Why and when should you pool? Analyzing Pooling in Recurrent Architectures

Pratyush Maini†, Keshav Kolluru†, Danish Pruthi‡, Mausam†

† Indian Institute of Technology, Delhi, India | ‡ Carnegie Mellon University, Pittsburgh, USA

Carnegie Mellon University

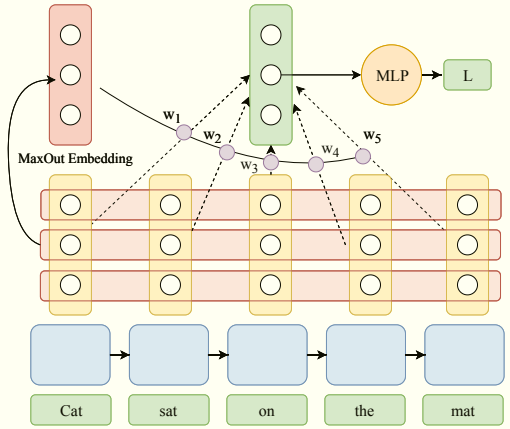
Key Results: 1. Pooling (and attention) based BiLSTMs demonstrate enhanced learning ability and positional invariance.
2. Pooling improves sample efficiency in low-resource settings and is beneficial when salient words lie towards the middle of the sentence
3. We propose max-attention which achieves higher accuracy on classification tasks & is more robust to distribution shift

CODE LINK IN PAPER

Motivation

- Pooling the hidden-states is standard practice in RNNs
- However, *Why* and *When* pooling benefits these models is largely unexamined.

Max-attention

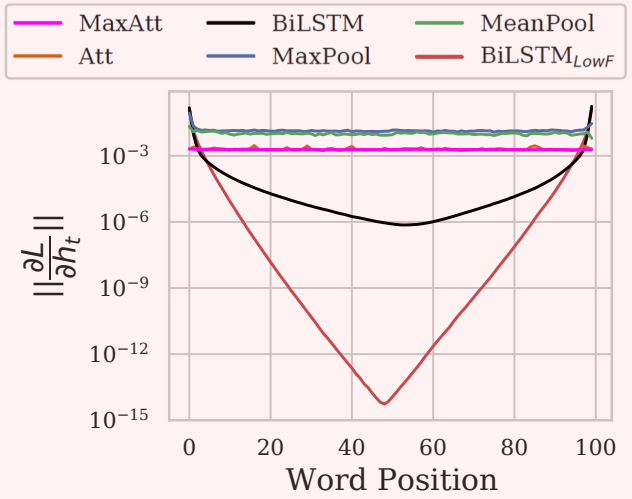


Unique locally-informed query vector = max-pooled embedding for every sentence.

Max-attention is the **best performing** model in ~80% of all experimental settings discussed.

Gradient Propagation

• **Beginning of training:** Gradients in BiLSTMs vanish towards the middle of the sentence (nearly 10^{-10} times that at the ends). Pooling mitigates vanishing.



• Vanishing decreases as training proceeds

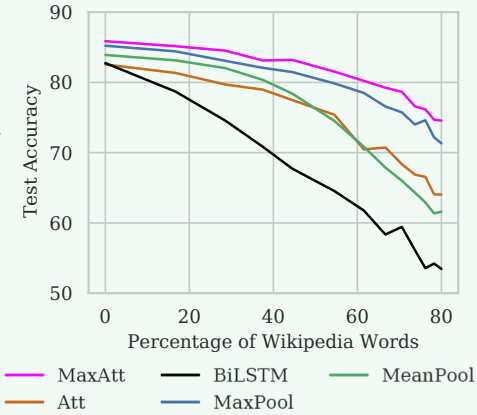
IN PAPER

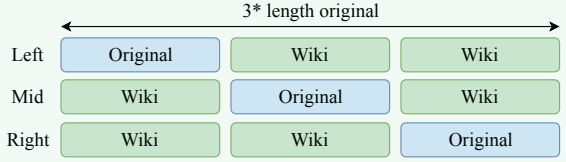
• **Low resource settings:** BiLSTMs prematurely *memorize* the training data -- based on the starting & ending few words

Positional Biases

Can organically trained RNNs skip over unimportant words?

• We append varying amounts of random Wikipedia sentences to the original data at test time.
• Performance ↓ significantly for BiLSTM & mean-pool.





How does the position of a word affect prediction?

- NWI metric to calculate per-position importance of words.
- **Pooled architectures:** No bias w.r.t. word position
- **BiLSTM:** Huge bias towards the end words *even* when the original sentence is in the mid

Can models be trained to skip unrelated words?

Not always! BiLSTM accuracy in mid setting = majority class baseline in low-resource datasets.