DINONAT: Exploring Self-Supervised training with Neighbourhood Attention Transformers

Vladimir V. Kniaz[1,2], Vladimir A. Knyaz[1,2], Petr Moshkantsev[1] and Sergey Melnikov[1]

Abstract

Data-driven methods achieved great progress in wide variety of machine vision and data analysis applications due to new possibilities for collecting, annotating and processing huge amounts of data, with supervised learning having the most impressive results. Unfortunately, the extremely time-consuming process of data annotation restricts wide applicability of deep learning in many applications.

Several approaches, such as unsupervised learning or weakly supervised learning has been proposed recently to overcome this problem. Nowadays self-supervised learning demonstrates state-of-the-art performance and outperforms supervised one for many tasks. Another state-of-the-art neural network models are transformer networks, that can rich high performance due to flexibility of the model.

Moreover, the quality of the annotation directly influences the quality of the network operating. From this point of view it is important to analyse what features the network uses during the training process. The study of the self attention mechanism allows to identify these features, and use it in annotation process.

The current study addresses the problem of self-supervised learning of transformer networks as a promise approach for making a step forward in self-adapting of neural network models. Specifically, we study the the cross-modal applicability of self-supervised learning using Transformer network pre-trained on color images for data distilling in thermal images datasets. The results of evaluation demonstrate that Transformer network based on self-attention mechanism identifies the same features both in color and in thermal image datasets.

Keywords

self-supervised learning, neural networks, local attention mechanisms

1. Introduction

Data-driven methods achieved great progress in wide variety of machine vision and data analysis applications due to new possibilities for collecting, annotating and processing huge amounts of data, with supervised learning having the most impressive results. Unfortunately, the extremely time-consuming process of data annotation restricts wide applicability of deep learning in many applications.

This drawback of the supervised learning methods strongly restrains the implementation of deep learning for many applications. Several approaches, such as unsupervised learning [1], semi-supervised learning [2], weakly supervised learning [3] and meta-learning [4] has been

GraphiCon 2023: 33rd International Conference on Computer Graphics and Vision, September 19–21, 2023,
V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, Moscow, Russia
vl.kniaz@gosniias.ru (V. V. Kniaz); knyaz.va@mipt.ru (V. A. Knyaz); petr$_{m}osh@gosniias.ru$ (P. Moshkantsev); melnikovsv@gosniias.ru (S. Melnikov)
https://gosniias.ru (V. V. Kniaz); https://gosniias.ru (S. Melnikov)
0000-0001-7116-9338 (V. V. Kniaz); 0000-0002-4466-244X (V. A. Knyaz); 0000-0001-9624-4322 (P. Moshkantsev); 0000-0002-4466-244X (S. Melnikov)

proposed recently to overcome this problem. Recently, self-supervised learning (SSL) did raise considerable attention in computer vision and achieved significant milestones towards the reduction of human supervision. Indeed, by distilling representative features from unlabelled data, SSL algorithms are already outperforming supervised pre-training on many problems [5].

Self-supervised training of neural networks is a training method based on the principle of using the knowledge that the network already knows about the image. Unlike traditional approaches, where data is fed into the network and its output is compared with the desired result, networks trained in the Self-Supervised mode transform each image into an embedding - a vector in some space that carries numerical information about the image.

The advantage of this training method is the ability to use unlabelled data and allow the neural network to independently choose the most significant areas of the image. In these methods, the visual transformers (ViT), based on the self-attention mechanism, is showing outstanding results. However, a significant drawback of classical implementations of visual transformers is the low speed of obtaining image feature maps due to the quadratic complexity of execution. One of the solutions to this problem is the use of local attention mechanisms. In this method, for each pixel, we obtain a weighted feature map with respect to only its nearest pixels. Through this, it is possible to obtain a weighted feature map in linear time.

Nowadays the Joint Embedding technique demonstrates impressive results in self-supervised learning. We use DINO self-supervised method [10] as a stating point in our study. The DINO framework, simplifies self-supervised training by directly predicting the output of a teacher network – built with a momentum encoder – by using a standard cross-entropy loss.

The main contributions of the study are:

(1) original framework DinoNAT for self-superwised learning based on knowledge distillation without labelling;

(2) evaluation of the performance of DinoNAT framework;

(3) demonstration of cross-modal performance of the DinoNAT framework for vision-thermal imagery.

2. Related work

2.1. Self-supervised learning

A large number of methods are based on different representation of vectors - embeddings of the same augmented image. For example, in [9] the joint information is maximized from different layers of the network. This method allows you to save information about the input image. Also, the distance between the augmented representations of the image is minimized, since the augmentation does not introduce strong distortions into the semantics of the image.

