# **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**.