Attention is All You Need

introduction to the transformer architecture

Felix Karg

March 11, 2023

Institute for Theoretical Informatics:
Artificial Intelligence for Materials Science

Plan

Individual Parts:

Plan

Individual Parts:

- Normal FeedForward MLP

Plan

Individual Parts:

- Normal FeedForward MLP
- Embedding: Input

Plan

Individual Parts:

- Normal FeedForward MLP
- Embedding: Input
- Embedding: Location

Plan

Individual Parts:

- Normal FeedForward MLP
- Embedding: Input
- Embedding: Location
- Basics of Attention (before transformer)

Plan

Individual Parts:

- Normal FeedForward MLP
- Embedding: Input
- Embedding: Location
- Basics of Attention (before transformer)
- Attention is All You Need [1]

Plan

Individual Parts:

- Normal FeedForward MLP
- Embedding: Input
- Embedding: Location
- Basics of Attention (before transformer)
- Attention is All You Need [1]
- FastFormer [2]

Plan

Individual Parts:

- Normal FeedForward MLP
- Embedding: Input
- Embedding: Location
- Basics of Attention (before transformer)
- Attention is All You Need [1]
- FastFormer [2]
- (fun:) One Model to Learn Them All [3]

Plan

Individual Parts:

- Normal FeedForward MLP
- Embedding: Input
- Embedding: Location
- Basics of Attention (before transformer)
- Attention is All You Need [1]
- FastFormer [2]
- (fun:) One Model to Learn Them All [3]
- Distillation / Quantization [4]

Overview

Background

Embedding

Attention

Transformer

Compression

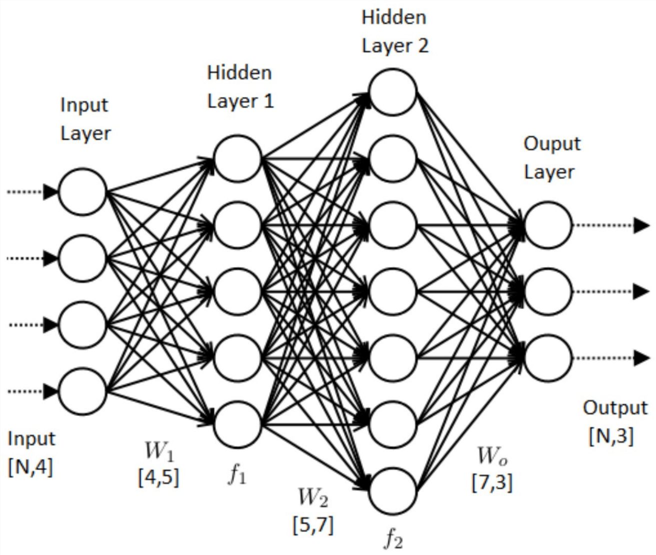
Successes

Conclusion

Background

- **Multi-Layer Perceptron**
- Dropout
- Residual Connections
- RegNet

Multi-Layer Perceptron



Source: Public Domain

Background

Multi-Layer Perceptron

Dropout

Residual Connections

RegNet

Powerful method for regularization during training Dropout [5]

Background

Skipping connections to propagate gradients Residual Connections [6]

Background

Multi-Layer Perceptron

Dropout

Residual Connections

RegNet

Self-Regulated Network [7]

Embedding

Input

Location

- Build Vocabulary

- Build Vocabulary
- Define Embedding Dimensions

- Build Vocabulary
- Define Embedding Dimensions
- Learn Mapping

Embedding

Input

Location

Basically sinosoidal grid-cell patterns
Simply getting added on top

Attention

Attention before Transformers

Something about Q, K, V matrices and effects

Yes we have empty context

Attention

Multi-Head-Attention

Attention

Basic Attention

Multi-Head Attention

Masked Attention

Masking Attention

Transformer

FF and Softmax over embedding
followed by topk selection

Transformer

Full Architecture Overview

Transformer

Even the original GPT didn't use 'full' transformers, but Sparse Transformer [8]

Transformer

Recently, people built a linear-cost attention mechanism: FastFormer [2]

Compression

Quantization

Distillation

Reducing resolution of models after training significantly, e.g. from fp32 to fp8

Compression

Quantization

Distillation

Training a smaller model on outputting similar outputs distributions for given inputs

Successes

BERT

GPT2

CLIP

Diffusion

generating language embeddings for NLP tasks

Successes

- BERT
- GPT2
- CLIP
- Diffusion

because architecture parameters and dimensions are known

Successes

- BERT
- GPT2
- **CLIP**
- Diffusion

Building semantic embeddings shared from pictures and text

Successes

- BERT
- GPT2
- CLIP
- **Diffusion**

Generating images based on shared semantic embeddings

Conclusion

We've seen:

Conclusion

We've seen:

- things

Sources i

- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, t. L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- C. Wu, F. Wu, T. Qi, Y. Huang, and X. Xie, "Fastformer: Additive Attention Can Be All You Need," Sept. 2021.
- L. Kaiser, A. N. Gomez, N. Shazeer, A. Vaswani, N. Parmar, L. Jones, and J. Uszkoreit, "One Model To Learn Them All," June 2017.

Sources ii

- A. Polino, R. Pascanu, and D. Alistarh, "Model compression via distillation and quantization," Feb. 2018.
- N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Sources iii

- J. Xu, Y. Pan, X. Pan, S. Hoi, Z. Yi, and Z. Xu, “RegNet: Self-Regulated Network for Image Classification,” Jan. 2021.
- R. Child, S. Gray, A. Radford, and I. Sutskever, “Generating Long Sequences with Sparse Transformers,” Apr. 2019.

End

Additional slide

without numbering, does not show up in normal numbers