QANet Analysis: Default Project (non-PCE)

Timothy Le

Stanford
Computer Science

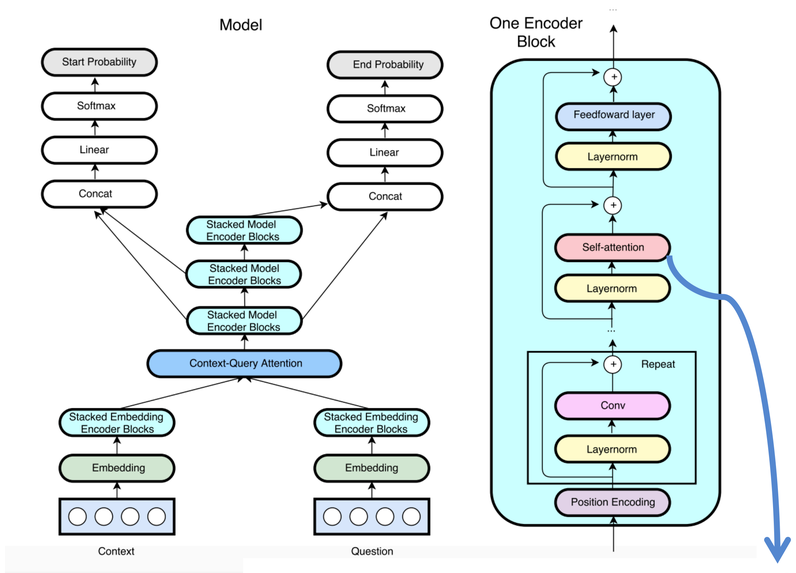
Problem: Given a passage and a query, predict the start and end indices in the passage to answer the question, or predict "No Answer"
Data: The SQuAD is used to approach this problem

- **Question:** What is the polish word for wreaths?
- **Context:** Several commemorative events take place every year. Gatherings of thousands of people on the banks of the Vistula on Midsummer's Night for a festival called Wianki (Polish for Wreaths) have become a tradition and a yearly event in the programme of cultural events in Warsaw. The festival traces its roots to a peaceful pagan ritual where maidens would float their wreaths of herbs on the water to predict when they would be married, and to whom. By the 19th century this tradition had become a festive event, and it continues today. The city council organize concerts and other events. Each Midsummer's Eve, apart from the official floating of wreaths, jumping over fires, looking for the fern flower, there are musical performances, dignitaries' speeches, fairs and fireworks by the river bank.
- **Answer:** Wianki
- **Prediction:** Wianki

Example output from baseline w/ character embeddings

QANet Architecture

Implemented Encoder, Model, and Output layers from QANet model:

