Visual Transformers and their applications as discriminators

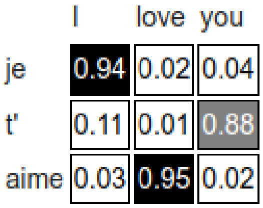
1st Evangelos Papamitsos
National Technical University of Athens
Athens, Greece
vpapamitsos@gmail.com

***Abstract*—One of the most important and revolutionary network architectures proposed in the field of neural networks is that of the Transformer. This network architecture has shown great success in sequence modeling tasks, especially in NLP, almost replacing previous state-of-the-art architectures such as Recurrent neural networks, long short-term memory and gated recurrent neural networks. The Transformer has also been the basis of the architecture of Generative Pre-trained Transformers (GPTs), Bidirectional Encoder Representations from Transformers (BERT) and many more that have become very popular in the community but also to the public. This success has inspired the development of a new variation of the Transformer, the Vision Transformer, aiming to convey the model's impressive results to the field of Computer Vision. In this paper, after we dive deeper in the architectures of the Transformer and Vision Transformer (ViT) and the concept of attention, we aim to present the proposed applications of ViTs specifically in deepfake detection , which has become a very important task, especially after the emergence of highly realistic generative models.**

***Index Terms*—transformers, vision transformers, deepfake detection**

I. Introduction to Transformers and Attention

The Transformer was first introduced in 2017 by the paper [10]. The main motivation behind the new architecture was to try to provide an alternative to sequential nature of existing state-of-the-art Recurrent models, that inherently have computational demands and memory constraints. This was achieved by utilizing the already existing attention mechanism [1] [7] [12], a technique used to model the dependencies between the input and output tokens (in the case of NLP: words), making them more "context-aware" in a global scope and not in a finite window like in the standard Reccurent models.



	I	love	you
je	0.94	0.02	0.04
t'	0.11	0.01	0.88
aime	0.03	0.95	0.02

Fig. 1. Simplified Attention Example

In the figure above we can see a simple visualization attention between input and output in a French-to-English translation task.

The architecture proposed for the model is the following:

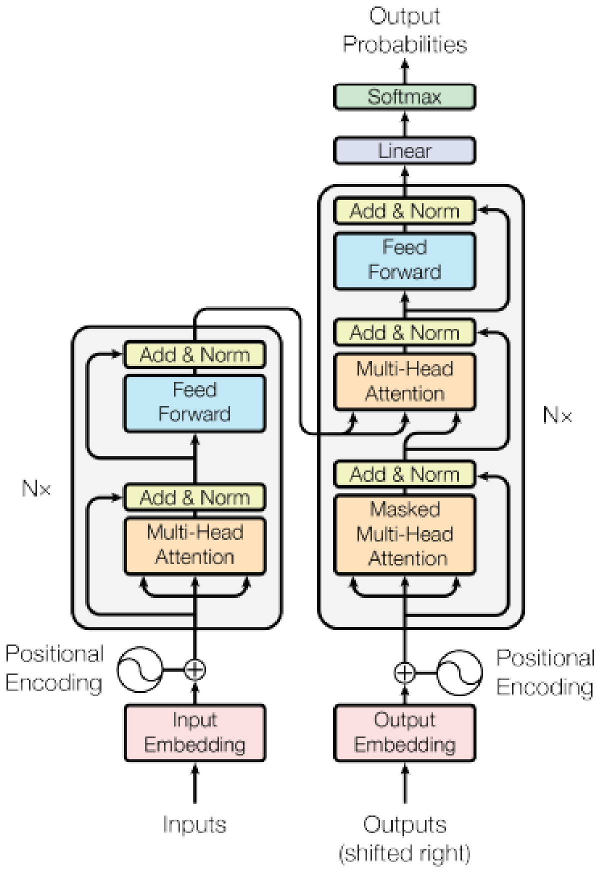


Fig. 2. Transformer Architecture

The model basically consists the encoder and the decoder (6 identical layers stacked). The concept of encoder-decoder architecture is not new and has been used in some form in previous sequential models [7] [12].

