



Self-supervised Rare Visual Token Enhancement

Adrien Chan-Hon-Tong

► To cite this version:

| Adrien Chan-Hon-Tong. Self-supervised Rare Visual Token Enhancement. 2024. hal-04662774

HAL Id: hal-04662774

<https://hal.science/hal-04662774v1>

Preprint submitted on 26 Jul 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

Self-supervised Rare Visual Token Enhancement

Adrien Chan-Hon-Tong

July 26, 2024

Abstract

This paper focuses on rare visual token enhancement: given an image the model is trained to select few visual tokens with the property of being largely distant from all the others. This idea, which is structurally more frugal than classical self-supervised methods, could be an interesting pretext task for both fast classification based on few tokens, or, interesting point selection.

1 Introduction

Self-supervision consists in selecting a label-free pretext task whose resolution eventually forces the model to organize an underlying representation of the datum / data. Two classical pretext tasks are contrastive learning (e.g. [3]) which aims to make all image representations distant from each other while grouping representations of an single image under data augmentations, or, masked prediction (e.g. [14, 2]) where the model should predicts parts of the datum (or related features) that are voluntary removed.

If those two pretext tasks have been successfully used to train very large foundation models (e.g. [16]), they may struggle when trained on smaller models [3]. Yet, both frugality at server level to save energy and frugality at edge level to run on limited hardware require lighter models for many industrial applications. And, despite that pruning [12], distillation [15] and/or quantization [9] of large models have allowed a rapid spread of large-made-light model on smartphone, considering structurally frugal approaches should be investigated to know if better speed / performance balance than compression of large networks can be found.

This paper considers a frugal-by-design label-free task: extracting from a datum only the rare tokens. Obviously, it is not clear that rare tokens capture the global information of the entire datum. In some situation, proportion of different common tokens may encode the main information. Yet, when information is carried by rare token, getting rid of all common-uninformative token may allow a critical speedup in the processing of the datum. This is why this paper focuses on self-supervised rare visual token enhancement: the model is trained to make some visual tokens of an image to be largely distant to all others (assuming the image is not uniform). Importantly, standard contrastive loss

can not operate for such task as a lot of tokens are similar and thus have to share a common representation. So making *all* tokens distant to each other can not be done. This is why the approach has to select *rare* tokens which can be distant from the other (and which should be present only once otherwise their distance to other is 0).

Despite this preliminary work does not answer the question whether those rare tokens are enough for Imagenet classification or useful for point-based geometric registration, the contribution of the paper is to show that separating some token from all others is possible on Imagenet with only few neural layers which is relevant for frugality. Qualitatively, the select token are visually relevant and matches to some extend with classical interest points.

This paper presents related works in next section before losses for the selected pretext task in section 3 and experiments in section 4.

2 Related works

2.1 Sparsity in image classification

The idea of selecting rare areas in image was somehow common before deep learning rising. Indeed, trend in bag-of-features approaches just before Alexnet [10] was about selecting relevant words or features from the codebook [17] i.e. processing the image using only the most relevant features. Since then, image classification is overwhelmingly performed in dense manner using deep network. Despite network efficiency is improved using somehow sparser architectures (e.g. introduction of block removing the need to improve the width of the layers in Resnet [8], grouped convolution and separation between spatial and spectral operation in state of the art ConvNext[13]), deep network tends to process the image entirely.

Recent transformer approaches [5] which perform tokenization of the images could easily reintroduce the idea of token selection. Yet, current trends mostly focuses on training larger networks on larger databases using better self-supervised pretext tasks.

2.2 Sparsity for vision based geometry

The closer works to this paper are related to vision based geometry. Indeed, simultaneous localization and mapping (SLAM [6]), 3D reconstruction[11], or image registration [7] still mostly rely on points: salient points are extracted from multiple images, then, points are matched across images based on their neighborhood to recover corresponding 3D points (1 3D point is associated to 1 2D point in each image).

Thus, this process requires the ability select rare or salient points and the ability to describe them in such way a same 3D points will be described similarly in multiple images i.e. such process should enhance visual rarity. Yet, only few works have considered this task with deep learning approaches. For example, it

worth notices that even recent advanced vision language mapping like [1] relies on hand crafted 2D point extractions to generate 3D point cloud.