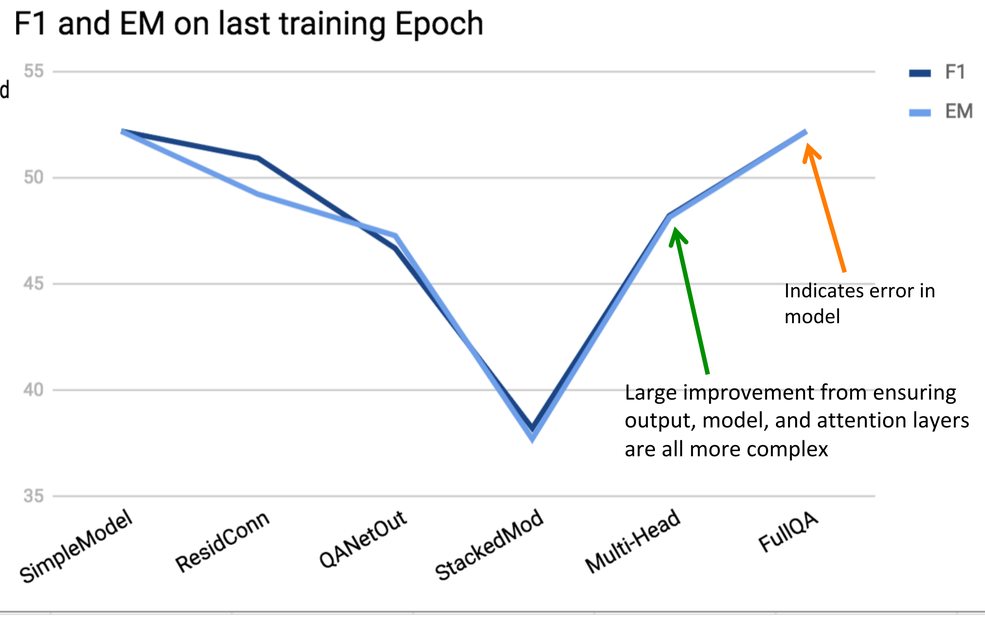


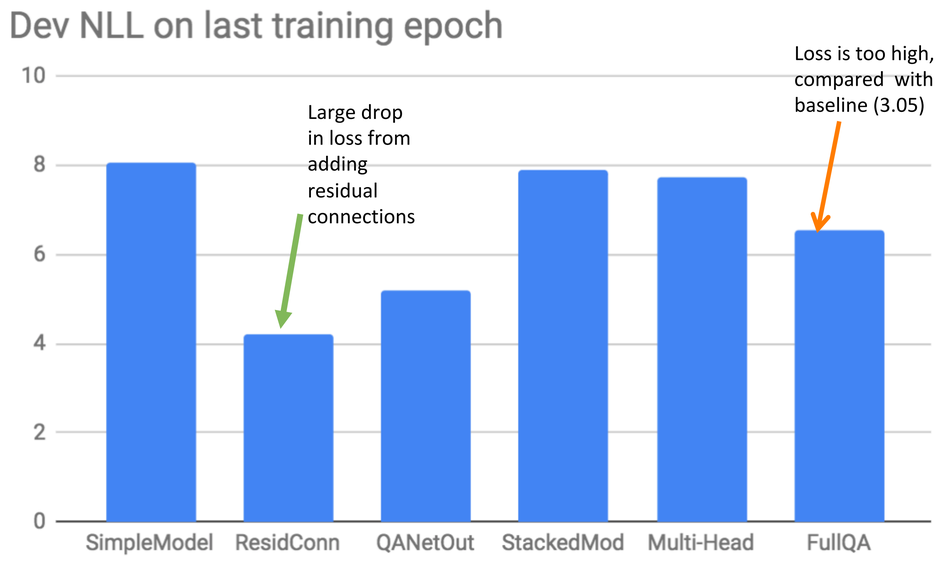
$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1,...,\text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

- QANet diagram from QANet paper[1]
- Attention equations from "Attention is all you need" paper[2]

Effects of making the QANet model more complex





SimpleModel: Single encoder block with single-head attention for encoder and decoder layers, with residual connection only in encoder feed-forward layer
ResidConn: Added residual connection for convolution and self-attention components
QANetOut: Replaced BiDAF output layer with QANet Output Layer, which now takes in three matrices
StackedMod: Replaced three single encoder blocks in model layer with three stacked encoder blocks (stack size=4)
Multi-Head: Replaced single-head attention with multi-head attention in self-attention encoder blocks
FullQA: Includes dropout, layer normalization, and stack size of 6 for stacked encoder blocks in model layer

Baseline:	**Best score:**
BiDAF model w/o character embed	BiDAF baseline with character embeddings:
Train:	**Train:**
Dev NLL: 3.05, F1: 61.53, EM: 58.24	Dev NLL: 3.10, F1: 61.57, EM: 58.12
Test:	**Test:**
NLL: 3.06, F1: 61.27, EM: 58.46	NLL: 2.94, F1: 62.29, EM: 59.10

Conclusion and Future Directions

- Residual connections throughout the encoding layer are vital in reducing the loss (Dev NLL)
- Increased complexity in one component of the model should have corresponding increased complexity in related components of the model
- Future: Tune parameter such as dropout rate to lower training loss and investigate layers since the fullQA model only predicts "No Answer"

References
[1]Adams Wei Yu et al. "Qanet: Combining local convolution with global self-attention for reading comprehension." In: *arXiv preprint arXiv:1804.09541* (2018).
[2]Ashish Vaswani et al. "Attention is all you need". In: *Advances in Neural Information Processing System* (2017).

Stanford University