**Attention-Based Models for Speech Recognition**

**Abstract:**

Recurrent sequence generators conditioned on input data through an attention mechanism have recently shown very good performance on a range of tasks including machine translation, handwriting synthesis [1, 2] and image caption generation [3]. We extend the attention-mechanism with features needed for speech recognition. We show that while an adaptation of the model used for machine translation in [2] reaches a competitive 18.7% phoneme error rate (PER) on the TIMIT phoneme recognition task, it can only be applied to utterances which are roughly as long as the ones it was trained on. We offer a qualitative explanation of this failure and propose a novel and generic method of adding location-awareness to the attention mechanism to alleviate this issue. The new method yields a model that is robust to long inputs and achieves 18% PER in single utterances and 20% in 10-times longer (repeated) utterances. Finally, we propose a change to the attention mechanism that prevents it from concentrating too much on single frames, which further reduces PER to 17.6% level.

**Conclusions:**

We proposed and evaluated a novel end-to-end trainable speech recognition architecture based on a hybrid attention mechanism which combines both content and location information in order to select the next position in the input sequence for decoding. One desirable property of the proposed model is that it can recognize utterances much longer than the ones it was trained on. In the future, we expect this model to be used to directly recognize text from speech [10, 17], in which case it may become important to incorporate a monolingual language model to the ARSG architecture [26]. This work has contributed two novel ideas for attention mechanisms: a better normalization approach yielding smoother alignments and a generic principle for extracting and using features from the previous alignments. Both of these can potentially be applied beyond speech recognition. For instance, the proposed attention can be used without modification in neural Turing machines, or by using 2–D convolution instead of 1–D, for improving image caption generation [3].