# CNN vs Transformer Architecture in Machine Translation

Maxi Fischer, Sophie Gräfnitz, Sriyal Jayasinghe (Team Fishbone)

## INTRODUCTION & BACKGROUND

Our project was to compare two relatively new Sequence to Sequence approaches in Machine Translation, namely Convolutional Neural Networks (CNN) and Transformer architecture. Our project was based on two paper publications. They are as follows:

- *Attention is all you need* (by Ashish Vaswani et al.) - This paper was composed by Google where it proposes a architecture known as **Transformer** which is based on the Attention concept.
- *Convolutional Sequence to Sequence Learning* (by Jonas Gehring et al.) - This paper was composed by Facebook AI Research (FAIR) and it proposes **fully convolutional neural network architecture model** that can be used for machine translation.

### Traditional Machine Translation

- models based on architectures such as RNN, RNN-LSTM are sequential models.
- Such models expect the tokens to be fed into the model in a sequential manner and the translations are also produced in a sequential manner.
- Due to the sequential nature such models have difficulties in learning long-range dependencies.

### Attention Architecture

- parallel learning on all input tokens is performed
- the model learns to distribute its attention to certain structures and words
- According to the architecture proposed even for a multi-head attention the computational effort is similar to one-head attention

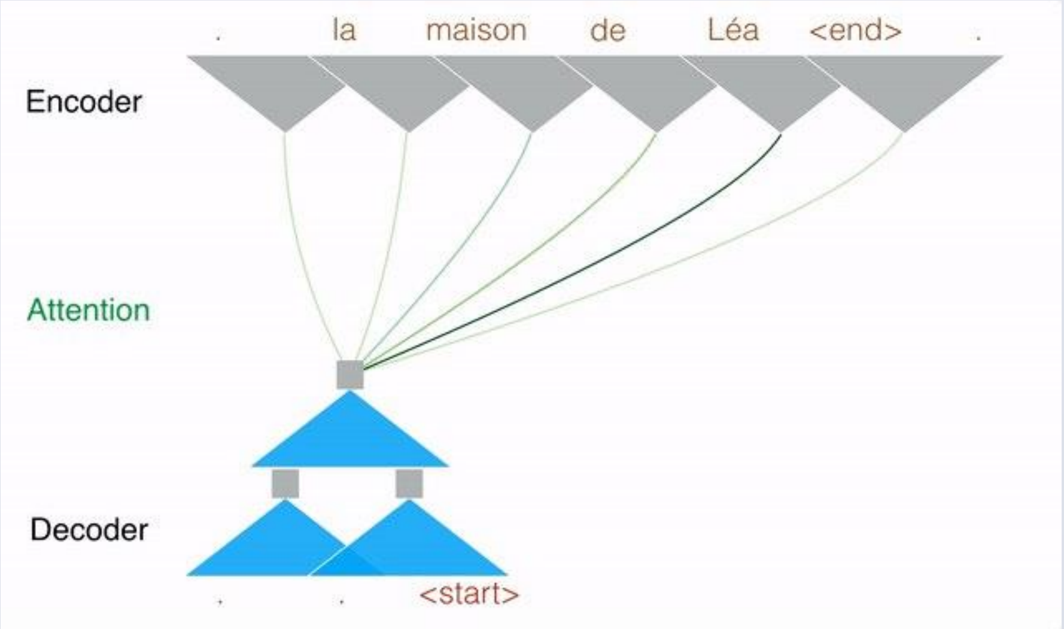
### What is multi-head attention?

- Since with only one attention head it is hard to learn multiple dependencies the paper proposes to apply multi-head attention.
- The model learns how to optimally choose k sentence substructures, that can be passed to k attention heads. Thus, k different internal dependencies can be learned

- In the paper "Attention is all you need" the network consists of 6 layers of encoder and decoder, where both consist of multiple stages and both including multi head attention and a Feed Forward Network which work on attention and positions.

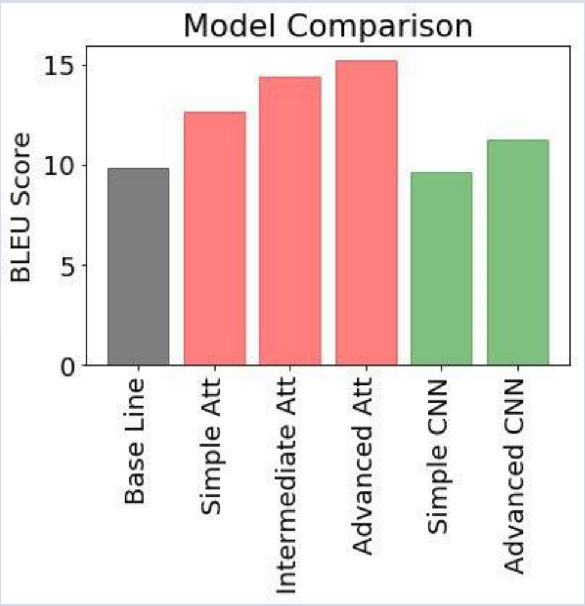
### Convolutional architecture

- is frequently applied in image processing.
- Long-term dependencies can be learned by applying filters on the original data, which means the greater the distance between two tokens, the later their dependence is modelled.
- The filtering steps are applied sequentially, while each filtering step means the convolution of subphrases.



