Summery chapter

Florian Weidner
Philipps-University Marburg, Germany
Department of Mathematics and Computer Science, Deep Learning Group
February 09, 2024

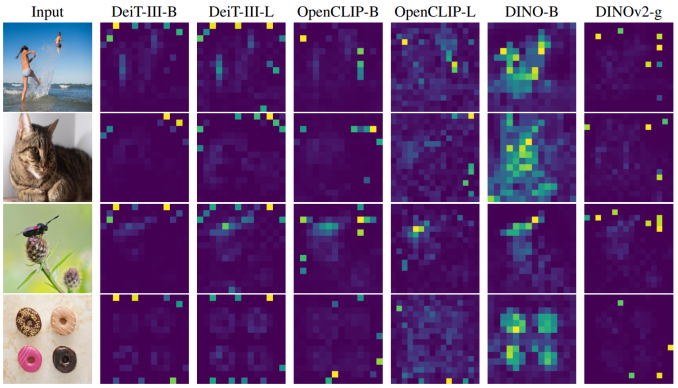


Fig. 1. Illustration of artifacts observed in the attention maps of modern vision transformers. [2]

***Abstract*—The abstract goes here.**

I. Summary

I this chapter we summerize the paper [2]. The paper discoverd artifacts and proposes to use additional register tokens for Vision Transformers (ViTs) to remove these artifacts. After introducing to ViTs like we did in this paper, the models they found the artifacts are introduced. The DINO algorithm is a self-supervised learning method, that uses two ViTs. A student network is predicting the output of a teacher network, to learn rich representaions of visual data without the need of manual annotations. [1] DINO is shown to produce models, that contain semantically consistent information in the last attention layer. Object discovery algorihtms like LOST, built on top of DINO, are using these attention maps, that often contains semantically interpretable information, used to detect objects without supervision. DINOv2 [3] is a improved followup focusing on dense predition tasks, which are tasks, where detailed outputs are required to provide fine-grained localized informations, like semantic segementation or depth estimation. Despite good performance on these dense tasks, the authors observed that DINOv2 is incompatible with LOST [2]. The different behaviour of DINO and DINOv2 can be observed in the artifacts in the last attention maps. In figure 1 you can see the different models and their artifacts on the last attention layer. While DINO shows no peak outlier values focusing the main object in the image, DINOv2 shows a lot of artifacts on the background of the images. This qualitatively observation can be also made for the label-supervised model DeiT-III and the text-supervised model OpenCLIP. Shown in figure 1, you can observe similar artifacts in the background. To explain why and where the artifacts of ViTs in attention maps appear, the paper focuses on DINOv2. Artifact patches show higher norm of their token embedding at the output of the model than other patches. In figure 2 you can see the distribution of the local feature norms over a small dataset. While for DINO, the norm stays under 100 for all patches, DINOv2 shows a lot of patches with a norm higher than 150. This cutoff value can vary across different models. They define artifacts as

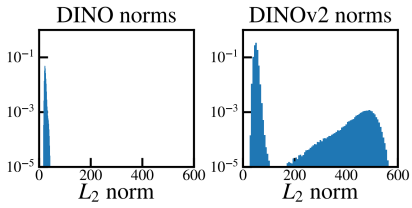


Fig. 2. Comparison of local feature norms for DINO ViT-B/16 and DINOv2 [2]

> ”tokens with norm higher than 150 will be considered as “high-norm” tokens” [2]

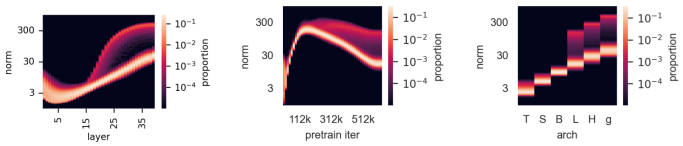
The authors found different conditions, when the artifacts appear in the training process of DINOv2. Figure 3 shows the following conditions:

- artifacts start appearing around layer 15 to 40.
- artifacts start appearing after on thrid of training.
- artifacts only appear in the three largest model versions

Another discovery is that the high-norm tokens appear where patch information is redundant. The authors tested the cosine similarity between high-nomr tokens and their four neighbors, directly after the image is emebdded. They observed, that the high norm patches appear where their cosine similarity to the neighbors is high. Compared to the observations, that shows that artifacts appear mostly in the background of images, high-norm pathes seem to have redundant information, that the model can ignore, to achive similar scores at the output.

References

[1] Mathilde Caron et al. *Emerging Properties in Self-Supervised Vision Transformers*. 2021. arXiv: 2104.14294 [cs.CV]. URL: https://arxiv.org/abs/2104.14294.



(a) Norms along layers. (b) Norms along iterations. (c) Norms across model size.

Fig. 3. Illustration of several properties of outlier tokens in the 40-layer DINOv2 ViT-g model [2]

[2] Timothée Darcet et al. *Vision Transformers Need Registers*. 2024. arXiv: 2309.16588 [cs.CV]. URL: https://arxiv.org/abs/2309.16588.

[3] Maxime Oquab et al. *DINOv2: Learning Robust Visual Features without Supervision*. 2024. arXiv: 2304.07193 [cs.CV]. URL: https://arxiv.org/abs/2304.07193.