The [12] article presents an approach based on feeding two different augmentations to two networks. In this case, the first network updates its weights using the backpropagation method, and the second - using the exponential moving average mechanism.

Our method is based on [10]. We present the process as a task of knowledge distillation. The student network and the teacher network are involved in the learning process. The teacher network receives only global patches (large chunks of the image), while the student network receives both global and local patches (small chunks of the image). During one epoch, the

student networks weights are updated by backpropagation while the teacher network weights are frozen. At the end of an epoch, the weights of the teacher networks are updated using an exponential moving average mechanism.

2.2. Transformer networks

Transformer networks [6], that were initially proposed for the tasks of natural language processing (NLP), are based on self-attention, thus allowing to exclude recurrent and convolution operation. The transformer networks currently demonstrates the state-of-the-art performance not only for natural language processing problem, but also for computer vision applications.

The applying of Transformer approach to image analysis task resulted in Vision Transformers (ViT) [8]. They demonstrate competitive performance comparing with convolutional networks, but have such drawbacks as high computational and training data needs, not unique features extracted.

Vision Transformer (ViT) [12] was proposed as an image classifier using only a Transformer Encoder operating on an embedded space of image patches, mostly for large-scale training. A number of other methods followed, attempting to increase data efficiency [13, 15, 28], eventually making such Transformer-like models the state of the art in ImageNet-1K classification (without pre-training on large-scale datasets such as JFT-300M).

Previous works, such as DETR [4], explored CNN-Transformer hybrids for object detection. ViT on the other hand proposed a model that would only rely on a single non-overlapping convolutional layer (patching and embedding). ViT was pre-trained primarily on the private JFT-300M dataset, and was shown to outperform state- of-the-art CNNs on many benchmarks. However, it was also added that when ViT is pre-trained on medium-scale datasets, such as ImageNet-1K and ImageNet-21K, it no longer achieves competitive results.

Data-efficient image Transformer (DeiT) model pushed ViT ahead with minimal architectural changes, and through the use of advanced augmentations and training techniques. Their efforts highlighted the true potential of a Transformer-based image classifier in medium-sized data regimes, and inspired many more to adopt their training techniques [21, 29].

2.3. Self Attention

Scaled dot-product attention was defined by Vaswani et al. [31] as an operation on a query and a set of key-value pairs. The dot product of query Q and key K is computed and scaled. Softmax is applied to the output in order to normalize attention weights, and is then applied to the values V. It can be expressed as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}}) \quad (1)$$

where d is embedding dimension. Self attention applies dot-product attention over linear projections of the same input as both the query and key-value pairs. In Transformers, the multi-headed variants of attention and self attention are typically applied. Multi-headed attention applies dot-product attention multiple times over different embeddings, hence forming attention heads.

Stand Alone Self Attention (SASA) [25], is one of the earliest sliding window self attention patterns, aimed to re- place convolutions in existing CNNs. It operates similarly to a convolution with zero padding, and extracts key-value pairs by striding the feature map. The authors reported a noticeable accuracy improvement, but observed that the implementation suffered high latency despite the lower theoretical cost.

SASA and its modifications were not able to scale to larger windows and models as a result of both computational overhead. Additionally, the reduced receptive field in corner cases caused by padding was not addressed. Window and Shifted Window (Swin) Attention [21] were introduced by Liu et al. as non-sliding window-based self attention mechanisms that partition feature maps and apply self attention to each partition separately. This operation has a similar theoretical complexity to SASA, but it can be easily parallelized through batched matrix multiplication. The shifted variant follows the regular, and as the name suggests shifts the partitioning to allow out-of-window interactions, which are necessary for receptive field growth. Their proposed model, Swin Transformer, is one of the earliest hierarchical vision transformers. It produces pyramid-like feature maps, reducing spatial dimensionality while increasing depth.

The proposed Neighbourhood Attention (NA) [7] localizes SA to each pixel's nearest neighbors, which is not necessarily a fixed window around the p ixel. This change in definition allows all pixels to maintain an identical attention span, which would otherwise be reduced for corner pixels in zero-padded alternatives (SASA). NA also approaches SA as its neighbourhood size grows, and is equivalent to SA at maximum neighbourhood. Additionally, NA has the added advantage of maintaining translational equivariance [30], unlike blocked and window self attention.

2.4. Local attention mechanisms

The methods of local attention mechanisms are based on the concept of obtaining a feature map within a certain window, instead of calculating it for the entire image. In article [13], an experiment was carried out to replace convolution operations with local attention mechanisms. As a result, this innovation increased the speed of inference of the model and metrics on the ImageNet sample.

