Transformers for Point-cloud Data
Week 4 Progress

Project lead: Aakash Kumar
By: Jordaine Wisdom
Main Tasks

- Understanding transformers
- Image transformers
- Point cloud transformers
Attention Is All You Need

Ashish Vaswani*  
Google Brain  
avasani@google.com

Noam Shazeer*  
Google Brain  
nno@google.com

Niki Parmar*  
Google Research  
niki@google.com

Jakub Lukasiak*  
Google Research  
uus@google.com

Llion Jones*  
Google Research  
ljones@google.com

Aidan N. Gomez  
University of Toronto  
asdmohan.utoronto.com

Lukasz Kaiser*  
Google Brain  
lukasz.kaiser@gmail.com

Iliia Polosukhin*  
iliia.polosukhin@gmail.com

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 self-attention mechanisms, eliminating the recurrence and convolutional filters 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 44.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

Introduction

Recurrent neural networks, long short-term memory [12] and gated recurrent [7] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and transduction problems such as language modeling and machine translation [29, 2, 5]. Numerous efforts have since continued to push the boundaries of recurrent language models and encoder-decoder architectures [31, 24, 13].

*Equal contribution. Listing order is random. Jakub proposal replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Boris, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and wrote the paper. Iliia also experimented with novel model variants, was responsible for our initial deployment, and efficiently implemented and visualized. Lukasz and Aidan spent countless long days running various parts of and implementing these/other ideas, and troubleshooting various bugs and issues. Lilian also experimented with novel model variants, was responsible for our initial deployment, and successfully accelerating training and validation on GPUs.
Transformer architecture analysis
Analyzing image transformer (DETR)
Analyzing point cloud transformer (3D-DETR)

**Figure 2: Approach.** (Left) 3D-DETR is an end-to-end trainable Transformer that takes a set of 3D points (point cloud) as input and outputs a set of 3D bounding boxes. The Transformer encoder produces a set of per-point features using multiple layers of self-attention. The point features and a set of ‘query’ embeddings are input to the Transformer decoder that produces a set of boxes. We match the predicted boxes to the ground truth and optimize a set loss. Our model does not use color information (used for visualization only). (Right) We randomly sample a set of ‘query’ points that are embedded and then converted into bounding box predictions by the decoder.
Next

- Training and testing 3DETR on ScanNet Dataset
- Gain intuitive understanding of 3DETR code