Research Investigation for T ransformer for Vision.

(Correlation (relationship) is the core of the core, 相关性才是核心中的核心)

**Summary**

(Correlation is the core of the core,相关性才是核心中的核心). Correlation is even more important than local features.

**Introduction**

1. Transformer for Vision is an excellent algorithm borrowed from the NLP domain for the CV domain. Transformer (Attention is all you need) replaces the traditional RNN structure in the NLP domain with the Attention mechanism, which plays a good role in the language processing of long sequence (long-term dependency). It has the special structure of Encoder and Decoder, breaking the bottleneck of the sequence length of the time with the old RNN and LSTM to excellent levels.[1] In 2020, Facebook's conformer（Transformers with convolutional context for ASR）with the help of CNN to extract short-time sequences. From the perspective of NLP field development, the combination of CNN and Attention has achieved good results.

2. ViT (vision transformer) [2] is a model proposed by Google in 2020 to directly apply transformer in 2020, and much of the subsequent Vision Transformer work is improved based on ViT. ViT idea is simple: directly divide the image into fixed size patchs, and then get patch embedding, through linear transformation this analogy to NLP words and word embedding, and then input patch embedding into Transformer encoder can do the corresponding image classification task.

3.Transformer can be seen [1] and the biggest feeling it gives me is (Correlation is the core of the core). All roads lead to Rome in the paper. The most important thing is not a certain core feature, but the relationship between the whole.

4. Why is the work important? CNN is special Self-Attention,Self-Attention is an outreach of CNN, which is very important.That is, we have more general and advanced means. This is a rigorous mathematical proof (https://arxiv.org/abs/1911.03584) On the Relationship between Self-Attention and Convolutional Layers.local multi-head dot-product self attention blocks can completely replace convolutions..The attention mechanism has a global view and can fully consider the correlations of the different characteristic parts, and it has been shown that the upper limit of the Self-Attention is higher and more futuristic when the training samples are relatively large. At the same time, it has not been relatively long to use transformer research CV in the Transformer for Vision field, the means are relatively primary, and there is still a lot of room for improvement. It let me see how, at least in the field of picture synthesis, how to solve the current relative location of synthetic picture features, with new ideas and new solutions. In the CV domain, both the Transformer for Vision training process and the number of training samples are high and require new means to solve these problems. Finally, Transformer provides a multimodal fusion capability, and CNN is not good at integrating other modal information, such as text, labels, speech, time, and other various related spatial discrete and abstract information, Transformer can.

5. Summarizes recent work in the addendum and I also noticed this latest progress in Bottleneck Transformers for Visual Recognition, replacing bottleneck in the fourth block in ResNet with the MHSA (Multi-Head Self-Attention) module, forming a new module named Bottleneck Transformer (BoT). The network structure eventually composed of block like BoT is called BotNet. Eventually the instance segmentation of BotNet based Mask R-CNN achieved 44.4% Mask AP,49.7% Box AP. on the coco dataset In the classification tasks, 84.7% top-1 accuracy was achieved on ImageNet. And it is 2.33 times faster than EfficientNet.Still using the CNN main structure, the performance is greatly improved, the computation is relatively not very large.

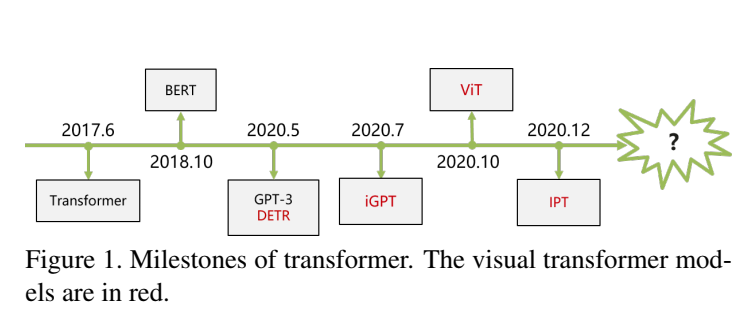
[GitHub:https://github.com/leaderj1001/BottleneckTransformers.](https://github.com/leaderj1001/BottleneckTransformers)

[https://github.com/lucidrains/bottleneck-transformer-pytorch.](https://github.com/lucidrains/bottleneck-transformer-pytorch)

CNN is combined with self-attention:

1. by augmenting feature maps for image classification.
2. by further processing the output of a CNN using self-attention.

