ECOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
MACHINE LEARNING - CS 433

Do CNNs and Vision Transformers learn visual representations similar to those of the monkey brain?

In Collaboration with:
PONCE LABORATORY, HARVARD MEDICAL SCHOOL

Supervisor:
PONCE CARLOS, MD, PHD

Authors:
XU MAOCHENG
MEKKI MALEK
RAKOTOMAHANINA RAPHAËL

DECEMBER 22, 2022

***Abstract*—Simulating visual neurons with Convolutional Neuron Networks have been of a great interest in order to comprehend the neuronal integration. Given an activation maximization pipeline generating so called prototypes (GAN- derived images obtained via neuronal responses), Vision Transformers are now investigated in order to testify their relevance for understanding similar integration processes. Transformers are thus able to develop super-stimulating prototypes. The prototypes converge quicker to maximum local values although these maxima are less significant in comparison with natural images. Eventually, Transformer's prototypes suggest a greater focus on shapes and on textures and have 'messier' generated images than the ones generated by CNNs.**

I. Introduction

The monkey visual cortex and in particular V1, V4 and inferotemporal (IT) neurons are believed to integrate patterns from the environment to translate semantic notions of the individual life such as faces, food or daily objects (paper). The historical use of simplified artificial images revealed preferences for some neurons, such as in the orientation tuning phenomenon [1], but the comprehension of the neuronal integration remains limited as these images are too simple to be associated with real-world representations. Thus, a current challenge is to experiment on synthetic images that are more sophisticated than the previously mentioned figures while being less difficult to manipulate than the highly complex natural images. To do so, a conventional approach consisted in experimenting on simulated neurons, namely Convolutional Neural Network (CNN) and interpret the visual features they learn from artificially optimized images for a designated layer [10]. However, with the increase in use of Vision Transformers (ViT) in image classification tasks [6], the question of their similarity with the monkey neuronal integration arises as well. In the following study, a comparison between a typical CNN, namely AlexNet and several ViTs of the PyTorch library is also operated regarding the visual learnt features they display on their respective Multilayer Perceptron (MLP) component.

II. Previous findings

The notion of tuning landscape has been defined as "the neuronal response function over the entire image manifold" [11]. This space being considered topologically compact, it is expected to dispose maximum values that are translated by excitatory images for a given neuron. Generative adversarial network (GAN)-generated images, commonly called prototype in the literature, have been proved to super-stimulate neurons across the visual system [8] and particularly the ventral stream in two monkeys. Neurons of the internal visual cortex, namely IT neurons, took longer to generate optimized prototypes compared to neurons of the primary visual regions [7]. This observation has been interpreted as the former may integrate more specific information from the environment than the latter. In parallel, convolutional neural network such as AlexNet also revealed that deeper layers took longer to generate their preferred prototypes [7]. In addition, the respective prototypes were simpler and more concentrated, resembling more to little image fragments or patches rather than to natural images. Eventually, mean shift clustering algorithm and discrete cosine transform showed that neurons encode motifs of intermediate complexity and less complex than standard neural networks representations [7], which led to the conclusion that prototypes are efficient to predict the monkeys neuronal behavior.

