# Building a QA system through local convolution with global self-attention (IID SQuAD track)

*Hugo Vergnes[1] & Hamza El Mosor[2]*

*[1] Statistics Department, Stanford University*

*[2]ICME, Stanford University*

## Abstract:

Recurrent Neural Networks were historically successful architecture for Machine Reading Comprehension. QANet was the first architecture to achieve higher performance with only convolutions and self-attention. In this paper, we discuss its implementation and architecture with an emphasis made on the model use of parameters. We tested the model and some variations of it on the Stanford Question Answering Dataset 2.0

## Problem & Dataset

Given the SQuAD 2.0 dataset [1], where every instance consist of the triplet (question, query, answer) , we want to build a model that predicts the answer from new question/query instance. The SQuAD dataset has some questions where the query isn't answered, in which case the model should outputs no answer.

The training dataset is constituted of:

- train (129,941 examples): All taken from the official SQuAD 2.0 training set
- dev (6078 examples): Roughly half of the official dev set, randomly selected.
- test (5915 examples): The remaining examples from the official dev set, plus hand-labeled examples.

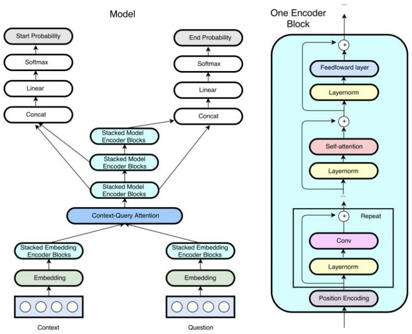
## Baseline & Final model

Our baseline was the BiDAF architecture [2] which was the state of the art when QANet was published. One of the main contribution of that paper was the bidirectional attention flow layer that coupled the query and the context to model that produce a set of query-aware vectors for each word in the context.

→Limitation: probabilities of the beginning and end of the answer are created independently, while one can expect it to be closely related.

## Experiments & Results:

We experimented with the QANet architecture, which consists of an embedding layer, encoder, layer context-query attention layer, modelling and output layers. The encoder block consists of a stack of encoders, and we played with the number of encoder we are putting in this layer. We achieved the highest performance with 6 blocks.



| *Models* | *Dev NLL* | *F1* | *EM* | *AvNA* |
|---|---|---|---|---|
| BiDAF | 3.05 | 56.69 | 56.07 | 66.88 |
| BiDAF w/ character embedding | 3.05 | 60.55 | 57.18 | 67.11 |
| C2Q attentions | 3.24 | 55.56 | 52.33 | 64.17 |
| CoAttention | 3.14 | 58.81 | 55.42 | 66.32 |
| CoAttention + LSTM dropout | 3.34 | 59.96 | 53.22 | 64.18 |
| QANet 6 blocks | **2.55** | **66.36** | **63.23** | **73.34** |
| QANet 5 blocks | 2.59 | 66.23 | 62.75 | 72.51 |
| QANet 4 blocks | 2.57 | 66.28 | 63.08 | 72.11 |
| QANet 3 blocks | 2.63 | 65.67 | 61.97 | 71.57 |
| QANet 5 blocks + 0.4 lr | 2.65 | 65.69 | 62.19 | 72.00 |

## References

[1] Konstantin Lopyrev Percy Liang Pranav Rajpurkar, Jian Zhang. Squad: 100,000+ questions for machine comprehension of text. 2016.
[2] Ali Farhadi Hannaneh Hajishirzi Minjoon Seo, Aniruddha Kembhavi. Bidirectional attention flow for machine comprehension. 2016.
[3] Niki Parmar Jakob Uszkoreit Llion Jones Aidan N. Gomez Lukasz Kaiser Illia Polosukhin Ashish Vaswani, Noam Shazeer. Attention is all you need. 2017.
[4] Minh-Thang Luong Rui Zhao Kai Chen Mohammad Norouzi Quoc V. Le Adams Wei Yu, David Dohan. Qanet: Combining local convolution with global self-attention for reading comprehension. 2018.