To better understand the architecture of the model, we can follow the process of the input passing through the Transformer to generate the output. Firstly, the input vectors (which are mapped in an embedding space) are given positional information by adding positional encoding vectors. In this model these vectors are produced using the following sine and cosine functions:

$$\mathrm{PE}_{(pos,2i)} = \sin\left(\frac{\mathrm{pos}}{10000^{\frac{2i}{dmodel}}}\right)$$

$$\mathrm{PE}_{(pos,2i+1)} = \cos\left(\frac{\mathrm{pos}}{10000^{\frac{2i}{dmodel}}}\right)$$

where pos is the position and i the dimension.

Then these vector, using learned weight matrices, produce the Query, Key, Value vectors that are used by the Multi-Head Attention Layer which is depicted in the following picture:

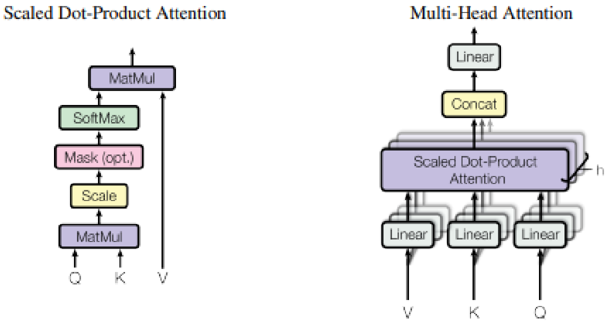
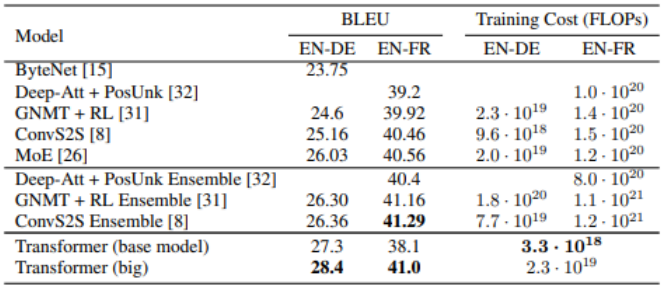


Fig. 3. Attention architecture

The result then is normalized and passed to a Feed Forward Network (also via residual connection) and we get a representation of the input vectors with added attention information. The Decoder operates in a similar way using the previous output embeddings and the Self Attention architecture with the addition of a Mask to avoid calcuating depedencies with subsequent outputs. The result is then used together with the encoder representations (as Key and Value) to a final Multi-Head Attention block similarly to the encoder and finally after the Feed Forward Network, Linear network and Softmax the output probabilities are acquired.



Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [15]	23.75			
Deep-Att + PosUnk [32]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [31]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [8]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [26]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [32]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [31]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [8]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	$\mathbf{3.3 \cdot 10^{18}}$	
Transformer (big)	**28.4**	**41.0**	$2.3 \cdot 10^{19}$	

Fig. 4. Training data and results

The model has 65×10^6 parameters (213×10^6 for the big model) and has outperfomed previous state-of-the-art technologies in the task of translation, while also showing major improvements in training time, capitalizing on fact that input depedencies on a global scale are now being modeled by the parallelizable Self Attention mechanism in contrast to the serial nature of previous models (something that had been part of the motivation behind Transformer, as we have already mentioned). We can see that in the table in Fig. 4.

Stepping upon this architecture, many new technologies have been developed making the Transformer dominate the filed of NLP, with the most notable being:

- BERT [3], developed in 2018 by Google consisting of multiple encoders (12 and 24 for the base-110M parameters and big-340 parameters respectively) and was pretrainded on the Toronto BookCorpus (800M words) and English Wikipedia (2,500M words) and scpecialized in tasks like question answering, masked word prediction etc. showcasing state-of-the-art language understanding.
- GPT models are large language models (LLMs) developed by OpenAI that are also based on the Transformer architecture (we know that the earlier models GPT-1 to GPT-3 consist mainly by Transformer decoders) and have revolutionized the field of natural language understanding and text generation.