The main paper on this idea is [4]. However, [4] relies on a geometric-based pretraining: geometric strong-corner solids are used to generated image on which corresponding 2D points should be extracted. Thus, this approaches is completely relevant for extracting interesting point but it does not directly tackle the task of enhancing rare visual token that may not be point.

Inversely, this paper tries a more general approach which should include interest points but which is more general.

3 Methodology

3.1 Overview

The task considered in this paper is directly to force a small part of the token to be far from all other tokens: formally is $f_1(\theta), \dots, f_N(\theta)$ are N tokens in dimension D extracted from 1 image by a few-layers encoder with weights θ , then, the objective of the training for this image is:

$$\max_{\theta} \left(\frac{\max_{(i_1 \neq \dots \neq i_K) \in \{1, \dots, N\}} \sum_{k \in \{1, \dots, K\}} \min_{n \neq i_k} \|f_{i_k}(\theta) - f_n(\theta)\|^2}{\max_{i, j \in \{1, \dots, N\}} \|f_i(\theta) - f_j(\theta)\|^2} \right) \quad (1)$$

One has to select K features $i_1, \dots, i_K$ with $K \ll N$ (typically $K = \frac{N}{100}$), such that f_{i_k} is distant to all f_n (except himself i.e. $n \neq i_k$). Thus, all $f_{i_k}, f_{i'_k}$ have to be distant from each other and from all other f_n but not-selected n can be close together. This is a very critical difference with classical contrastive learning: similar parts of the image have to lead to similar features. Thus wanting all features to be distant to all other is a non sense, but, wanting that K are distant to all other implies to select K rare tokens only present a single time, and, to make them distant to all others. Let just remark that if θ allows to perform a linear transformation on f , then, eq.1 is equivalent (to a $\frac{1}{2}$ factor) to:

$$\max_{\theta, \forall n, \|f_n(\theta)\| \leq 1} \max_{(i_1 \neq \dots \neq i_K) \in \{1, \dots, N\}} \sum_{k \in \{1, \dots, K\}} \min_{n \neq i_k} \|f_{i_k}(\theta) - f_n(\theta)\|^2 \quad (2)$$

which means: one wants to move some points on the unit disk of dimension D so that some points are each the most possible distant to all others.

3.2 Greedy strategy

This problem is highly not convex as one wants to optimize a distance only between some points that should be selected. Yet, a greedy approach is to consider, for a given image, the currently most distant $i_1, \dots, i_K$, and, to enhance their distance to others. Indeed, for a frozen θ ,

$$\max_{(i_1 \neq \dots \neq i_K) \in \{1, \dots, N\}} \sum_{k \in \{1, \dots, K\}} \min_{n \neq i_k} \|f_{i_k} - f_n\|^2$$

can be trivially solved by:

- computing $\|f_i - f_j\|^2 \forall i \neq j \in \{1, \dots, N\}$
- computing $d_i = \min_{j \neq i} \|f_i - f_j\|^2 \forall i \in \{1, \dots, N\}$
- sorting $\{1, \dots, N\}$ decreasingly according to d_i
- selecting the top- K

Thus, as, stochastic gradient descent will introduce a lot of noise in the selection of the next batch of images and in the optimization process, this greedy approach of enhancing the distance of the current most-far-points may in practice allows to try to enhance the rarity of all possible candidates.

Thus, in practice given $f_1(\theta), \dots, f_N(\theta)$, one will first compute I the set of size K of currently most distant features and J the corresponding closer neighbor, then,

$$\max_{(i_1 \neq \dots \neq i_K) \in \{1, \dots, N\}} \sum_{k \in \{1, \dots, K\}} \min_{n \neq i_k} \|f_{i_k} - f_n\|^2 = \sum \|f_I - f_J\|^2 \quad (3)$$

Importantly, the computation of I, J can be done without retaining the gradient flow.

3.3 Loss

Thus, the most straightforward way is to perform the offered task with stochastic gradient descent is to consider the loss

$$L(x, \theta) = - \sum \|F_{I_x}(x, \theta) - F_{J_x}(x, \theta)\|^2 \quad (4)$$

where x is the current image, F the few-layers encoder and I_x, J_x are computed as pointed in 3.2.

