Investigating Model Combination to Improve Performance on SQuAD
by, Matt Riedman
Stanford Department of Computer Science

Introduction
- Question answering is a useful tool to get info from large databases (e.g., Wikipedia, libraries)
- Present challenge is ensuring a model can efficiently identify most relevant words
- In LSTM-based models, attention layer is key to finding relevant context words

Background
- SQuAD: Widely used Q/A dataset
- 100,000 pairs of context and queries
- Includes unanswerable questions (mimics real life)
- Standard metrics: F1, EM
- BiDAF Model (Seo et al.)

Methods
- Train two similar established models with different attention layers
- BiDAF, Dynamic Coattention
- Try to leverage likelihoods of start and end position from each to produce better predictions than either alone
- General approach: Train models separately, then create a hybrid model using original up to the softmax layer
- Provides framework for leveraging strengths of different models

Experiments
- First: assume softmax output represents actual probability, use normalized elementwise product (i.e., take product of probabilities from each)
- Second: run through training set to see “actual” probabilities given softmax values, use those in first method
- Third: concatenate hidden layer prior to softmax and use nonlinearity then softmax

Results and Analysis
- BiDAF EM: 57.60, F1: 67.95
- Coattention EM: 63.20, F1: 73.56
- Strangely, experiments 1 and 2 both provide lower EM, F1 metrics than either of the original models
- I suspect in cases where one model is wrong, softmax probability drags down the other’s predictions
- Experiment 3 results pending

Conclusions
- Combining models is harder than anticipated
- Rather than producing a summative effect where models combine their predictive capability, naive approaches seem to collide

References