Hierarchical Neural Story Generation: Supplementary Material

1 Model Architectures

1.1 GCNN Language Model + Self-Attention

9 layers with hidden unit sizes $512 \times 4, 768 \times 2, 1024 \times 3$ and convolutional kernel widths $4 \times 2, 1, 4 \times 3, 1, 3 \times 2$. Learning rate 1, momentum 0.99, dropout 0.1, embedding size 300, l2 normalization $1e^{-7}$, 4 decoder self-attention heads.

1.2 Conv seq2seq + self-attention

3 layers in encoder with hidden unit sizes $128 \times 2, 512$ and convolutional kernel widths 3×3. 8 layers in the decoder with hidden unit sizes $512 \times 4, 768 \times 2, 1024$ with convolutional kernel widths 4×8. Learning rate 0.25, momentum 0.99, dropout 0.3, embedding size 256, output embedding size 256, l2 nomalization $1e^{-7}$, 4 decoder self-attention heads.

1.3 Ensemble: Conv seq2seq + self-attention

Two different Conv seq2seq models were trained and ensembled together by averaging with equal weights.

1.4 Fusion: Conv seq2seq + self-attention

The pretrained seq2seq model is the model in Section 1.2. The additional fused model has the following architecture:

5 layers in the encoder with hidden unit sizes $128 \times 2, 512 \times 3$ and convolutional kernel widths 3×5. 5 layers in the decoder with hidden unit sizes $512 \times 3, 768 \times 2$ and convolutional kernel widths 4×5. Learning rate 0.25, momentum 0.99, dropout 0.3, embedding size 256, output embedding size 256, l2 normalization $1e^{-7}$, 4 decoder self-attention heads.