This loss is computed after linear normalization of the features to force them to be in the unit disk - a small auxiliary loss is used so that the normalization becomes native.

Importantly, one target is that rarity will be somehow represented similarly in 2 similar images, thus, robustness to geometric data augmentation D is enforced by a regularity loss:

$$R_D(x, \theta) = \|F(D(x), \theta) - D(F(x, \theta))\|^2 \quad (5)$$

where D can be rotation, symmetry, scaling or crop¹.

The final loss is the sum of the 2 $\mathcal{L}(x, \theta) = R_D(x, \theta) + L(x, \theta)$.

¹Cropping of a relative - and not absolute - part can commute with an encoder even if it reduces the spatial size.



Figure 1: Example of output on Imagenet images. Boxes are the selected tokens, green point are Harris corners extracted with default scikit-image params.

4 Experiments

4.1 Setting

Experiments are performed using the 2 first block of EfficientNet-b5. Those block only reduces spatial size of the image by a factor 2 following by 2 convolutions with kernel 3x3. This ensures the features are mostly local (but with an information at image level brings by squeeze excitation). In order to reduce the influence of a pixel to a single token, a 8x8 pooling is added at the end of those 2 blocs.

This small network (3 trainable full convolutions and some squeeze excitation) is trained with adam optimizer for 2 epoch on batch of size 32 on Imagenet with the loss presented in section 3.

K (the number of rare tokens) is 7 in all the experiments.

4.2 Results

The main result is that both supervised weights or random weights produces a very dense point cloud with almost no native separation between token (distance is respectively 0.095371 and less than 0.001 for Imagenet and random weights).

Inversely, with only two epoch, the network trained with the rarity enhancement loss allows to achieve a distance of 1.144922 in average on Imagenet val between selected tokens and all others.

As the tokens are enforced to belong to the unit circle, it means the absolute optimal distance is 4 while the presence of a point at the center of the circle reduce this optimal distance at 1 (bellow the one found). So this results is not trivial as it has to breaks natural symmetries of the feature distribution to achieve more than 1 in average distance.

Visually, the resulting rare token are humanly acceptable, and, matches to some extent Harris corner.

5 Conclusion

This preliminary works presents a new pretext task that enhances the distance of rare visual tokens to other tokens in images.

Despite many questions are left open (usefulness for image classification or computer geometry based on interesting points, possible alternative losses for a better gradient propagation, ...), this paper proves that this task is optimisable even on middle scale datasets like Imagenet.

References

- [1] Armen Avetisyan, Christopher Xie, Henry Howard-Jenkins, Tsun-Yi Yang, Samir Aroudj, Suvam Patra, Fuyang Zhang, Duncan Frost, Luke Holland, Campbell Orme, et al. Scenescrypt: Reconstructing scenes with an autoregressive structured language model. *arXiv preprint arXiv:2403.13064*, 2024.
- [2] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabat, Yann LeCun, Mido Assran, and Nicolas Ballas. V-jepa: Latent video prediction for visual representation learning. *arXiv*, 2023.
- [3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.
- [4] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018.
- [5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [6] Hugh Durrant-Whyte and Tim Bailey. Simultaneous localization and mapping: part i. *IEEE robotics & automation magazine*, 13(2):99–110, 2006.
- [7] J Michael Fitzpatrick, Derek LG Hill, Calvin R Maurer, et al. Image registration. *Handbook of medical imaging*, 2:447–513, 2000.
- [8] 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.

- [9] Qinghao Hu, Peisong Wang, and Jian Cheng. From hashing to cnns: Training binary weight networks via hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [11] Patrick Labatut, Jean-Philippe Pons, and Renaud Keriven. Efficient multi-view reconstruction of large-scale scenes using interest points, delaunay triangulation and graph cuts. In *2007 IEEE 11th international conference on computer vision*, pages 1–8. IEEE, 2007.
- [12] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. *arXiv preprint arXiv:1608.08710*, 2016.
- [13] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.
- [14] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- [15] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.
- [16] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [17] Lijun Zhang, Chun Chen, Jiajun Bu, Zhengguang Chen, Shulong Tan, and Xiaofei He. Discriminative codeword selection for image representation. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 173–182, 2010.