Attention Is All You Need!

Prepared by : Vaishnav Krishna P(F14158802)

Year: 2017
Publication / Venue: NeurIPS 2017

Author

Ashish Vaswani
Noam Shazeer
Niki Parmar
Jakob Uszkoreit
Llion Jones
Aidan N. Gomez
Łukasz Kaiser
Illia Polosukhin

Abstract

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 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

1. **Context(background / what we have)**

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.

2. **Task - introducing new architecture**

We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely.

3. **Objective - How Transformers are superior than other**

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.

4. **Finding - performance accuracy & efficiency**

- **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 41.0 after training for 3.5 days on eight GPUs**, a small fraction of the training costs of the best models from the literature.

5. **Conclusion (not stated directly)**

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

The Transformer achieves **state-of-the-art performance** on both **WMT 2014 English-to-German** and **English-to-French translation tasks**, outperforming previous models, including ensembles, while being **significantly faster to train** than recurrent or convolution-based architectures.

6. **Perspective / Future outlook(not stated directly)**

The Transformer opens a new direction for **attention-based models** across multiple domains. Future work includes:

- Applying Transformers to **tasks beyond text**, such as **images, audio, and video**.
- Investigating **local or restricted attention mechanisms** to efficiently handle **large inputs and outputs**.