**Methodological Approachs**



1. **CNN.**

Without a global vision, one of the core algorithms of deep learning has achieved great achievements.

Reasonable: CNN filters invalid information to characterize valid information. It has a strong feature extraction ability, but the vision is too narrow, like a leopard in a hole.

1. **VIT:AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE.**

In fact, the core problem is to consider how to sequence the image data H\*W\*C, into a word that structure, naturally think of the picture crop into a patch, assuming that there is N patch, dimension for p\*p\*C,reshape plus concate and then become a N\*p ^ 2C, is a similar word vector.

1. **. Image transfer sequence. Position embeddings ③.. learnable embedding ④.. Enter the transformer encoder**

Experimental results:

Training on mid-sized data, the accuracy is a few percentage points lower than the same scale ResNet network;

Training on large scale datasets, migration to smaller scale datasets on the result> = state of the art. (In particular, the best model reached 88.55% on ImageNet, 90.72% on ImageNet-ReaL, 94.55% on CIFAR-100, and 77.63% on the VTAB suite for 19 tasks. ).

Direct project directly: https://github.com/lukemelas/PyTorch-Pretrained-ViT

rationality: Apply NLP's pure transformer to the CV realm.**The different query in ViT is share key set which makes memory access very friendly and greatly accelerated.The presence of ViT changes a lot of intrinsic cognitive 1. locality(locality); 2. translation invariance(translation invariance).**

Problem: Complex, low computational efficiency, and difficult convergence.

**3.Look Closer to See Better:Recurrent Attention Convolutional Neural Network for Fine-grained Image Recognition.**

In this paper, the authors propose a CNN-based attention mechanism, called recurrent attention convolutional neural network (RA-CNN), which recursively analyzes the local information and extracts the necessary features from the local information. Meanwhile, there is a classification structure in subnetworks (sub-network) in RA-CNN, that is, a probability of dividing bird species can be obtained from images of different regions. In addition, the attention mechanism was introduced to make the entire network structure focus not only on the overall information but also on local information, called Attention Proposal Sub-Network (APN). This APN structure is starting from the whole picture (full-image), iteratively generating subregions and making the necessary predictions for these subregions and integrating the predictions obtained from the subregions to obtain the classification prediction probability of the whole picture.

reasonableness: The idea is interesting. the important part gets closer to zoom in. What shortcomings, if there are two and more important goals?

**4 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.**

Original paper:  
[BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.](https://arxiv.org/abs/1810.04805)

Official code and pre-training model  
[Github: https://github.com/google-research/bert.](https://github.com/google-research/bert)

**BERT Benefits**

Transformer Encoder Because of the Self-attention-mechanism, BERT is bidirectional

Because of the two-way functionality and the multi-layer Self-attention mechanism, BERT has to use the Cloze version of the language model Masked-LM to complete the token-level pretraining

To obtain semantic representations of sentence level higher than words, BERT joined Next Sentence Prediction to do joint training with Masked-LM

To adapt migration learning under multi-tasks, BERT designed more versatile input and output layers

Fine-tuning cost is small

**BERT disadvantages**

The random occlusion strategy is slightly rugged, and reading the Data Nosing As Smoothing In Neural Network Language Models》 is recommended

[MASK] tags do not appear in actual prediction and use too much during training [MASK] affects model performance;

Only 15% of the token per batch is predicted, so the BERT converges slower than the left-to-right model (they predict each token)

BERT has a great consumption of hardware resources (larger models take 16 tpu, for four days; larger models take 64 tpu, for four days.

