Deep learning

13.3. Transformer Networks

François Fleuret
https://fleuret.org/dlc/
Vaswani et al. (2017) proposed to go one step further: instead of using attention mechanisms as a supplement to standard convolutional and recurrent operations, they designed a models combining only attention layers.

They designed this “transformer” for a sequence-to-sequence translation task, but it is currently key to state-of-the-art approaches across NLP tasks.
They first introduce a multi-head attention module.

(Vaswani et al., 2017)

\[
\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{Q K^\top}{\sqrt{d_k}} \right) V
\]

\[
\text{MultiHead}(Q, K, V) = \text{Concat} \left( H_1, \ldots, H_h \right) W^O
\]

\[
H_i = \text{Attention} \left( Q W_i^Q, K W_i^K, V W_i^V \right), \quad i = 1, \ldots, h
\]

with

\[
W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}, \quad W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}, \quad W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}, \quad W^O \in \mathbb{R}^{h d_v \times d_{\text{model}}}
\]
Their complete model is composed of:

- An encoder that combines $N = 6$ modules each composed of a multi-head attention sub-module, and a [per-component] one hidden-layer MLP, with residual pass-through and layer normalization.
- A decoder with a similar structure, but with causal attention layers to allow for regression training, and additional attention layers that attend to the layers of the encoder.
Their complete model is composed of:

- An encoder that combines $N = 6$ modules each composed of a multi-head attention sub-module, and a [per-component] one hidden-layer MLP, with residual pass-through and layer normalization.
- A decoder with a similar structure, but with causal attention layers to allow for regression training, and additional attention layers that attend to the layers of the encoder.

Positional information is provided through an additive positional encoding of same dimension $d_{\text{model}}$ as the internal representation, and is of the form

$$PE_{t,2i} = \sin \left( \frac{t}{10,000 \frac{2i}{d_{\text{model}}}} \right)$$

$$PE_{t,2i+1} = \cos \left( \frac{t}{10,000 \frac{2i+1}{d_{\text{model}}}} \right).$$
3.1 Encoder and Decoder Stacks

Encoder: The encoder is composed of a stack of $N = 6$ identical layers. Each layer has two sub-layers. The first is a multi-head self-attention mechanism, and the second is a simple, position-wise fully connected feed-forward network. We employ a residual connection \cite{11} around each of the two sub-layers, followed by layer normalization \cite{1}. That is, the output of each sub-layer is $\text{LayerNorm}(x + \text{Sublayer}(x))$, where $\text{Sublayer}(x)$ is the function implemented by the sub-layer itself. To facilitate these residual connections, all sub-layers in the model, as well as the embedding layers, produce outputs of dimension $d_{\text{model}} = 512$.

Decoder: The decoder is also composed of a stack of $N = 6$ identical layers. In addition to the two sub-layers in each encoder layer, the decoder inserts a third sub-layer, which performs multi-head attention over the output of the encoder stack. Similar to the encoder, we employ residual connections around each of the sub-layers, followed by layer normalization. We also modify the self-attention sub-layer in the decoder stack to prevent positions from attending to subsequent positions. This masking, combined with the fact that the output embeddings are offset by one position, ensures that the predictions for position $i$ can depend only on the known outputs at positions less than $i$.

3.2 Attention

An attention function can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key.
The architecture is tested on English-to-German and English-to-French translation using the standard WMT2014 datasets.