It is only natural, therefore, for the Transformers' effectiveness to be researched also in the field of Computer Vision.

II. Introduction to Vision Transformers

The Vision Transformer architecture (ViT) was proposed in 2020 by the paper [5]. The motivation behind it was to try to benefit from the advantages of the Transformers (global Self Attention mechanism, computational efficiency, scalability etc.) and apply them directly to Computer Vision tasks, where CNNs were still the dominant models. The naive approach of using the Transformer architecture as is and treating each pixel as an input token each other pixel can attend through Self attention is obviously computationaly impossible with realistic image sizes. The proposed solution was to split the image in 16×16 patches which are then linearly projected in a fixed length and added with learnable positional embeddings. Then, they are fed to the Transformer encoders and finally with the use of a standard MLP the classification is complete. In the following picture we can see the architecture:

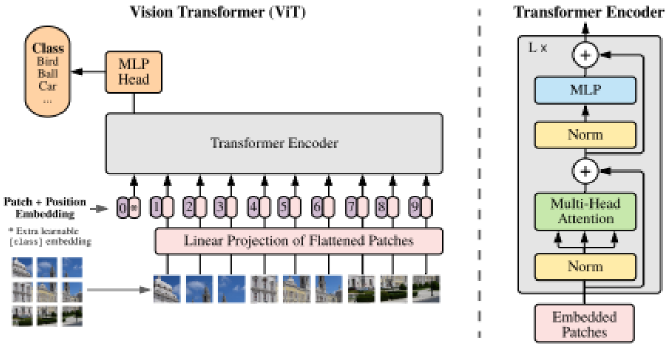
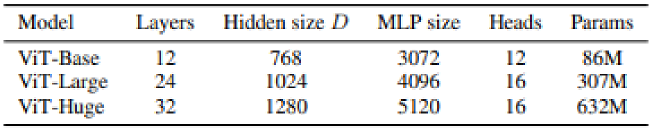


Fig. 5. ViT architecture

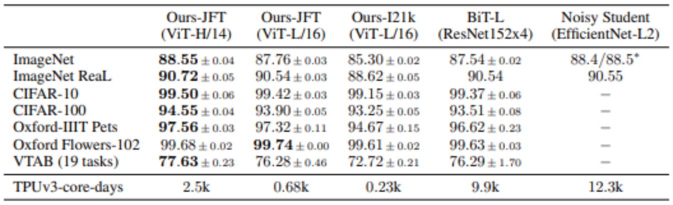
An interesting takeaway is that in the CNN we inherently have an inductive bias due to locality that is nearly absent in the Transformer. This leads to observing mediocre results (with the Transformer) in smaller datasets, as the model does not generalize well, but excellent results in large datasets (14M - 300M images) with fine-tune in smaller/specific ones.



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. 6. Different ViT models

In the following table is presented how ViTs compare with other state-of-the-art models.



	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	–
CIFAR-100	**94.55** ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	–
Oxford-IIIT Pets	**97.56** ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02	**99.74** ± 0.00	99.61 ± 0.02	99.63 ± 0.03	–
VTAB (19 tasks)	**77.63** ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

Fig. 7. Results and comparisons

As we can see, there are improvements across different datasets and tasks, but also major improvement in training time.

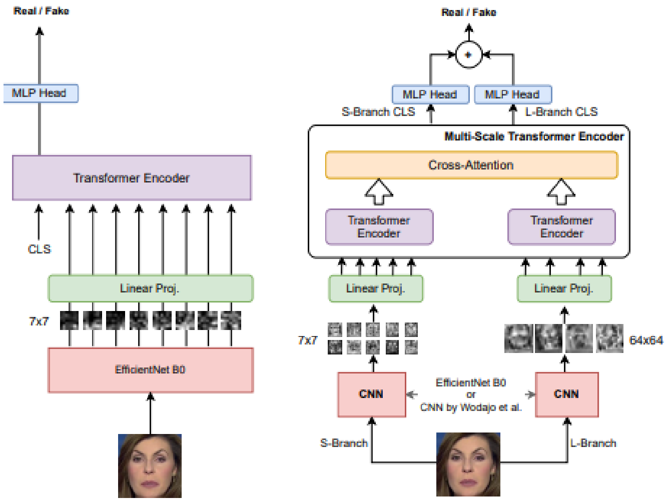
The architecture has inspired follow up research on ViTs and the development of many models showing impressive results, in many cases state-of-the-art, in different areas of computer vision including Image Classification, Segmentation etc.