In [7] a new neural network architecture based on the local attention mechanism called the Neighborhood Attention Transformer (NAT) was proposed. The authors managed to obtain a competitive architecture, which showed higher rates of classification speed and accuracy in comparison with the Swin model.

3. Method

This section describes the learning process of the selected neural network in the self-supervised learning mode and its controlled additional training. Our method is based on [10]. We present the process as a task of knowledge distillation.

3.1. Network training

Firstly, two copies of the same model are created to obtain the embedding vectors. The Neighbourhood Attention Transformer architecture was taken as the student and teacher models, the schematic of which is shown in Fig. 1.

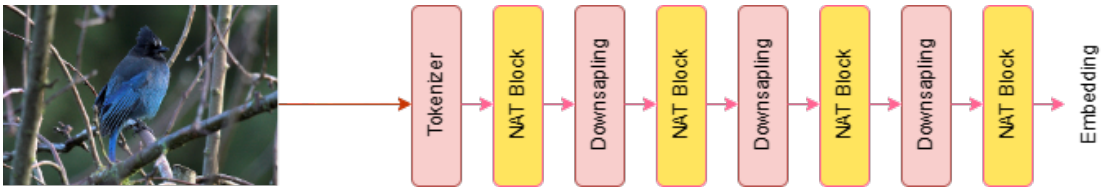

Figure 1: External scheme of the NAT architecture

Further, at each iteration of network training, global and local patches are obtained from the image. Global patches are parts of the image that cover more than 50% of the image (224x224 pixels). Local patches - parts of the image of smaller size (96x96 pixels). They undergo augmentation and are fed to the input of the networks. The input of the teacher network is a global patch, the input of the student network is a global or local patch, but not the same as the input of the teacher network. The output of the networks are embeddings of the same dimensions.

Based on the obtained vectors, we obtain the value of the error function represented by the formula:

$$\min_{\theta_S} \sum_{x \in \{x_1^g, x_2^g\}} \sum_{x' \in V, x' \neq x} H(P_t(x), P_s(x')), \tag{2}$$

where θ – weights,
S – network-student,
x, x' – different patches fed to the input to the networks,
$H(a, b)$ – error function - cross entropy,
$P_a(b)$ – network output vector a.

Based on the obtained value of the error function, the weights of the student network are updated when the weights of the teacher network are frozen.

At the end of each epoch, the weights of the teacher network are updated based on the weights of the student network using a moving average:

$$W_t = aW_t + (1 - a)W_s, \tag{3}$$

where W_t – model-teacher weights,
W_s – model-student weights,
a – some number between 0 and 1.

The scheme of DINO training is shown in Fig. 2.

It can be presented as Algorithm 1.

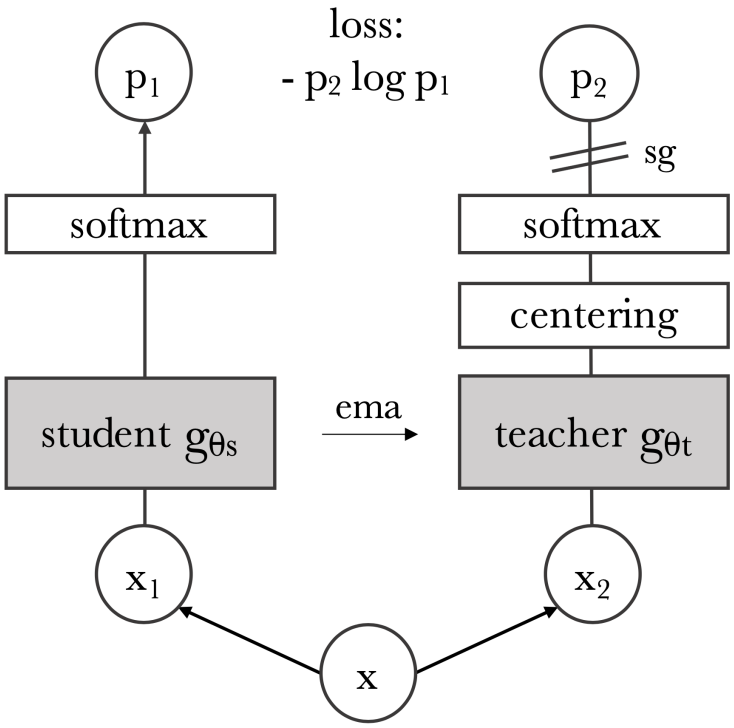

Figure 2: Illustration of DINO work. Different student and teacher augmentations are applied to the input image, after which the input image is passed through neural networks to produce embedding vectors. The teacher's output vector is subjected to sharpening and centring operations to avoid collapse problems. The teacher and student vectors are then passed through Softmax, and the loss function is calculated by updating the student weights. The teacher weights are updated every epoch using the moving average method.

