# **Research Papers**

## **Attention Is All You Need**

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the **Transformer**, **based solely on attention mechanisms**, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be **superior in quality** while being **more parallelizable** and requiring **significantly less time to train**. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after **training** for **3.5 days on eight GPUs**, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

### Training

We trained on the **standard WMT 2014 English-German dataset** consisting of about 4.5 million sentence pairs. **Sentences were encoded using ‘byte-pair encoding’**, which has a shared source-target vocabulary of about 37000 tokens. For **English-French**, we used the significantly **larger WMT 2014 English-French dataset** consisting of 36M sentences and split tokens into a 32000 word-piece vocabulary. Sentence pairs were batched together by approximate sequence length. Each training batch contained a set of sentence pairs containing approximately 25000 source tokens and 25000 target tokens.

We trained our models on one machine with **8 NVIDIA P100 GPUs**. For our base models using the hyperparameters described throughout the paper, each **training step took about 0.4 seconds**. We trained the base models for a total of **100,000 steps or 12 hours**. For our **big models, step time was 1.0 seconds**. The big models were **trained for 300,000 steps** (3.5 days).

We used the **Adam optimizer** with β1 = 0:9, β2 = 0:98 and ꜫ = 10^-9. We varied the learning rate over the course of training. This corresponds to increasing the learning rate linearly for the first warmup\_steps training steps and decreasing it thereafter proportionally to the inverse square root of the step number. We used warmup\_steps = 4000.

**We employ three types of regularization during training:**

**Residual Dropout**: We apply dropout to the output of each sub-layer before it is added to the sub-layer input and normalized. In addition, we apply dropout to the sums of the embeddings and the positional encodings in both the encoder and decoder stacks. For the base model, we use a rate of Pdrop = 0.1.

**Label Smoothing**: During training, we employed label smoothing of value ꜫls = 0.1. This hurts perplexity, as the model learns to be more unsure, but improves accuracy and BLEU score.

### Results

On the WMT 2014 English-to-German translation task, the **big transformer model outperforms the best previously reported models (including ensembles)** by more than 2.0 BLEU, **establishing a new state-of-the-art BLEU score of 28.4**. Training took 3.5 days on 8 P100 GPUs. Even our base model surpasses all previously published models and ensembles, at a fraction of the training cost of any of the competitive models.

On the WMT 2014 English-to-French translation task, our **big model achieves a BLEU score of 41.0**, outperforming all of the previously published single models, at **less than 1/4 the training cost of the previous state-of-the-art model**. The Transformer (big) model trained for English-to-French used dropout rate Pdrop = 0.1, instead of 0.3.

For the base models, we used a single model obtained by averaging the last 5 checkpoints, which were written at 10-minute intervals. For the big models, we averaged the last 20 checkpoints. We used beam search with a beam size of 4 and length penalty α = 0.6. These hyperparameters were chosen after experimentation on the development set. We set the maximum output length during inference to input length + 50 but terminate early when possible. We estimate the number of floating point operations used to train a model by multiplying the training time, the number of GPUs used, and an estimate of the sustained single-precision floating-point capacity of each GPU 5.

### Conclusion

In this work, we presented the Transformer, the **first sequence transduction model based entirely on attention**, replacing the recurrent layers most commonly used in encoder-decoder architectures with **multi-headed self-attention**.

For translation tasks, the Transformer can be trained significantly faster than architectures based on recurrent or convolutional layers. On both WMT 2014 English-to-German and WMT 2014 English-to-French translation tasks, we achieve a new state of the art. In the former task our best model outperforms even all previously reported ensembles.

We are excited about the future of attention-based models and plan to apply them to other tasks. We plan to extend the Transformer to problems involving input and output modalities other than text and to investigate local, restricted attention mechanisms to efficiently handle large inputs and outputs such as **images, audio, and video**. Making generation less sequential is another research goals of ours.

## **MViTv2: Improved Multiscale Vision Transformers for Classification and Detection**

In this paper, we study **Multiscale Vision Transformers (MViTv2)** as a unified architecture **for image and video classification**, as well as **object detection**. We present an **improved version of MViT** that incorporates **decomposed relative positional embeddings** and **residual pooling connections**. We instantiate this architecture in five sizes and evaluate it for **ImageNet classification**, **COCO detection** and **Kinetics video recognition** where it outperforms prior work. We further **compare MViTv2’s pooling attention to window attention mechanisms** where it outperforms the latter in accuracy/compute. Without bells-and-whistles, MViTv2 has state-of-the-art performance in 3 domains: **88.8% accuracy on ImageNet classification**, **58.7 APbox on COCO object detection** as well as **86.1% on Kinetics-400 video classification**.

## **xViTCOS: Explainable Vision Transformer Based COVID-19 Screening Using Radiography**

In this paper, they proposed the use of **vision transformers** (instead of convolutional networks (CNN)) for COVID-19 screening using the X-ray and CT images. They employed a **multi-stage transfer learning technique** to address the issue of data scarcity. Furthermore, they showed that the features learned by their transformer networks are **explainable**. Results: They demonstrated that their method not only quantitatively outperforms the recent benchmarks but also focuses on meaningful regions in the images for detection (as confirmed by Radiologists), aiding not only in accurate diagnosis of COVID-19 but also in **localization of the infected area**. The code for their implementation can be found here - <https://github.com/arnabkmondal/xViTCOS>. Conclusion: The proposed method will help in timely identification of COVID-19 and efficient utilization of limited resources.