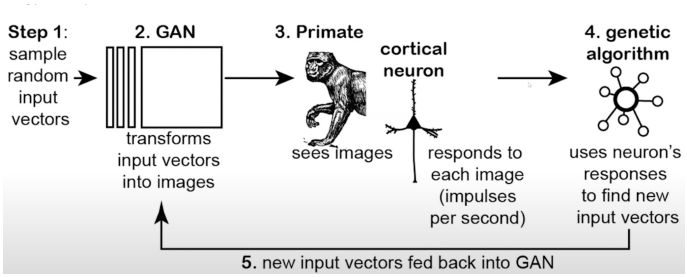

Fig. 1: *Activation maximization pipeline*

III. Methods

This analysis aims to pursue the strategy implemented in Ponce laboratory's previous experiments. More precisely in in-silico experiment, a random "shapeless, texture generated image" [11] is firstly produced by a GAN. Then, one MLP layer that is supposedly simulating a monkey neuron behavior is selected. A stimulating score for the given image and the given layer is then computed from the input vector once it has been processed by the classifier of interest (CNN or ViT). Notice that the classifier is usually pre-trained on an image dataset such as ImageNet or CIFAR 10. At the output of the classifier, a covariance matrix adaptation evolution strategy (CMA-ES) is provided, knowing that the optimization is not a convex problem. This CMA-ES especially allows to maximize the layer response at the next generation by optimizing the input vector of the synthetic image, creating a loop 1. The experiment is eventually terminated once it is suspected that a local maximum score has been reached, which is highly expected due to the compact nature of the image manifold. This overall process is commonly referred as activation maximization [10]. In this study, the CNNs are replaced with common vision transformer (ViTs) algorithms while the layer of interest is the last one of the classifying MLP defined in the following section.

IV. Datasets

The training of ViTs and AlexNet have been accomplished with the Imagenette dataset, a subset of the ImageNet dataset. In other words, this first dataset consists on 10 classes, of a thousand 320pixels colored images each, that are supposely easy to distinguish [4]. The use of this dataset in particular is motivated by a greater speed of training, the set being one hundred times less important than the original ImageNet dataset. In further experiments, training has also been performed on the CIFAR10 dataset, which is composed of 60.000 colored 32x32 images divided into 10 classes [5].

V. Models

ViTs were originally used for natural language processing tasks but became particularly relevant for images classification as well [6]. To complete the latter, the input image is subdivided according to a fixed patch size and flattened while keeping the positional information of the initial input. The subimages are then processed in the encoder defined in 2. The last MLP layer is later in charge of the classification problem and is considered as a simulated visual neuron in this study. The transformers that are taken into account are ImageNet pre-trained models provided by the Pytorch library, namely vit b 16 and vit l 16 [9]. The b and l letters correspond to the complexity of the transformer according to the provided table 3 while the number coincides with the size of patches, 16 being translating as 16x16 pixels for example.

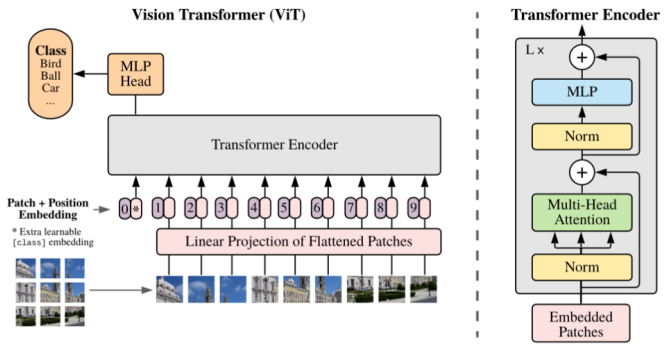

Fig. 2: *Architecture of a Vision Transformer and its application to image classification given a labelled image dataset such as ImageNet. [2]*

Hence, one base model and one large model have been used in this analysis. Furthermore, we implement our own ViT model trained on CIFAR-10 dataset in order to deeply understand the architecture of Vits. The model achieves 76% accuracy when tested on the same dataset. Such implementation have been permitted thanks to the official pytorch documentation [9] and to Alexey Dosovitskiy paper on Vits [2]. Regarding the model hyper-parameters, the patch size affects the input sequence length for the Transformer and can thus impact both computation time and model performance. After testing several patch sizes of 2,4 and 8 (which result in 256, 64 and 8 input sequences respectively), a patch size of 4 was selected as it obtained the best performance. Another aspect to consider when implementing ViTs is that each key has the feature dimensionality of embed dim over num heads [2]. Since we have an input sequence length of 64, a minimum size for the key vector would be 16 or 32. We chose an embedding dimensionality of 256 since a number of heads greater than 8 did not appear to be necessary. Notice that hidden dimensionality has to be larger than embedding dimensionality by at least a factor of 2 (and up to 4). Hence, this dimensionality was set to 512. Finally a dropout value of 0.2 has been set to avoid overfitting.

