Assignment 8

Fudi(Fred) Wang

December 2019

Q1

The pointer-generator network is a mixture of the baseline model(Seq2Seq + Attention) and a pointer network. Not only can it allow copying words by pointing, but it can also generate words from a fixed vocabulary.

The addition of a generation probability p_{gen} for each decoder timestep t weights the probability of generating words from the vocabulary versus copying words from the source text. Then the probability distribution over the extended vocabulary is just the sum of p_{gen} times P_{vocab}, which is the probability distribution over all words in the vocabulary and $(1 - p_{gen})$ times the attention distribution. Other components such as the loss function, remains the same as the baseline model. Note that the ability to produce OOV words is one of the advantages of pointer-generator models.

Q2

The Coverage mechanism mainly solves the problem of repetition in previous models. We maintain a coverage vector c^t, which is the sum of attention distributions over all previous decoder timesteps and is used as extra input to the attention mechanism. We also define a coverage loss to penalize repeatedly attending to the same locations. The coverage loss, reweighted by some hyperparameter λ, is added to the primary loss function to yield a new composite loss function.