Supplementary Material

1 Model settings

1.1 MT

Model. In the experiments with Transformer, we employ 512 dimensions for word embedding, 6 layers for both the encoders and the decoder with 512 units, and feed-froward dimension of 1,024. In order to avoid over-fitting, we use attention and residual dropout by setting the dropout probability to 0.2, along with the label-smoothing with parameter equal to 0.1.

Optimization. We trained the model using the Adam optimizer with batch size of 3,072 tokens, learning rate of 2.0 and the warm-up strategy and exponential decay introduced by (Vaswani et al., 2017) with the warm-up steps equal to 8,000. We also employ the beam search with beam width of 4 to sample hypotheses from the model.

1.2 ASR and STL

The end-to-end ASR and STL systems are built and trained using the same settings.

Model. The first two fully-connected layers have size, respectively, of 256 and 128 units and they are both followed by tanh non-linearity. The fully-connected layers are followed by two 2D convolutional layers, each having 16 output channels and stride 2 in both dimensions. After the convolutions, the output tensor is flattened with dimension 512, and it is passed as input to a stack of 3 bi-directional LSTMs with hidden size of 512 (256 for each direction). The decoder consists of a two-layered deep transition LSTM with size 512 in both layers. The attention network is placed between the two decoder LSTM layers and uses a general attention scoring function. The LSTM output is concatenated with the attention output and the current character embedding and processed by a final fully-connected layer with output size 256 and tanh non-linearity. The embedding size is 256. The total number of parameters is about $9.1M$. The initial states of the encoder LSTMs are learnable parameters of the network, while the first state of the decoder is initialized with last encoder layer state.

Optimization. We trained the models using the Adam optimizer with β values of (0.9, 0.999) and an initial learning rate 0.001 and we clip the gradient when it exceeds a norm value of 5. We set the dropout at rate 0.2 in all the layers, except for the input in both encoder and decoder, and no dropout is set in the recurrent connections.

2 Sample of the MuST-C Corpus

A sample of the dataset is uploaded in the *softconf* system. The sample includes the following folders:

- **txt**, containing – for each of the 8 language directions included in the corpus:
  1. ten sentence pairs (*src* + *tgt* files)
  2. a yaml file which contains – for each sentence pair – information about the audio corresponding to the English sentence, namely its duration, offset, the speaker id, and the name of the *wav* file of the TED Talk including it;
- **wav**, containing the audio file of the TED talk that includes the English sentences of the pairs in the *txt* folder. In this uploaded sample, only a 3-minute subset of the entire TED talk is released;
- **h5**, containing the log Mel 40-dimensional filter-bank features for the audio segments corresponding to the English-Portuguese section of the sample.

This subset of the corpus is an example of all the

information that will be made available in the official MuST-C release.

References

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of NIPS 2017*, pages 5998–6008, Long Beach, CA, USA.