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Fig. 3: *Description of the different Vision Transformers main architectures [2]*

VI. Results

A. Image Scores Distributions

The activation maximization is not stated to be working when applied to ViTs yet. For that purpose, the scores of 150 randomly sampled images from the Imagenette dataset have been plotted with respect to the classifier MLP layer of interest 4. In parallel, 50 prototypes have been generated following the GAN/CMA-ES pipeline. As a result, the maximum value that has been reached among the 50 prototypes appears to be significantly greater than their corresponding distributions for every classifier. More precisely, an enhance in stimulation for the neuron of interest has been estimated to 168% for the AlexNet algorithm when compared to the mean value while this increase is between 33% and 38% for the tested ViTs.

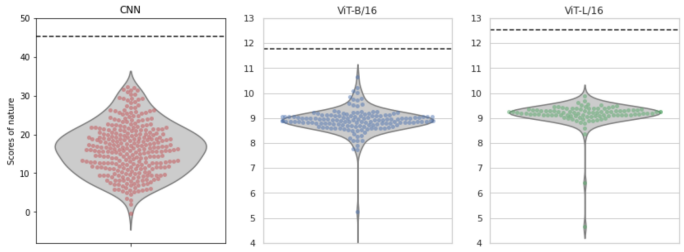

Fig. 4: *Distribution of the scores for 150 images of the Imagenette dataset given the last layer of the MLP component for the implemented ViTs. The dashed lines correspond to the maximum scores reached through the activation maximization pipeline (50 generations).*

B. Activation maximization profile

The evolution of the scores for prototypes during the activation maximization process has been investigated to compare the performance of the different models and to gain insight into the speed of convergence. A min-max scale was used to normalize the scores of every prototypes evolutions with respect to the classifiers. In that manner, the focus is rather based on the convergence than on the absolute values. Notably, the large Vit model architecture (Vit-l-16) performed better than its base counterpart (Vit-b-16) for a fixed patch size. In fact, when displaying the minimum generation for which the score reached half of its maximum value, that is here equals to 1, significant improvements appear. More precisely, while Alexnet reached its half-maximum value at the 41 generation, ViTs performs twice as much better by reaching their corresponding values at the 26 generation for the base model and 14 generations for the large architecture. To further the reasoning, while Alexnet displays quite a linear profile 5, ViTs behave like sigmoids functions so that reaching 80% of their maxima took them less than 40 generations i.e the number of generations needed for AlexNet to reach 50%. To summarize, ViTs prototypes seem to be converging faster by a factor of approximately 2 when compared to a robust CNN.

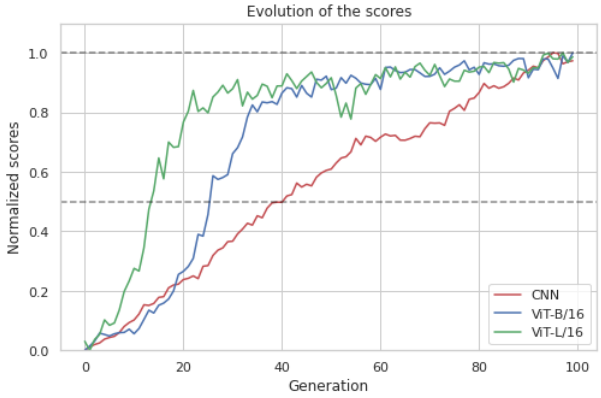

Fig. 5: *Prototype score progression for 100 loops, given AlexNet, ViT B 16 and ViT L 16 classifiers.*

C. External validation to the CIFAR10 dataset

To ensure that the activation maximization is functional with ViTs regardless to the training dataset, another distribution of the scores has been operated as followed in the VI-A section with a sample of the CIFAR10 dataset and with the implemented ViT described in the V section. As expected, the prototype displays a better score than most of the natural images 6.

