Homework 3: Object Recognition and Computer Vision (RecVis24)

Javier Alejandro Lopetegui González
ENS Paris-Saclay
javier.lopetegui_gonzalez@ens-paris-saclay.fr

Abstract

Image classification is challenging, especially with many specific classes and non-natural images, like those in the ImageNet-Sketch dataset. To tackle this, I applied transfer learning using vision transformer models, experimenting with embedding methods, model sizes, and fine-tuning approaches to enhance performance.

1. Introduction

Vision transformers [1] models appeared as a way to adapt the transformer architecture, successfully and extensively applied in NLP, for Computer Vision tasks. By simply applying transfer learning techniques, it is possible to achieve state-of-the-art results in several vision tasks, such as image classification. Therefore, I decided to focus on experimenting with different transfer learning settings using the model Dinov2 [3], a novel and efficient vision transformer model.

2. Solution description

2.1. Baseline

Our initial baseline used the ResNet-50 [2] model, a CNN pre-trained on ImageNet with a residual learning approach. We adapted it to our task by adding a linear layer to its pre-trained weights.

2.2. Dinov2-Based Solution

To enhance the baseline solution I used the Dinov2 [3] model. It uses knowledge distillation to train smaller models initialized with weights from the largest model, enhancing performance. Dinov2 offers four architectures: ViT-S (384,6,12), ViT-B (768,12,18), ViT-L (1024,16,24), and ViT-g (1536,24,40), where triplets denote embedding size, attention heads, and blocks per head.

I focused on the largest three, starting from ViT-B, adding a linear layer for classification. Transfer learning customizations included different embedding strategies (CLS token, average-pooled embedding, or their concatenation), freezing strategies (either all parameters or all but the last attention head), data augmentation (random crops and horizontal flipping controlled via a boolean flag), and configurable dropout applied before the linear layer.

3. Results

After testing various configurations, I observed that adding dropout or data augmentation did not lead to any improvement. The most effective embedding strategy was the concatenation of CLS and average-pooled embeddings, suggesting unique information is captured by both. Table 1 presents the results for the baseline and freezing strategies across model variants. Detailed training metrics are available on `wandb`.[1]

Model	Freezing	val	test
ResNet	”all”	70.28	72.30
ViT-B	”all”	90.00	88.62
ViT-B	”n-1_att”	91.52	90.99
ViT-L	”all”	91.40	92.14
ViT-g	”all”	91.68	**92.89**
ViT-g	”n-1_att”	**91.82**	92.47

Table 1. Results for the baseline and freezing strategies for each Dinov2 model architecture.

4. Discussion and Conclusions

The results confirm that Vision Transformers outperform traditional CNNs in transfer learning scenarios. Larger Dinov2 models achieve better performance, consistent with the original paper’s findings. While fine-tuning the last attention head slightly improved validation results, the improvement was marginal and inconsistent on the test set. Better regularization techniques and additional experiments, including extending the original Dinov2’s pre-training task or leveraging ensemble approaches with other ViT models, could further enhance performance. Computational limitations prevented exhaustive experimentation.

[1] Training report: report

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[3] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 1