### Comparative Experiment

- *Training dataset*: Europarl (Estonian transcript of a Europian Parliament speech and the corresponding English translation, approx. 650000 sentences)
- *Test and dev dataset*: English dev set and Estonian dev set, each had 2000 sentences, taken from Estonian-English news translations)



| | Base Line | Simple Attention | Intermediate Attention | Advanced Attention | Simple CNN | Advanced CNN |
|---|---|---|---|---|---|---|
| **Encoder and decoder** | RNN-LSTM | Transformer | | | CNN | |
| **Hidden layers** | 600 | 512 | | | | |
| **Activation Type** | Relu | | | | Glu | |
| **Positional embedding** | Fixed | | | | Learned | |
| **Initial learning rate** | 0.0002 | | 0.0001 | 0.0003 | 0.0002 | |
| **Individual** | | | 2048 hidden units in feed forward layers 2048,<br>8 attention heads,<br>Vocabulory size 50000 | 2048 hidden units in feed forward layers 2048,<br>8 attention heads,<br>Weight-tying ON,<br>Learning-rate-warmup 50000,<br>Label-smoothing 0.2 | 6 convol. layers | 8 convol. layers |

## LARGE SCALE EXPERIMENT

### Which model won?

- Best BLEU score: 15.2 ( Advanced Attention)
- The flexible attention on sentence structure outweighs the hierarchical representation in CNN.
- Therefore we trained this model on a combination of 4 corpora and performed tests on the same test set like in the comparative experiment and on another test set.

| CNN Advantages | Transformer Advantages |
|---|---|
| - parallelizable | - parallelizable |
| - Hierarchical representation of sentence structures | - Flexible attentioned sentence structure representation |
| - For words of distance n, kernel size k, O(n/k) convolutions are needed | - For words of distance n, O(1) operations are needed |

| **Data Set** | Training | Early Stopping | In Domain Test | Out of Domain Test |
|---|---|---|---|---|
| **Source** | EMEA, Europarl, JRC-Acquis, OpenSubtitles2018 | EMEA, Europarl, JRC-Acquis, OpenSubtitles2018 | EMEA, Europarl, JRC-Acquis, OpenSubtitles2018 | Estonian-English News |
| **Size** | Approx. 5 Mio sentences, whole corpora | Approx. 2000 sentences, 500 per corpus | Approx. 4000 sentences, 1000 per corpus | 2000 sentences |

### Performance

- BLEU score of 36.17 on the in-domain dataset and
- BLEU score 14.45 on the out-of-domain dataset
- Outperformed the baseline with a BLEU score of 9.83,
- Did not outperform the small dataset model with a BLEU score of 15.20.

### Results of Manual Error Analysis

Our task was to compare our final model's translations of the out of domain data set with the translations provided by *TartuNLP Translator,* also named Neurotolge.

| | Advanced Attention is better | Neurotolge is better |
|---|---|---|
| **Reference** | You just have to rest up calmly and deduce why you're so tired (if you are). | Over the past year, prodded by the government, cellphones have added new tools to counteract unwanted "robocalls." |
| **Neurotolge** | It is simply necessary to rest calmly and to find out why this fatigue is so great (as it is). | Over the last year, new tools have been added to mobile phones to prevent unwanted 'robotic calls' from being encouraged by the government. |
| **Baseline** | It is simply necessary to break out calmly and to haunt why this fatigue is so high (if it). | In the last year, the government inspired by the mobile phones has been included in the new tool for preventing unwanted 'Roma'. |
| **Advanced Attention** | You just have to get out easy and wonder why this fatigue is as big as it is. | Over the last year, the government's inspiration has been added to the Jääs to block new tools to obstruct last year's Bible speeches. |

### Problems of Advanced Attention Model

- proper names are not translated
- long sentences suffered from unnecessary repetitions and grammatical nonsense.
- fluent translations, but still had content errors, informal translations, missing or invented words.
- In comparison to the *TartuNLP Translator* our model was preferred in 21% of the analysed sentences.

### Explanations

- The Advanced Attention model was mainly trained on informal data (OpenSubtitles) lacked formal sentence structures and complicated long sentences.
- Neurotolge was optimized to translate polite and casual style texts

### Conclusions

- Transformer architecture performs far better in comparison to CNN architecture for seq2seq tasks.
- This is mainly due to better sentence structure modelling
- Still, it is not straightforward to train an universal attention only model for formal and informal occasions. An attention model will spread its attention always "in the style" like it was trained. The bad out-of domain performance shows that

### References

*Attention is all you need. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. . In Advances in Neural Information Processing Systems (pp. 5998-6008).*

*Convolutional sequence to sequence learning. Gehring, J., Auli, M., Grangier, D., Yarats, D. and Dauphin, Y.N., 2017.arXiv preprint arXiv:1705.03122.*

*https://github.com/mt2018-tartu-shared-task/final-report-fishbone*