Abstract

We describe hyperparameters and details of our model training and evaluation.

3 The Learning Framework

3.1 Base Language Model

We use a 2-layer GRU (Cho et al., 2014) with a hidden size of 1024 for each layer. Following (Inan et al., 2017) we tie the input and output embedding layers’ parameters. We use an Adaptive Softmax for the final layer (Grave et al., 2016), which factorizes the prediction of a token into first predicting the probability of $k$ (in our case $k = 3$) clusters of words that partition the vocabulary and then the probability of each word in a given cluster. To regularize we dropout (Srivastava et al., 2014) cells in the output layer of the first layer with probability 0.2. We use mini-batch stochastic gradient descent (SGD) and anneal the learning rate when the validation set performance fails to improve, checking every 1000 batches. Learning rate, annealing rate, and batch size were tuned on the validation set for each dataset. Gradients are backpropagated 35 time steps and clipped to a maximum value of 0.25.

3.2 Cooperative Communication Models

For all the models except the entailment model, training is performed with Adam (Kingma and Ba, 2015) with batch size 64 and learning rate 0.01. The classifier’s hidden layer size is 300. Dropout is performed on both the input word embeddings and the non-linear hidden layer before classification with rate 0.5.

Word embeddings are kept fixed during training for the repetition model, but are fine-tuned for all the other models.

3.2.2 Entailment Model

We mostly follow the hyperparameters of Parikh et al. (2016): Word embeddings are projected to a hidden size of 200, which are used throughout the model. Optimization is performed with AdaGrad (Duchi et al., 2011) with initial learning rate 1.0 and batch size 16. Dropout is performed at rate 0.2 on the hidden layers of the 2-layer MLPs in the model.

Our entailment classifier obtains 82% accuracy on the SNLI validation set and 68% accuracy on the MultiNLI validation set.

3.2.3 Relevance Model

The convolutional layer is a one-dimensional convolution with filter size 3 and stride 1; the input sequences are padded such that the input and output lengths are equal.

4 Experiments

4.1 Corpora

For the language model and discriminators we use a vocabulary of 100,000 words – we found empirically that larger vocabularies lead to better generation quality. To train our discriminators and evaluate our models, we use segments of length 10, using the first 5 sentences as context and the second 5 as the reference continuation. For TripAdvisor we use the first 10 sentences of reviews of length at least 10. For the BookCorpus we split books into segments of length 10. We select 20% of each corpus as held-out data (the rest is used for language model training). From the held-out data we select a test set of 2000 examples and two validation sets of 1000 examples each, one of which is used to train the mixture weights of the decoding objective. The rest of the held-out data is used to train the discriminative classifiers.
4.2 Baselines

**CACHELM**  Due to memory constraints, we use a vocabulary size of 50k for CACHELM. Beam search decoding is used, with a beam size 5.

**SEQGAN**  The implementation we used adds a number of modelling extensions to the original SeqGAN. In order to make training tractable, the vocabulary is restricted to 25k words, the maximum sequence length is restricted to 250, Monte Carlo rollouts to length 4, and the discriminator updated once for every 10 generator training steps. Greedy decoding sampling with temperature 0.7 was found to work better than beam search.

**SEQ2SEQ**  Due to memory constraints, we use a vocabulary size of 50k for SEQ2SEQ. Beam search decoding is used, with a beam size 5.

4.3 Evaluation Setup

The forms used on Amazon Mechanical Turk are pictured in Tables 1, 2, 3, and 4.

References


Table 1: The first half of the form for the BookCorpus human evaluation.
Table 2: The second half of the form for the BookCorpus human evaluation.
TripAdvisor

Table 3: The first half of the form for the TripAdvisor human evaluation.
Table 4: The second half of the form for the TripAdvisor human evaluation.