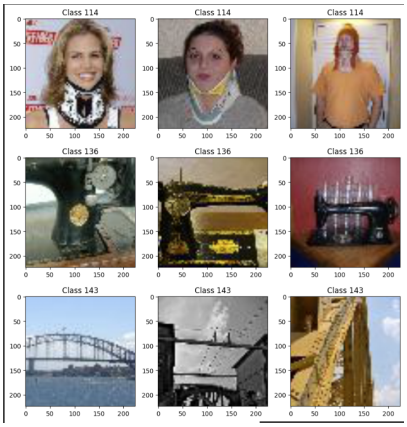

Figure 1: Prediction by finetuned DINO model

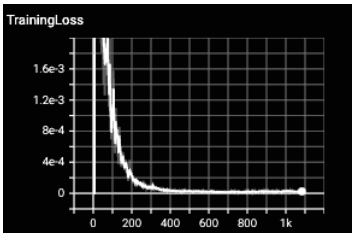

Figure 2: Train Loss

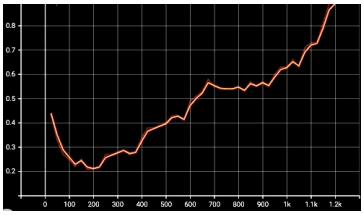

Figure 3: Validation Accuracy

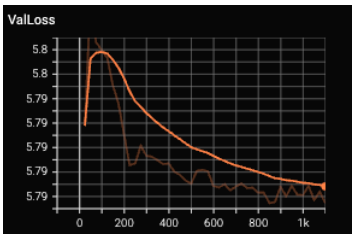

Figure 4: Validation Loss

4 Discussion

Our project demonstrates the effectiveness of using DINOv2 for unsupervised image classification tasks. The choice of data augmentation, optimization techniques, and model architecture played a significant role in achieving competitive performance.

We are also interested in exploring the combination of DINOv2 with the recently proposed SPiCe method [6], which could potentially lead to further improvements in unsupervised learning performance.

Overall, our project successfully showcases the potential of DINOv2 for unsupervised image classification tasks, and future work could focus on exploring other unsupervised learning techniques or model architectures, as well as investigating the combination of DINOv2 with SPiCe.

Acknowledgements

We would like to thank our instructors and teaching assistants for their guidance and support throughout this project.

References

[1] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A Simple Framework for Contrastive Learning of Visual Representations. *Proceedings of the 37th International Conference on Machine Learning.*

[2] He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum Contrast for Unsupervised Visual Representation Learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[3] Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., ... & Péré, A. (2020). Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. *Advances in Neural Information Processing Systems.*

[4] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging Properties in Self-Supervised Vision Transformers. *Proceedings of the International Conference on Computer Vision (ICCV).*

[5] Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., ... & Bojanowski, P. (2023). DINOv2: Learning Robust Visual Features without Supervision. *arXiv:2304.07193.*

[6] Niu, C., & Wang, G. (2021). SPICE: Semantic Pseudo-labeling for Image Clustering. *arXiv:2103.09382.*

[7] Zhuang, C., Lin Zhai, A., Yamins, D. (2019). Local Aggregation for Unsupervised Learning of Visual Embeddings. *arXiv:1903.12355.*

[8] Wu, Z., Efros, A., Yu, S. (2018). Improving Generalization via Scalable Neighborhood Component Analysis. *arXiv:1808.04699v1.*

Local Aggregation testing code available here.