We aim to discover their current applications as discriminators, their ability, that is, to detect AI generated images and provide further insight.

III. AI GENERATED IMAGE DETECTION, DEEPFAKE, VITS AND MORE

With the raise of generative AI in recent years and technologies that can produce realistic images and manipulated video and while the results can be interesting, impressive or funny, they raise concerns about privacy and trust demands effective ways to discriminate real and AI generated content.

Recognizing that need, the Facebook AI team has developed a dataset [4] containing over 100,000 deepfake videos aiming to discover if a model trained on the Dataset can generalize to real world Deepfake. The paper was supported with a challenge on ***Kaggle***. The results showed that the best models achieved very good perfomance in the dataset's video (test dataset) and decreased, but yet significant, accuracy of around 0.753 on real word deepfake and real videos which indicates that training on the Dataset can indeed train sufficient models. The best model was [8] an ensemble technique that was based on the EfficentNet B7 architecture. EfficientNets [9] are a family of networks designed for efficiency and accuracy by intelligently scaling the model's depth, width, and resolution and have been widely used in Computer Vision. The EfficientNet B0 has also been used paired with a Vision Transformer for deepfake detection [2]. Specifically, in the paper the EfficientNet B0 is used as a feature extractor already pre-extracted faces using the state-of-the-art MTCNN [13]. The visual features are converted to 7 × 7 and then processed by the ViT and finally through an MLP we get the binary classification result. An alternative architecture was also proposed, where two distinct branches are being used using with different size chunks for having a wider receptive field. Similarly, the chunks are processed by the Transformer encoder and combined through cross attention. The architectures are shown below:



(a) Efficient ViT architecture.

(b) Convolutional Cross ViT architecture.

Fig. 8. Architectures

The models are tested on the DFDC test dataset and compared with the [8] but also two other architectures we will discuss later in the paper.

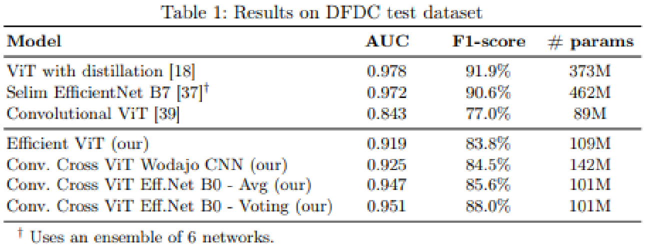


Table 1: Results on DFDC test dataset

Model	AUC	F1-score	# params
ViT with distillation [18]	0.978	91.9%	373M
Selim EfficientNet B7 [37]†	0.972	90.6%	462M
Convolutional ViT [39]	0.843	77.0%	89M
Efficient ViT (our)	0.919	83.8%	109M
Conv. Cross ViT Wodajo CNN (our)	0.925	84.5%	142M
Conv. Cross ViT Eff.Net B0 - Avg (our)	0.947	85.6%	101M
Conv. Cross ViT Eff.Net B0 - Voting (our)	0.951	88.0%	101M

† Uses an ensemble of 6 networks.

Fig. 9. Enter Caption

As we can see, state-of-the-art results in the DFDC dataset were produced by a *ViT with distillation* [6]. The paper proposes a network architecture that is based on Vision Transformer but is trained with the help of a teacher network (thus the distillation) through a distillation token trained by the latter. The teacher network is [8], a rational choice because it has the best results as we saw.