3.2. Classifier retraining

The nearest neighbour method as well as the linear classifier were used as the classifier At each iteration of pre-training, the augmented image was fed to the input of the pre-trained network with frozen weights, the resulting embedding vector was fed to the input of the classifier. The value of the classification error was calculated, and on its basis the classifier weights were updated.

4. CONCLUSION

This work is devoted to the exploring of self-supervised learning of transformers using mechanisms of local attention. Self-Supervised learning is a model learning mode in which the markup is formed based on the internal structure of the objects themselves, or from basic knowledge about objects. The advantage of this method is that there is no need for additional

Algorithm 1: DINO training

```
/* This is comment                                                    */
Input:
Teacher network model G_t with θ_t parameters
Student network model G_s with θ_s parameters
tp_s, tp_t: student and teacher temperatures
C – center (K)
q, m – network and center momentum rates
Output:
θ_s parameters, that match probability P_t(x)

Training procedure ;
Procedure Training():
    /* load a minibatch x with n samples                              */
    for x ∈ Loader do
        x_1 = augment(x)
        x_2 = augment(x)
        s_1 = G_s(x_1),  s_2 = G_s(x_2) ;          /* G_s output      */
        t_1 = G_t(x_1),  t_2 = G_t(x_2) ;          /* G_t output      */
        loss = H(t_1, s_2)/2 + H(t_2, s_1)/2
        loss.backward()
        update(G_s); ;                                  /* SGD        */
        G_t.params = q · G_t.params + (1 − q) · G_s.params
        C = m · C + (1 − m) · cat([t1, t2]).mean(dim = 0)
    return θ_s;

Loss Function ;
Function Loss(t,s):
    t = t.detach() ;                            /* stop gradient      */
    s = softmax(s / tps, dim=1)
    t = softmax((t - C) / tpt, dim=1) ;         /* center + sharpen   */
    return (t ∗ log(s)).sum(dim = 1).mean();
```

markup. As a learning network, we use Neighborhood Attention Transformer, a transformer based on the use of the mechanism of local attention. As part of this work, the Neighborhood Attention Transformer neural network was trained on an ImageNet sample using the method of self-controlled training of DINO neural networks. The results obtained as a result of training were compared with the results presented in the original article of the DINO method, and conclusions were drawn about the quality of the network (Fig. 3).