### Scope and contributions

1) They proposed a vision transformer based deep neural classifier, xViTCOS for screening of COVID-19 from chest radiography.

2) They provided explanability-driven, clinically interpretable visualizations where the patches responsible for the model's prediction are highlighted on the input image.

3) They employed a multi-stage transfer learning approach to address the problem of need for large-scale data.

4) They demonstrated the efficacy of the proposed framework in distinguishing COVID-19 positive cases from non-COVID-19 Pneumonia and Normal control using both chest CT scan and X-ray modality, through several experiments on benchmark datasets.

### Experiments and results

They chose the ViT-B/16 network as the most suitable amongst those tested for further experimentation. The model parameters are initialized with the parameters of a model pretrained on ImageNet-21k and fine-tuned on ImageNet-2012. While training xViTCOS-CXR, for the intermediate finetuning step using CheXpert, they used standard binary cross-entropy loss & While training xViTCOS-CT, they utilized categorical cross-entropy.

They computed and reported Accuracy, Precision (Positive Prediction Value), Recall (Sensitivity), F1 score, Specificity, and Negative Prediction Value (NPV).

For xViTCOS-CT, the proposed method achieves the best accuracy score of 98.1%, surpassing the current state of art methods. xViTCOS-CT achieves a high value of recall or sensitivity at 96%. The proposed model attains high specificity and NPV values of 98.8% for the COVID-19 case.

For xViTCOS-CXR, in terms of classi\_cation accuracy, xViTCOS-CXR achieves an accuracy of 96%, outperforming the baseline methods by a considerable margin. xViTCOS-CXR achieves high recall (100%) and precision values (99%) on the COVID-19 cases. The proposed model attains high specificity and NPV values of almost 100% for the COVID-19.

### Conclusion

They have empirically demonstrated the efficacy of the proposed method over CNN based SOTA methods as measured by various metrics such as precision, recall, F1 score. Additionally, we examine the predictive performance of xViTCOS utilizing explanability-driven heatmap plot to highlight the important factors for the predictive decision it makes. These interpretable visual cues are not only a step towards explainable AI, also might aid practicing radiologists in diagnosis. Furthermore, they also analyzed the failure cases of their proposed method. Thus, to enhance the effectiveness of diagnosis we suggest that xViTCOS be used to complement RT-PCR testing.

## **Top 3 Transformer models 2021**

1. **Google Switch Transformers**

Google’s Switch transformers were the first published work to exceed a trillion parameters, weighing in at 1.6 trillion. The best part is that they devised a mechanism to boost the number of parameters to 1.6 trillion while maintaining the number of Floating point operations per second which is the golden standard for measuring the computation demand of Neural networks.

They have done so using many machine learning concepts. The first one is known as the ‘Mixture Of Experts (MoE)’.

1. **WuDao 2.0 natural language processing model**

WuDao is a Chinese NLP model that has around 1.75 trillion parameters. It was released after Google’s Switch transformers. When it comes to US and China, we all know that it can get competitive sometimes. Wu Dao 2.0 is 10x larger than GPT-3. When it was first released it was the largest neural network ever created and I don’t think there has been a larger model ever since.

Wu Dao 2.0 is “multimodal” which means it can tackle more than one different task such as learning from text and images. It also uses a MoE algorithm (same as Google’s one).

1. **Microsoft ZeRO-Offload: Democratizing Billion-Scale Model Training**

Coming at the 3rd place with a much smaller size, but still quite impressive performance is Microsoft’s Zero-Offload. The model is mainly focused on minimizing the offload costs from/to GPUs/CPUs through novel optimization tricks. It is capable of training models with up to 70 billion parameters.

One of the unique ideas about this transformer is the way they approached the solution. They decided to model this communication problem using data-flow graphs and first principle analysis. And the bulk of the work is done to effectively partition the graphs between the CPU and GPU devices.

# **3-Research papers based on Transformers applications:**

## **CvT: Introducing Convolutions to Vision Transformers**

We present in this paper a new architecture, named **Convolutional vision Transformer (CvT)**, that improves Vision Transformer (ViT) in performance and efficiency by introducing **convolutions into ViT** to yield the best of both designs. This is accomplished through two primary modifications: a hierarchy of Transformers containing a new **convolutional token embedding**, and a **convolutional Transformer block** leveraging a **convolutional projection**.

These changes introduce desirable properties of convolutional neural networks (CNNs) to the ViT architecture (i.e., shift, scale, and distortion invariance) while **maintaining the merits of Transformers** (i.e., dynamic attention, global context, and better generalization). Introducing convolutions into the model, **allowing local context to be captured**.

Finally, our results show that the **positional encoding**, a crucial component in existing Vision Transformers, can be **safely removed** in our model, simplifying the design for higher resolution vision tasks.