[**5. GPT Model (Generative Pre-Training)**](https://www.cnblogs.com/yifanrensheng/p/13167796.html)

Both are the structure of transformer, while GPT is unidirectional.

GPT used the Decoder structure of Transformer and made some changes to Transformer Decoder, the original Decoder contained two Multi-Head Attention structures, GPT only retained Mask Multi-Head Attention, as shown in the figure below.

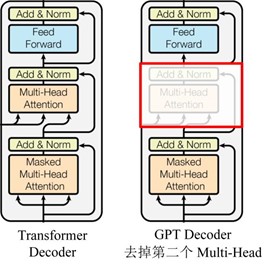


Figure 2 The GPT model

Image GPT.

Key Points of the Image GPT:

Use the same transformer schema as the GPT-2 in NLP

Unsupervised learning, without manual labeling

More calculations are needed to generate competitive representations

The learned features achieve SOTA performance on classification benchmarks for low-resolution datasets

rationality: Image GPT3 breaks the language and visual boundaries



Figure 3 The GPT3 generates a picture in one sentence

https://www.youtube.com/watch?. v= -6Xn4nKm-Qw.

**6.End-to-end object detection with transformers (DERT)**

That is, the N inference results are obtained simultaneously, then calculate the difference between the true values and the prediction two sets, and use the dipartite graph loss to match the prediction and the truth objects. However, due to the slow speed of DERT, there are solutions such as ACT, and even introduce the MTKD method of distillation network. Disadvantages: long training cycle, poor detection effect on small objects.

**7. Pre-Trained Image Processing Transformer (IPT)**

[paper:https://arxiv.org/pdf/2012.00364.pdf.](https://arxiv.org/pdf/2012.00364.pdf)

A pre-training model for handling low-level visual tasks (super-resolution, image demogging, image denoising) is proposed.