- English-to-German: 4.5M sentence pairs, 37k tokens vocabulary.
- English-to-French: 36M sentence pairs, 32k tokens vocabulary.
- 8 P100 GPUs (150 TFlops FP16), 0.5 day for the small model, 3.5 days for the large one.
Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU</th>
<th>Training Cost (FLOPs)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EN-DE</td>
<td>EN-FR</td>
</tr>
<tr>
<td>ByteNet [18]</td>
<td>23.75</td>
<td></td>
</tr>
<tr>
<td>Deep-Att + PosUnk [39]</td>
<td>39.2</td>
<td></td>
</tr>
<tr>
<td>GNMT + RL [38]</td>
<td>24.6</td>
<td>39.92</td>
</tr>
<tr>
<td>ConvS2S [9]</td>
<td>25.16</td>
<td>40.46</td>
</tr>
<tr>
<td>MoE [32]</td>
<td>26.03</td>
<td>40.56</td>
</tr>
<tr>
<td>Deep-Att + PosUnk Ensemble [39]</td>
<td>40.4</td>
<td></td>
</tr>
<tr>
<td>GNMT + RL Ensemble [38]</td>
<td>26.30</td>
<td>41.16</td>
</tr>
<tr>
<td>ConvS2S Ensemble [9]</td>
<td>26.36</td>
<td>41.29</td>
</tr>
<tr>
<td>Transformer (base model)</td>
<td>27.3</td>
<td>38.1</td>
</tr>
<tr>
<td>Transformer (big)</td>
<td><strong>28.4</strong></td>
<td><strong>41.8</strong></td>
</tr>
</tbody>
</table>

(Vaswani et al., 2017)
The Law will never be perfect, but its application should be just—this is what we are missing, in my opinion.

Figure 4: Two attention heads, also in layer 5 of 6, apparently involved in anaphora resolution. Top: Full attentions for head 5. Bottom: Isolated attentions from just the word 'its' for attention heads 5 and 6. Note that the attentions are very sharp for this word.

(Vaswani et al., 2017)
The Law will never be perfect, but its application should be just—this is what we are missing, in my opinion.

(Vaswani et al., 2017)
The Universal Transformer (Dehghani et al., 2018) is a similar model where all the blocks are identical, resulting in a recurrent model that iterates over consecutive revisions of the representation instead of positions.

Additionally the number of steps is modulated per position dynamically.
Transformer self-training and fine-tuning for NLP
The transformer networks were introduced for translation, and trained with a supervised procedure, from pairs of sentences.

However, as for word embeddings, they can be trained in an unsupervised manner, for auto-regression or as denoising auto-encoders, from very large data-sets, and fine-tuned on supervised tasks with small data-sets.
Figure 3: Differences in pre-training model architectures. BERT uses a bidirectional Transformer. OpenAI GPT uses a left-to-right Transformer. ELMo uses the concatenation of independently trained left-to-right and right-to-left LSTMs to generate features for downstream tasks. Among the three, only BERT representations are jointly conditioned on both left and right context in all layers. In addition to the architecture differences, BERT and OpenAI GPT are fine-tuning approaches, while ELMo is a feature-based approach.

(Devlin et al., 2018)
GPT (Generative Pre-Training, Radford, 2018) is a transformer trained for auto-regressive text generation.
“GPT-2 is a large transformer-based language model with 1.5 billion parameters, trained on a dataset of 8 million web pages. GPT-2 is trained with a simple objective: predict the next word, given all of the previous words within some text. The diversity of the dataset causes this simple goal to contain naturally occurring demonstrations of many tasks across diverse domains. GPT-2 is a direct scale-up of GPT, with more than 10X the parameters and trained on more than 10X the amount of data.”

(Radford et al., 2019)
“GPT-2 is a large transformer-based language model with 1.5 billion parameters, trained on a dataset of 8 million web pages. GPT-2 is trained with a simple objective: predict the next word, given all of the previous words within some text. The diversity of the dataset causes this simple goal to contain naturally occurring demonstrations of many tasks across diverse domains. GPT-2 is a direct scale-up of GPT, with more than 10X the parameters and trained on more than 10X the amount of data.”

(Radford et al., 2019)

The GPT-3 model has 175B parameters and was trained on 300B tokens from various sources (Brown et al., 2020).
We can use HuggingFace's pre-trained models (https://huggingface.co/).

```python
import torch

from transformers import GPT2Tokenizer, GPT2LMHeadModel

tokenizer = GPT2Tokenizer.from_pretrained('gpt2')
model = GPT2LMHeadModel.from_pretrained('gpt2')
model.eval()

tokens = tokenizer.encode('Studying Deep-Learning is')

for k in range(100):  # no more than 100 tokens
    outputs = model(torch.tensor([tokens])).logits
    next_token = torch.argmax(outputs[0, -1])
    print(k, next_token)
    tokens.append(next_token)
    if tokenizer.decode([next_token]) == '.': break

print(tokenizer.decode(tokens))
```

prints

```
Studying Deep-Learning is a great way to learn about the world around you.
```
Large models of this class have been shown to exhibit some “zero shot learning” capabilities when they are properly “primed” (Brown et al., 2020).

