On using attention maps and self-supervised learning to improve semantic segmentation results
RECVIS 2021: Project Proposal

Clément Apavou
MVA Master's student
École Normale Supérieure
clement.apavou@ens-paris-saclay.fr

Arthur Zucker
MVA Master's student
École Normale Supérieure
arthur.zucker@ens-paris-saclay.fr

Abstract

Self-supervised learning using transformers [2] has shown interesting emerging properties [1] and learn rich embeddings without annotations. Most recently, Barlow Twins [8] proposed an elegant self-supervised learning technique using a ResNet50 backbone which achieved competitive results when fine-tuned on downstream tasks. The network learns deep image embeddings based on a cross correlation loss which pushes a similarity between the embeddings of two crops of the same image. We propose to study a transformer based Barlow Twins architecture, while analyzing the learned embeddings and testing our method on semantic segmentation tasks. Our goal is to check the reputability of the relevant attention maps produced by [1] using the self supervised training method proposed in Barlow Twins. Then, we want to use such attention maps (either from DINO or our experiment) in order to assist the segmentation and create a new attention aware decoding architecture.

1. Motivation

1.1. Barlow Twins: self-supervised learning

Barlow Twins [8] is a self-supervised method relying on a single neural network trained to produce embeddings from images. Two parts of an image are augmented and fed to the network's backbone while the loss is a cross-correlation between the predicted embeddings. This formulation of self-supervision does not collapse and learns meaningful embeddings, as the fine-tuning on various downstream tasks achieved competitive results. The proposed backbone is a *ResNet50* which does not use attention mechanism.

1.2. Attention based networks

1.2.1 Transformers

Transformers [6] architectures achieved state-of-the-art results ever since their introduction to the field of natural language processing, and recently most recently on computer vision benchmarks with the ViT [2] and Swin Transformer [4]. Given 1, we will use the Swin-T transformer and its variations based on our computing power (4) and the convergence of the model. That is to say that we will switch between the T, S and B variations depending on the capabilities of the GPU. If required, we will use heavy resizing on the image dataset in order to be able to train the network and let it reach convergence. The downstream task will also be evaluated on the potentially resized images in accordance to the training.

1.2.2 Convolution & Attention models

In [5] BoTNet (Bottle neck transfomer Network) is introduced combining convolutions and attention layers using bottleneck transformers instead of ResNet bottlenecks in the classic ResNet [3] architectures. Their experiments show that bottleneck transformer backbones improve results compared to ResNet bottleneck on COCO-stuff semantic segmentation. Using these will allow us to train a simpler architecture but still use attention modules and study the learned attention maps.

2. Proposition

In this project, we propose to analyze and implement Barlow Twins [8] and then use the learned embeddings for a semantic segmentation task on COCO-stuff dataset. We will **first** check that the properties from [1] can be reproduced using **Barlow Twins with transformers architecture**. Thus we will use the pretrained *ResNet50* weights to test our implementation of [8] before modify the bottleneck layers using transformers bottleneck from *BoTNet* [5]. This will allow us to use parts of the pretrained weights from *ResNet50*. Then, we will use a vision transformer (*Swin ViT*) backbone. In order to compare our results, we will visualize the learned embeddings (*e.g* T-SNE) and the attention maps. Our work can also be compared to previous self-supervised tasks which rely on SSL Swin transformers [7].

Second, we want to **study the usage of learned attention heads to create an attention aware decoder for the semantic segmentation** on COCO-stuff. This makes use of the properties described in [1], which show that the attention heads an their maps yield meaningful semantic information. Figure 2 show a representation of the learned attention heads. We could also explore training the attention maps to resemble a class specific semantic segmentation by adding a loss term. If the first step consisting of training Swin-T using Barlow Twins method is too computationally-expensive, we will use the pre-trained attention maps from DINO.

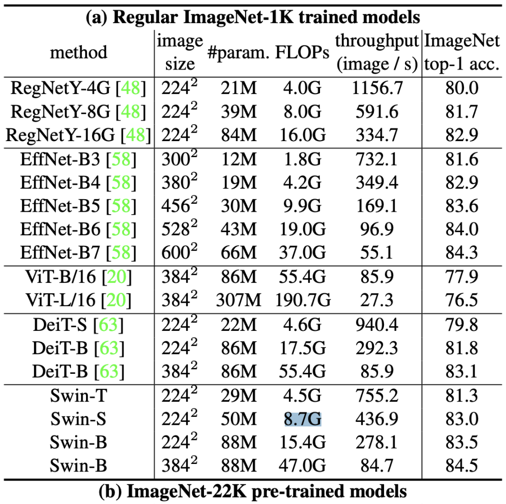
3. Workload

We will implement Barlow twins together, and will then split the different implementation of the backbones.

4. Materials available

Materials available : Nvidia RTX 3080 (10Go) and Nividia RTX 3050 (4Go), Google Colab GPU's (Tesla K80).

5. Figures



(a) Regular ImageNet-1K trained models

method	image size	#param.	FLOPs	throughput (image / s)	ImageNet top-1 acc.
RegNetY-4G [48]	224^2	21M	4.0G	1156.7	80.0
RegNetY-8G [48]	224^2	39M	8.0G	591.6	81.7
RegNetY-16G [48]	224^2	84M	16.0G	334.7	82.9
EffNet-B3 [58]	300^2	12M	1.8G	732.1	81.6
EffNet-B4 [58]	380^2	19M	4.2G	349.4	82.9
EffNet-B5 [58]	456^2	30M	9.9G	169.1	83.6
EffNet-B6 [58]	528^2	43M	19.0G	96.9	84.0
EffNet-B7 [58]	600^2	66M	37.0G	55.1	84.3
ViT-B/16 [20]	384^2	86M	55.4G	85.9	77.9
ViT-L/16 [20]	384^2	307M	190.7G	27.3	76.5
DeiT-S [63]	224^2	22M	4.6G	940.4	79.8
DeiT-B [63]	224^2	86M	17.5G	292.3	81.8
DeiT-B [63]	384^2	86M	55.4G	85.9	83.1
Swin-T	224^2	29M	4.5G	755.2	81.3
Swin-S	224^2	50M	8.7G	436.9	83.0
Swin-B	224^2	88M	15.4G	278.1	83.5
Swin-B	384^2	88M	47.0G	84.7	84.5

(b) ImageNet-22K pre-trained models

Figure 1. Performances of vision transformers and their variations on ImageNet-22K [4]

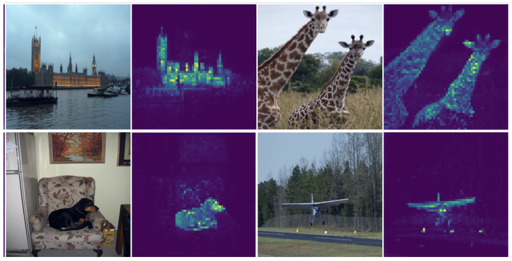


Figure 2. Pixel-wise attention maps from [1]

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. 1, 2

[2] 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

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 1

[4] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution, 2021. 1, 2

[5] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16519–16529, June 2021. 1

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. 1

[7] Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, and Han Hu. Self-supervised learning with swin transformers, 2021. 2

[8] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *CoRR*, abs/2103.03230, 2021. 1