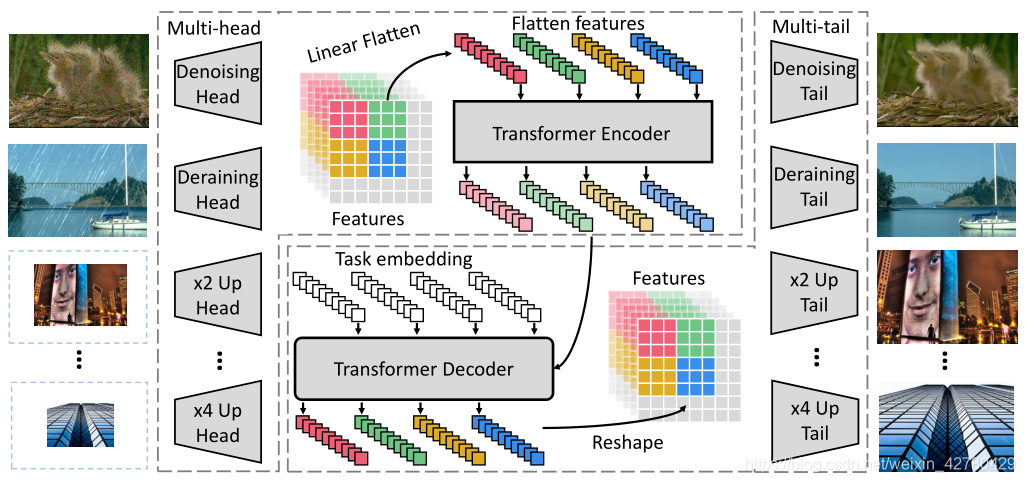


Figure 4 IPT structure diagram

1.Heads

2.Transformer encoder

3.Transformer decoder

4.Tails

**SOTA on super resolution, denoising, rain tasks!** The performance is better than that of IGNN, RDN and rcdnet

**Findings:**

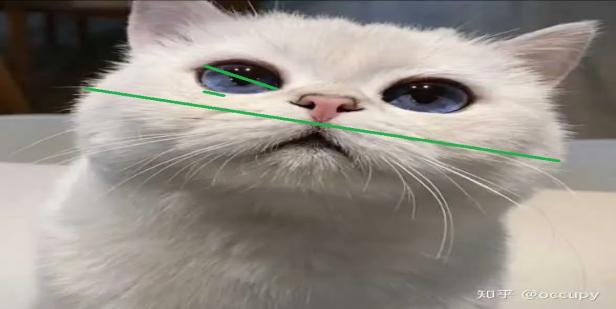


Figure 5 A schematic diagram of a cat's face

Short distances, for example Example of medium distance

When we show a picture of a cat, there is a correlation between the dots that constitute the cat, **otherwise the picture of the dots cannot be called a cat.** There is a strong correlation between the points that constitute the cat's eyes, as does the mouth and nose. Expanding from the stochastic theory, any point in the lattice diagram of the cat has a different size correlation **with itself and any other non-itself point.** The distance between pixels is different, the minimum is adjacent pixels, the largest can span the whole picture, the cat is little distance between the eyes and mouth. The distance between pixels is small and the correlation distance is generally small, but there are also special examples, such as the point distance at the edge of the cat's face is small, but the correlation distance is very large.**Point correlations with short correlation distances are stronger, and the points in cat faces and those in non-cat faces are weak, indicating a longer correlation distance between them. And cats have eyes and have short correlation distances.** Points with strong correlations converge together in higher dimensional spaces via nonlinear representations to form their respective correlation spaces. The volumes, dimensions and characteristics occupied by different related spaces are different, but they each converge and contain each other.

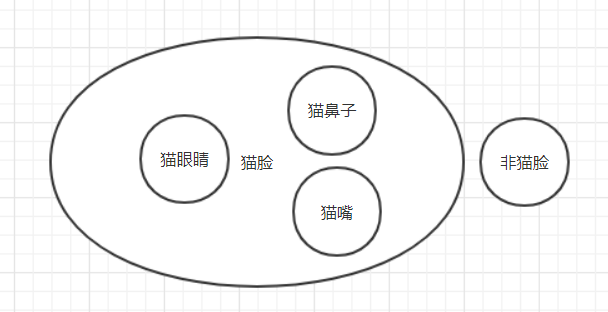


Fig. 6 Correlation convergence space diagram

From the figure below, it is seen that Transformer's self-attention is indeed the correlation between the computational sequences. By learning about the transformer network, we suppress the unimportant non-cat face part and highlight the cat face part. This is actually in dividing the correlation space of convergence, while suppressing part of the correlation space (the non-cat face part also converges to the correlation space different from the cat face). At this point we "attention" to cat face.