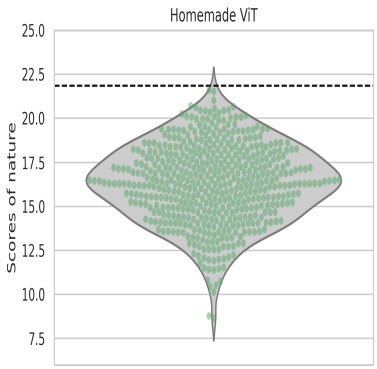

Fig. 6: *Distribution of the scores for 150 images of the CIFAR10 dataset given the last layer of the MLP component for the implemented ViT. The dashed lines correspond to the maximum scores reached through the activation maximization pipeline.*

However, this difference is not as important as viewed with Imagenette, the corresponding prototype score visually not being an outlier. For instance, CIFAR10 is a bigger dataset than Imagenette by a factor of 6. It remains quite uncertain why this behavior occurs but the number of generations one used to create a super-stimulating prototype as well as stochastic properties inherent with the activation maximization protocol might be explanatory variables.

D. Vits generated images interpretations

To qualitatively interpret the content of the images generated by Vits, we propagate the evolutions into a separate neural network (ResNet101) and measure the similarity of the evolutions with real world images, namely the Imagenette dataset. Hence, to find the closest natural images to the Vits-produced prototype of interest using ResNet101, we compare the activation vectors for the fc6 layer as an example. This can be operated by calculating the euclidian distance or even the cosine similarity in this case between the activation vectors. The natural images with the smallest distances to the prototype of interest is considered as the closest in terms of activation for the fc6 layer. To ensure robustness for the qualitative interpretations, two classes of the Imagenette dataset, namely golf ball and tench have been investigated. In that manner, we expect to determine features that are commonly identified by ViTs for both classes. In fact, for the tench class 7, ViTs tend to associate an additional shape to the fish so that it may consider fishermen as part of the tench object. Indeed, when displaying the nearest matches, most of the corresponding pictures monitor a person holding the fish as a trophy.

Fig. 7: *Comparison between a ViT-generated prototype with the Imagenette dataset based on their activation vectors once incorporated in ResNet101. The class of interest that served to generate the prototype is "tench", i.e a fish species. The closest images are the ones with the greater cosine similarity with respect to the prototype.*

Fig. 8: *Comparison between a ViT-generated prototype with the Imagenette dataset based on their activation vectors once incorporated in ResNet101. The class of interest that served to generate the prototype is "golf ball". The closest images are the ones with the greater cosine similarity with respect to the prototype.*

Regarding the golf ball class 8, the round shape appears to remain as the most important feature while the white color does not seem to be favored. To conclude, these two classes allow us to determine that ViTs are particularly focused on the shapes that define a given object while textures and colors are of smaller importance. It is particularly interesting knowing that CNNs may have a texture bias when trained on non-robust ImageNet datasets [3].

VII. Conclusion

Recall that activation maximization involves generating an input image that causes a particular neuron or layer in a neural network to produce a maximal output activation. This procedure helps to understand what features the network has learned to recognize. Actually, activation maximization combined with GANs allows to state that vision transformers (VITs) are also able to generate highly stimulating images. More precisely, the pipeline for convolutional neural networks (CNNs) is sufficiently efficient to create super-stimulating images with ViTs as well. Althought VITs can achieve super-stimuli two times faster than CNNs, there is still further development to make in order to comprehend their relatively low relative values, estimated at 30% increase when compared with their CNNs counterparts reaching 160% for AlexNet. Nevertheless, this greater speed of convergence indicates that when trained on a large dataset, VITs could be able to outperform CNNs. The propagation of activations from VITs into a separate neural network followed by the measure in similarity for activations input vector with respect to natural images exposes a qualitative preference of ViTs towards shapes. A qualitative comparison with the activations from CNNs showed that VITs have messier generated images than CNNs. The latter have shown to remove more background information as displayed for one AlexNet prototype example associated with the "tench" Imagenette dataset 9 (more generated images can be found in the GitHub repository). Thus, ViTs confirm their relevance for image-based machine learning problems but especially in visual neurons simulation. Their decomposition of images into linearly positioned sub-images is a promising architecture as at the output, the algorithm succeeds to focus on shapes rather than textures. Eventually, we encourage further translational studies between monkey brain and ViTs as well as studies on more diverse ViTs architectures. The more complex the ViT is, the more rapid it generates super-stimuli so that training huge-architectures ViTs on even more important dataset such as the original ImageNet dataset could lead to even sharper conclusions.