For instance using HuggingFace’s gpt2-xl model with 1.6B parameters, we can get these sentence completions, where the priming text is between <>:

<Cherry is red, lettuce is green, lemon is> yellow, and orange is blue.

<Cherry is sweet, lettuce is bland, lemon is> sour, and orange is bitter.

<Cherry is a fruit, lettuce is a vegetable, lemon is> a fruit, and so on.
BERT (Bidirectional Encoder Representation from Transformers, Devlin et al., 2018) is a transformer pre-trained with:

- Masked Language Model (MLM), that consists in predicting [15% of] words which have been replaced with a “MASK” token.
- Next Sentence Prediction (NSP), which consists in predicting if a certain sentence follows the current one.

It is then fine-tuned on multiple NLP tasks.
Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. \([\text{CLS}]\) is a special symbol added in front of every input example, and \([\text{SEP}]\) is a special separator token (e.g. separating questions/answers).

(Devlin et al., 2018)
The Stanford Sentiment Treebank is a binary single-sentence classification task consisting of sentences extracted from movie reviews with human annotations of their sentiment (Socher et al., 2013).

CoLA
The Corpus of Linguistic Acceptability is a binary single-sentence classification task, where the goal is to predict whether an English sentence is linguistically "acceptable" or not (Warstadt et al., 2018).

STS-B
The Semantic Textual Similarity Benchmark is a collection of sentence pairs drawn from news headlines and other sources (Cer et al., 2017). They were annotated with a score from 1 to 5 denoting how similar the two sentences are in terms of semantic meaning.

MRPC
Microsoft Research Paraphrase Corpus consists of sentence pairs automatically extracted from online news sources, with human annotations for whether the sentences in the pair are semantically equivalent (Dolan and Brockett, 2005).

RTE
Recognizing Textual Entailment is a binary entailment task similar to MNLI, but with much less training data (Bentivogli et al., 2009).

WNLI
Winograd NLI is a small natural language inference dataset (Levesque et al., 2011).

The GLUE webpage notes that there are issues with the construction of this dataset, and every trained system that's been submitted to GLUE has performed worse than the 65.1 baseline accuracy of predicting the majority class. We therefore exclude this set to be fair to OpenAI GPT. For our GLUE submission, we always predicted the majority class.

Note that we only report single-task fine-tuning results in this paper. A multitask fine-tuning approach could potentially push the performance even further. For example, we did observe substantial improvements on RTE from multi-task training with MNLI.
Head 9-6
- Prepositions attend to their objects
- 76.3% accuracy at the pobj relation

Head 8-11
- Noun modifiers (e.g., determiners) attend to their noun
- 94.3% accuracy at the det relation

(Clarke et al., 2019)
- **Possessive pronouns** and apostrophes attend to the head of the corresponding NP
- 80.5% accuracy at the **poss** relation

- **Prepositions** attend to their objects
- 76.3% accuracy at the **pobj** relation

- **Noun modifiers** (e.g., determiners) attend to their noun
- 94.3% accuracy at the **det** relation

- **Direct objects** attend to their verbs
- 86.8% accuracy at the **dobj** relation

- **Passive auxiliary verbs** attend to the verb they modify
- 82.5% accuracy at the **auxpass** relation

- **Coreferent mentions** attend to their antecedents
- 65.1% accuracy at linking the head of a coreferent mention to the head of an antecedent

(Clarke et al., 2019)
Head 9-6
- **Prepositions** attend to their objects
- 76.3% accuracy at the `pobj` relation

Head 5-4
- **Coreferent** mentions attend to their antecedents
- 65.1% accuracy at linking the head of a coreferent mention to the head of an antecedent

(Clark et al., 2019)
Attention in computer vision
Wang et al. (2018) proposed an attention mechanism for images, following the model from Vaswani et al. (2017).

\[ y = \text{softmax} \left( (W_\theta x)^\top (W_\phi x) \right) W_g x. \]
They insert “non-local blocks” in residual architectures and get improvements on both video and images classification.