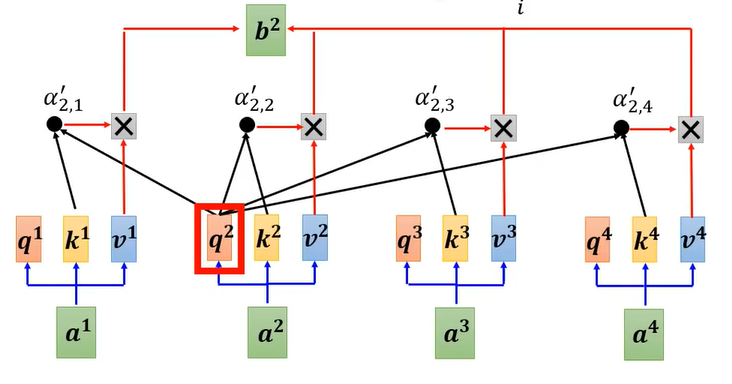


Figure 7 Part transformer calculation method



Figure 8. Original picture



Figure 9. Picture after attention

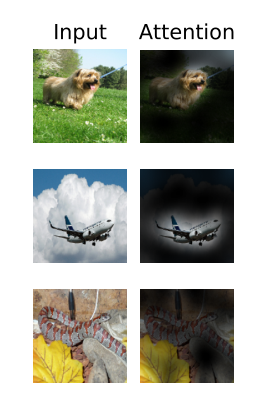


Figure 10 For another example

From the CV point of view, the overall experience of the wild is particularly important. Replace the Attention mechanism for the specific purpose in the CV field is the foreseeable direction in the coming years. There is a theoretical strict proof that CNN is special to Self-Attention, which is very important.One problem of CNN conv operator is that sensory field is limited. In order to expand the attention area of the network, the convolutional-pooling-stacked multi-layer structure, the problem is "effective / real" sensory field is Gaussian at the origin, so the usual effective attention of CNN is one or two more important parts. in the graph

We need to note that the text and speech, especially the text itself has already extracted the information, is relatively abstract, the nonlinear space we need to characterize the speech information is relatively small, the network needed can not be particularly deep. And the picture is just a lattice diagram, no feature extraction, it contains a variety of useless information, all we need is a part of the effective information. In CNN, we use nonlinear properties to construct deep networks to nonlinear characterize the space of pictures, and use convolutional networks to feature extract (filter) pictures to filter useless information. This contains two points, one filtering useless information and a nonlinear characterization of the information. Today some papers propose to use transformer to process images not only large parameters and computations, but also difficult to train to achieve convergence. In fact, there is a nonlinear layer in the structure of T ransformer, which has a certain feature extraction and characterization ability. For the abstract characteristics of speech and text, the characteristics themselves are extracted, so the nonlinear ability of transformer is enough. But it's difficult to converge for the pictures. At the same time, it is intuitively said that the local characteristics that transformer sees are only general, not particularly clear. Local features are not clear, so it is not easy to overfit in training.

What about the relevance of the picture? From the recognition point of view is not so important, because as long as I see some features, such as cat eyes, cat's mouth and nose I can basically judge that this is a cat, and the arrangement of facial features is not important, which is based on the association and completion of partial features. But in the perspective of picture synthesis and further improving the recognition rate, the placement of the cat's eyes, mouth and nose is important, with a correlation between them, only a specific arrangement can consider this as a cat, turning the mouth or nose back, a problem often found in the CNN algorithm. When we can see the relative arrangement of the cat's face (the correlation between each dot), even if I don't see the specific local structure of the cat's eyes, my judgment ability is actually better than only looking at the local characteristics of the eyes. Can we increase the ability of transformor by increasing the number of layers, I think not, because it extracts correlations between dots, which are not transitive, but strong correlation between two points and the other two points. From the correlation perspective, we can even use transformer to see if the two pictures are partially taken from the same picture.

Local features are also important, and CNN network is particularly suitable for this. It is important to extract and reduce the number of parameters and improve the nonlinear ability. Combine CNN and transformer structure together, even using the existing conformer to replace the network structure in Transformer for Vision, you can try it and should be improved. I also firmly believe that this combined approach is bound to make a greater breakthrough in recent years. In particular, using transformer to replace the old way to achieve the global feeling, can certainly replace the traditional deep method to make a breakthrough.

There is a question whether for video to combine CNN and Transformer to achieve better features. After all, transformer is where the Sequence-to-Sequence network is deployed in the time domain, is there a chance to increase the relevance of the core semantics for each frame of the video?

**References:**

[1] (Video lessons by Li Hongyi and Wang Shusen,

https://www.bilibili.com/video/BV1Xp4y1b7ih?. from=search&seid=10483150581334169827,https://www.bilibili.com/video/BV1BQ4y1R7V7? from=search&seid=4328378749261091290)

Or, (Original Paper [Attention Is All You Need]

https://papers.nips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf). Ashish Vaswani \*, Noam Shazeer\*,. Niki Parmar\*,. Jakob Uszkoreit\*,. Llion Jones\*,. Aidan N. Gomez\*,. Łukasz Kaiser\*,. Illia Polosukhin\*. NIPs 2017.)