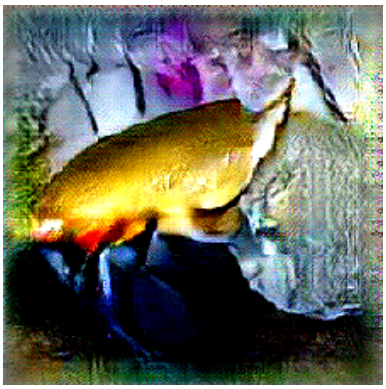

Fig. 9: *Alexnet-generated prototype.*

VIII. Acknowledgement

We would like to thank Dr. Carlos Ponce for hosting this project and allowing us to explore this interesting topic. We are grateful for his excellent guidance throughout the project and for sharing his knowledge and ideas with great enthusiasm.

References

[1] R Ben-Yishai, R L Bar-Or, and H Sompolinsky. "Theory of orientation tuning in visual cortex." In: *Proceedings of the National Academy of Sciences* 92.9 (1995), pp. 3844–3848. DOI: 10.1073/pnas.92.9.3844. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.92.9.3844. URL: https://www.pnas.org/doi/abs/10.1073/pnas.92.9.3844.

[2] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: (2020). DOI: 10.48550/ARXIV.2010.11929. URL: https://arxiv.org/abs/2010.11929.

[3] Robert Geirhos et al. *ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness*. 2018. DOI: 10.48550/ARXIV.1811.12231. URL: https://arxiv.org/abs/1811.12231.

[4] *Imagenette GitHub documentation*. URL: https://github.com/fastai/imagenette.

[5] Alex Krizhevsky. *The CIFAR-10 dataset*. 2009. URL: https://www.cs.toronto.edu/~kriz/cifar.html.

[6] Van-Anh Nguyen et al. "Vision Transformer Visualization: What Neurons Tell and How Neurons Behave?" In: (2022). DOI: 10.48550/ARXIV.2210.07646. URL: https://arxiv.org/abs/2210.07646.

[7] Carlos Ponce. *As simple as possible, but not simpler: features of the neural code for visual recognition*. Dec. 2020. URL: https://cbmm.mit.edu/video/simple-possible-not-simpler-features-neural-code-visual-recognition.

[8] Carlos R. Ponce et al. "Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principles and Neuronal Preferences". In: *Cell* 177.4 (2019), 999–1009.e10. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2019.04.005. URL: https://www.sciencedirect.com/science/article/pii/S0092867419303915.

[9] *VisionTransformer Pytorch documentation*. URL: https://pytorch.org/vision/main/models/vision_transformer.html.

[10] Binxu Wang. *Develop High-Performance Evolutionary Algorithms for Online Neuronal Control*. URL: https://animadversio.github.io/ActMax-Optimizer-Dev/.

[11] Binxu Wang and Carlos R. Ponce. "Tuning landscapes of the ventral stream". In: *Cell Reports* 41.6 (2022), p. 111595. ISSN: 2211-1247. DOI: https://doi.org/10.1016/j.celrep.2022.111595. URL: https://www.sciencedirect.com/science/article/pii/S2211124722014607.