ViT lacks certain desirable properties inherently built into the CNN architecture that make CNNs uniquely suited to solve vision tasks. For example, images have a strong 2D local structure: spatially neighboring pixels are usually highly correlated. The CNN architecture forces the capture of this local structure by using local receptive fields, shared weights, and spatial subsampling, and thus also achieves some degree of shift, scale, and distortion invariance. In addition, the hierarchical structure of convolutional kernels learns visual patterns that take into account local spatial context at varying levels of complexity, from simple low-level edges and textures to higher order semantic patterns.

In summary, our proposed Convolutional vision Transformer (CvT) employs all the benefits of CNNs: local receptive fields, shared weights, and spatial subsampling, while keeping all the advantages of Transformers: dynamic attention, global context fusion, and better generalization.

We introduce two convolution-based operations into the Vision Transformer architecture: the **Convolutional Token Embedding** and **Convolutional Projection**.

**Convolutional Token Embedding** allows each stage to progressively reduce the number of tokens (i.e., feature resolution) while simultaneously increasing the width of the tokens (i.e., feature dimension), thus achieving spatial downsampling and increased richness of representation, similar to the design of CNNs.

**Convolutional Projection** is applied for query, key, and value embeddings respectively, instead of the standard position-wise linear projection in ViT. Additionally, the classification token is added only in the last stage. Finally, an MLP (i.e., fully connected) Head is utilized upon the classification token of the final stage output to predict the class.

Introducing convolutions into the Vision Transformer architecture to merge the benefits of Transformers with the benefits of CNNs for image recognition tasks. Extensive experiments demonstrate that the introduced convolutional token embedding and convolutional projection, along with the multi-stage design of the network enabled by convolutions, make our CvT architecture achieve superior performance while maintaining computational efficiency. Furthermore, due to the built-in local context structure introduced by convolutions, CvT no longer requires a position embedding, giving it a potential advantage for adaption to a wide range of vision tasks requiring variable input resolution.

## **A ConvNet for the 2020s**

Vanilla ViT faces difficulties when applied to general computer vision tasks such as object detection and semantic segmentation. It is the hierarchical Transformers (e.g., Swin Transformers) that reintroduced several ConvNet priors, making Transformers practically viable as a generic vision backbone and demonstrating remarkable performance on a wide variety of vision tasks.

In this work, we re-examine the design spaces and test the limits of what a pure ConvNet can achieve. We gradually “modernize” a standard ResNet toward the design of a vision Transformer and discover several key components that contribute to the performance difference along the way.

ConvNeXts compete favourably with Transformers in terms of accuracy and scalability, achieving 87.8% ImageNet top-1 accuracy and outperforming Swin Transformers on COCO detection and ADE20K segmentation, while maintaining the simplicity and efficiency of standard ConvNets.

ConvNets in computer vision was not a coincidence: in many application scenarios, a “**sliding window**” strategy is intrinsic to visual processing, particularly when **working with high-resolution images**. ConvNets have **several built-in inductive biases** that make them **well-suited to a wide variety of computer vision applications**.

The most important one is **translation equivariance**, which is a desirable property for tasks like **objection detection**. ConvNets are also inherently efficient due to the fact that when used in a sliding-window manner, the computations are shared.

ViT introduces **no image-specific inductive bias** and makes minimal changes to the original NLP Transformers. One primary focus of ViT is on the **scaling behaviour**: with the **help of larger model** and **dataset sizes**, transformers can **outperform standard ResNets** by a significant margin.

Without the **ConvNet inductive biases**, a vanilla ViT model faces many challenges in being adopted as a generic vision backbone. The biggest challenge is **ViT’s global attention design**, which has a quadratic complexity with respect to the input size. This might be acceptable for ImageNet classification, but quickly becomes **intractable with higher-resolution inputs**. Hierarchical Transformers employ a hybrid approach to bridge this gap. For example, the “sliding window” strategy (e.g., attention within local windows) was reintroduced to Transformers, allowing them to behave more similarly to ConvNets. **Swin Transformer’s** success and rapid adoption also revealed one thing: the essence of convolution is not becoming irrelevant; rather, it remains much desired and has never faded.

Transformers for computer vision have been aimed at bringing back convolutions. These attempts, however, come at a cost: a naive implementation of sliding window self-attention can be expensive; with advanced approaches such as cyclic shifting, the speed can be optimized but the system becomes more sophisticated in design.

The only reason ConvNets appear to be losing steam is that (**hierarchical**) **Transformers surpass them in many vision** **tasks**, and the performance difference is usually attributed to the **superior scaling behaviour of Transformers**, with **multi-head self-attention** being the **key component**.

ConvNets and hierarchical vision Transformers become different and similar at the same time: they are both equipped with **similar inductive biases** but **differ significantly in the training procedure and macro/micro-level architecture design**. In this work, we investigate the architectural distinctions between ConvNets and Transformers and try to identify the confounding variables when comparing the network performance. Our research is intended to bridge the gap between the pre-ViT and post-ViT eras for ConvNets, as well as to test the limits of what a pure ConvNet can achieve.