The basic architecture is shown in the following picture:

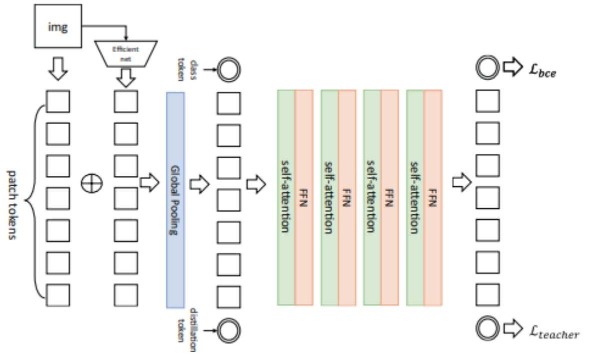


Fig. 10. Vision Transformer and Distillation architecture

The model showed state-of-the-art results on DFDC outperforming [8], predicting more clearly on fake videos, as we can see in the following confusion matrix:

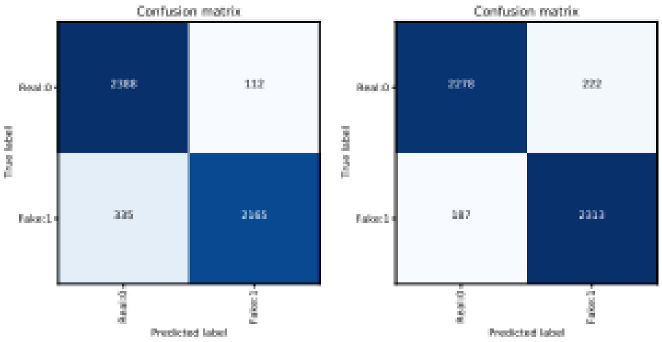


Fig. 11. Confusion matrices, Left-EfficientB7 ensemble / Right-ViT with distillation

Another earlier architecture is the Convolutional ViT [11]. The model was also used as part of the *Convolutional Cross ViT architecture* in [2] for feauture extraction and their results have been compared in *Fig. 9*. Basically, the model is also a combination of two components the preprocessing component and the detection component. The preprocessing component focuses on face extraction and data augmentation while the detection component is based on the actual Convolutional Vision Transformer. The architecture is shown in *Fig. 12*.

IV. Conclusion and Goals

In this paper, we have presented an overview of two very important architectures the Transformer and the Vision Transformer. We also examined how the latter is being used in a very important task in Computer Vision nowdays, that of the deepfake detection with emphasis on videos of people that raise, arguably, the most concern. We aim to further continue our research, after this introduction, by using ViT models for AI generated image detection, explore different datasets and generational models and provide further insight.

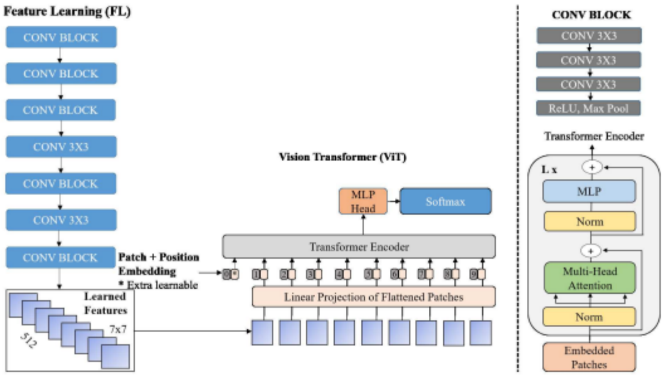


Fig. 12. ConvViT

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[2] Davide Alessandro Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. Combining efficientnet and vision transformers for video deepfake detection. In *International conference on image analysis and processing*, pages 219–229. Springer, 2022.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020.

[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] Young-Jin Heo, Young-Ju Choi, Young-Woon Lee, and Byung-Gyu Kim. Deepfake detection scheme based on vision transformer and distillation. *arXiv preprint arXiv:2104.01353*, 2021.

[7] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[8] Selim Seferbekov. Dfdc 1st place solution (2020), https://github.com/selimsef/dfdc_deepfake_challenge. 2020.

[9] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[11] Deressa Wodajo and Solomon Atnafu. Deepfake video detection using convolutional vision transformer. *arXiv preprint arXiv:2102.11126*, 2021.

[12] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[13] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.