Figure 2. A spacetime **non-local block**. The feature maps are shown as the shape of their tensors, e.g., $T \times H \times W \times 1024$ for 1024 channels (proper reshaping is performed when noted). “$\odot$” denotes matrix multiplication, and “$\oplus$” denotes element-wise sum. The softmax operation is performed on each row. The blue boxes denote $1 \times 1 \times 1$ convolutions. Here we show the embedded Gaussian version, with a bottleneck of 512 channels. The vanilla Gaussian version can be done by removing $\theta$ and $\phi$, and the dot-product version can be done by replacing softmax with scaling by $1/N$.

(Wang et al., 2018)
Figure 3. Examples of the behavior of a non-local block in res3 computed by a 5-block non-local model trained on Kinetics. These examples are from held-out validation videos. The starting point of arrows represents one $x_i$, and the ending points represent $x_j$. The 20 highest weighted arrows for each $x_i$ are visualized. The 4 frames are from a 32-frame input, shown with a stride of 8 frames. These visualizations show how the model finds related clues to support its prediction.

(Wang et al., 2018)
Ramachandran et al. (2019) replaced convolutions with local attention.

\[
y_{i,j} = \sum_{(a,b) \in \mathcal{N}(i,j)} W_{i-a,j-b} x_{a,b} \quad \text{(Convolution)}
\]

\[
y_{i,j} = \sum_{(a,b) \in \mathcal{N}(i,j)} \text{softmax}_{a,b} \left( (W_Q x_{i,j})^\top (W_K x_{a,b}) \right) v_{a,b} \quad \text{(Local attention)}
\]

Figure 2: An example of a 3 × 3 convolution. The output is the inner product between the local window and the learned weights.

Figure 3: An example of a local attention layer over spatial extent of \( k = 3 \).

(Ramachandran et al., 2019)
Table 1: ImageNet classification results for a ResNet network with different depths.

Baseline is a standard ResNet, Conv-stem + Attention uses spatial convolution in the stem and attention everywhere else, and Full Attention uses attention everywhere including the stem. The attention models outperform the baseline across all depths while having 12% fewer FLOPS and 29% fewer parameters.

Figure 5: Comparing parameters and FLOPS against accuracy on ImageNet classification across a range of network widths for ResNet-50. Attention models have fewer parameters and FLOPS while improving upon the accuracy of the baseline.

(Ramachandran et al., 2019)
“A fully attentional network based off of the proposed stand-alone local self-attention layer achieves competitive predictive performance on ImageNet classification and COCO object detection tasks while requiring fewer parameters and floating point operations than the corresponding convolution baselines.”

(Ramachandran et al., 2019)
Cordonnier et al. (2020) showed that provided with proper positional encoding multi-head multiplicative attention layers can encode convolutions with filter support of size the number of heads:

“A multi head self-attention layer with $N_h$ heads of dimension $D_h$, output dimension $D_{out}$ and a relative positional encoding of dimension $D_p \geq 3$ can express any convolutional layer of kernel size $\sqrt{N_h} \times \sqrt{N_h}$ and $\min(D_h, D_{out})$ output channels.”

(Cordonnier et al., 2020)
Figure 5: Attention probabilities of each head (column) at each layer (row) using learned relative positional encoding without content-based attention. The central black square is the query pixel. We reordered the heads for visualization and zoomed on the 7x7 pixels around the query pixel.

Figure 6: Attention probabilities for a model with 6 layers (rows) and 9 heads (columns) using learned relative positional encoding and content-content based attention. Attention maps are averaged over 100 test images to display head behavior and remove the dependence on the input content. The black square is the query pixel. More examples are presented in Appendix A.

(Cordonnier et al., 2020)
References