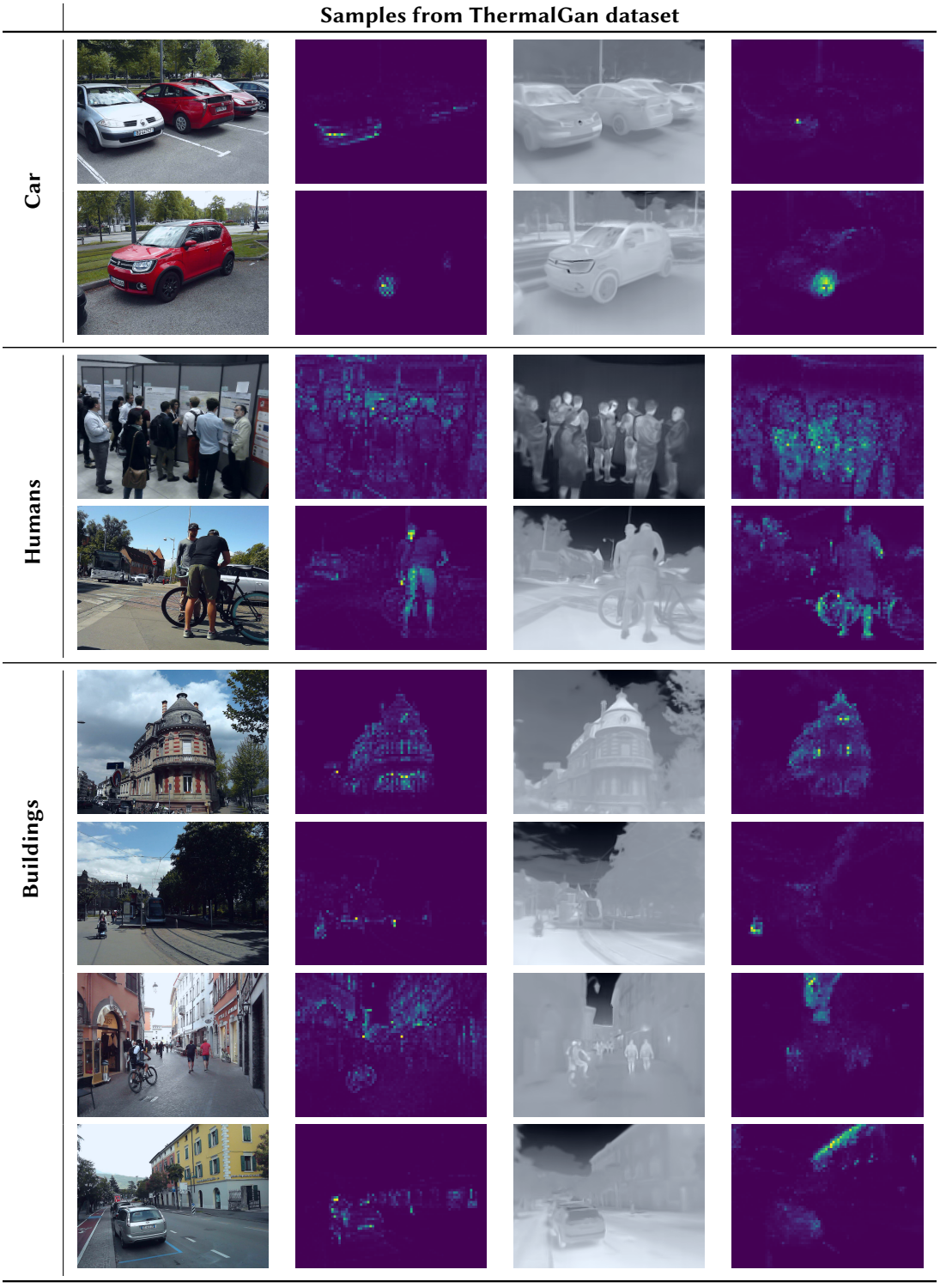

Figure 3: Examples of images and resulting attention maps.

References

[1] Happiness Ugochi Dike et al. “Unsupervised learning based on artificial neural network: A review”. In: 2018 IEEE International Conference on Cyborg and Bionic Systems (CBS). IEEE. 2018, pp. 322–327

[2] Jesper E Van Engelen and Holger H Hoos. “A survey on semi-supervised learning”. In: Machine Learning 109.2 (2020), pp. 373–440

[3] Zhi-Hua Zhou. “A brief introduction to weakly supervised learning”. In: National science review 5.1 (2018), pp. 44–53

[4] Joaquin Vanschoren. “Meta-learning”. In: Automated Machine Learning. Springer, Cham, 2019, pp. 35–61

[5] Priya Goyal and Mathilde Caron and Benjamin Lefaudeux and Min Xu and Pengchao Wang and Vivek Pai and Mannat Singh and Vitaliy Liptchinsky and Ishan Misra and Armand Joulin and Piotr Bojanowski, Self-supervised Pretraining of Visual Features in the Wild, arXiv:2103.01988 [cs.CV], 2021, https://doi.org/10.48550/arXiv.2103.01988

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in Neural Information Processing Systems, 30, 2017

[7] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, Humphrey Shi; Neighborhood Attention Transformer. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6185-6194

[8] 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. preprint arXiv:2010.11929, 2020

[9] Bachman, Philip and Hjelm, R Devon and Buchwalter, William, Learning Representations by Maximizing Mutual Information Across Views

[10] Caron, Mathilde and Touvron, Hugo and Misra, Ishan and Jégou, Hervé and Mairal, Julien and Bojanowski, Piotr and Joulin, Armand, Emerging Properties in Self-Supervised Vision Transformers, arXiv:2104.14294 [cs.CV], 2021, https://doi.org/10.48550/arXiv.2104.14294

[11] Maxime Oquab et al., DINOv2: Learning Robust Visual Features without Supervision, arXiv:2304.07193 [cs.CV], 2023, https://doi.org/10.48550/arXiv.2304.07193

[12] Grill, Jean-Bastien and Strub, Florian and Altché, Florent and Tallec, Corentin and Richemond, Pierre and Buchatskaya, Elena and Doersch, Carl and Avila Pires, Bernardo and Guo, Zhaohan and Gheshlaghi Azar, Mohammad and Piot, Bilal and kavukcuoglu, koray and Munos, Remi and Valko, Michal, Advances in Neural Information Processing Systems

[13] Prajit Ramachandran and Niki Parmar and Ashish Vaswani and Irwan Bello and Anselm Levskaya and Jonathon Shlens, Stand-Alone Self-Attention in Vision Models