The Addendum

Classification

[2][AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE](https://arxiv.org/pdf/2010.11929.pdf). Alexey Dosovitskiy∗, Lucas Beyer∗, Alexander Kolesnikov∗, Dirk Weissenborn∗, Xiaohua Zhai∗, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby. Arxiv 2020.

An alternative way to scale attentionis to apply it in Blocks of varying sizes(Weissenbornetal.,2019), in the extreme case only along individual axes(Hoet al., 2019; Wang et al., 2020a).

Detection

[Fast Convergence of DETR with Spatially Modulated Co-Attention](https://arxiv.org/pdf/2101.07448.pdf). Peng Gao, Minghang Zheng, Xiaogang Wang, Jifeng Dai, Hongsheng Li. Arxiv 2021.

[End-to-End Object Detection with Adaptive Clustering Transformer](https://arxiv.org/pdf/2011.09315.pdf). Minghang Zheng, Peng Gao, Xiaogang Wang, Hongsheng Li, Hao Dong. Arxiv 2020.

[Toward Transformer-Based Object Detection](https://arxiv.org/pdf/2012.09958.pdf). Josh Beal\*, Eric Kim\*, Eric Tzeng, Dong Huk Park, Andrew Zhai, Dmitry Kislyuk. Arxiv 2020.

[Rethinking Transformer-based Set Prediction for Object Detection](https://arxiv.org/pdf/2011.10881.pdf). Zhiqing Sun\*, Shengcao Cao\*, Yiming Yang, Kris Kitani. Arxiv 2020.

[UP-DETR: Unsupervised Pre-training for Object Detection with Transformers](https://arxiv.org/pdf/2011.09094.pdf). Zhigang Dai1, Bolun Cai, Yugeng Lin, Junying Chen. Arxiv 2020.

[DEFORMABLE DETR: DEFORMABLE TRANSFORMERS FOR END-TO-END OBJECT DETECTION](https://arxiv.org/pdf/2010.04159.pdf). Xizhou Zhu∗, Weijie Su∗, Lewei Lu, Bin Li, Xiaogang Wang, Jifeng Dai. Arxiv 2020.

[End-to-End Object Detection with Transformers](https://arxiv.org/pdf/2005.12872.pdf). Nicolas Carion\*, Francisco Massa\*, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. ECCV 2020.

**Segmentation**

[Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers](https://arxiv.org/pdf/2012.15840.pdf). Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, Li Zhang. Arxiv 2020.

[End-to-End Video Instance Segmentation with Transformers](https://arxiv.org/pdf/2011.14503.pdf). Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, Huaxia Xia. Arxiv 2020.

**Tracking**

[TransTrack: Multiple-Object Tracking with Transformer](https://arxiv.org/pdf/2012.15460.pdf). Peize Sun, Yi Jiang, Rufeng Zhang, Enze Xie, Jinkun Cao, Xinting Hu, Tao Kong, Zehuan Yuan, Changhu Wang, Ping Luo. Arxiv 2020.

**Image Synthesis**

[Taming Transformers for High-Resolution Image Synthesis](https://arxiv.org/pdf/2012.09841.pdf). Patrick Esser*, Robin Rombach*, Bjorn Ommer. Arxiv 2020.

**Action Understanding**

[Video Action Transformer Network](https://arxiv.org/pdf/1812.02707.pdf). Rohit Girdhar, Joao Carreira, Carl Doersch, Andrew Zisserman. CVPR 2019.

**3D Vision Tasks**

**Point Cloud Processing**

[PCT: Point Cloud Transformer](https://arxiv.org/pdf/2012.09688.pdf). Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R. Martin, Shi-Min Hu. Arxiv 2020.

[Point Transformer](https://arxiv.org/pdf/2012.09164.pdf). Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip Torr, Vladlen Koltun. Arxiv 2020.