CS7015-Deep Learning
Programming Assignment 4

Jeshuren Chelladurai (CS17M017)
Ajith Kumar M (CS17M009)

April 22, 2018

1. **Mathematical Formulation for Basic Model without attention:**
   Consider W_{ei} to be the word embedding matrices.
   Encoder:
   $x_{it} = W_e w_{it}, t \in [1, T]$
   $\overrightarrow{h_{it}} = \overrightarrow{LSTM}(x_{it}), t \in [1, T]$
   $\overleftarrow{h_{it}} = \overleftarrow{LSTM}(x_{it}), t \in [1, T]$
   $h_{it} = [\overrightarrow{h_{it}}, \overleftarrow{h_{it}}]$
   Decoder:
   $z_{it} = W_{e2} h_{it}, t = 0$
   $z_{it} = W_{e2} y_{it-1}, t \in [1, T']$
   $(y_i, s_i) = softmax(LSTM(z_{it-1}, s_{it-1})), t \in [1, T']$

   Mathematical Formulation for Hierarchical Model without attention:
   Encoder:
   for j = 1 to len(fieldWords)

   1) $x_{it} = W_e w_{it}, t \in [1, T_f]$
   2) $\overrightarrow{h_{it}} = \overrightarrow{LSTM}(x_{it}), t \in [1, T_f]$
   3) $\overleftarrow{h_{it}} = \overleftarrow{LSTM}(x_{it}), t \in [1, T_f]$
   4) $h_{it} = [\overrightarrow{h_{it}}, \overleftarrow{h_{it}}]$
   5) $emb_{it} = [h_{it}, W_e w^j_{it}]$

   $g_{it} = LSTM_2(emb_{it}, S_{it-1}), t \in [1, T_{fw}]$
   Decoder:
   $z_{it} = W_{e2} g_{it}, t = 0$
   $z_{it} = W_{e2} y_{it-1}, t \in [1, T']$
   $(y_i, s_i) = softmax(LSTM(z_{it-1}, s_{it-1})), t \in [1, T']$

2. **Learning curve for Basic encoder-decoder model:**

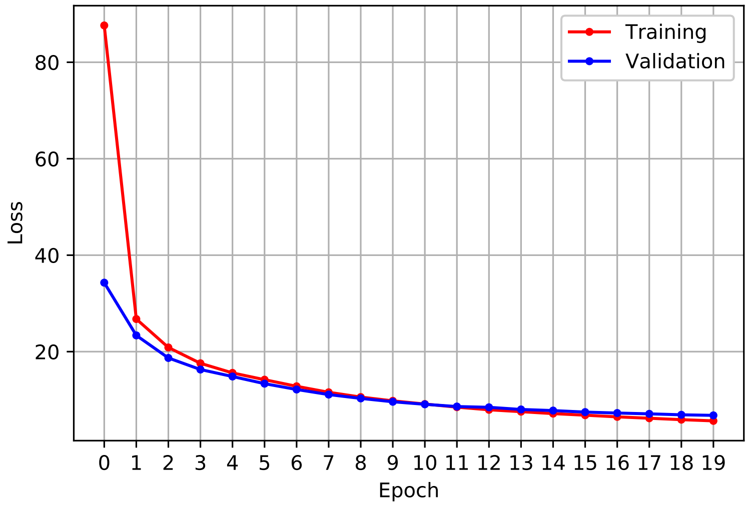

Figure 1: Learning curve for Basic model.

3. **Learning curve for Hierarchical encoder-decoder model:** The architecture was,

    (a) Encoder embedding size: 256

    (b) Level 1 bidirectional LSTM encoder: 512 units

    (c) Dropout with 0.5 probability

    (d) Level 2 bidirectional LSTM encoder: 512 units

    (e) LSTM Decoder : 1024 units

    (f) Encoder embedding size: 256

    (g) Dense Layer : Decoder Vocabulary Size

    (h) Batch Size: 250

    (i) Optimizer: Adam

    (j) Learning Rate: 0.001

    (k) Epochs: 20

    We also tried adding attention on top of the level 2 encoder. We were able to get a BLEU score of 0.8 on the test data

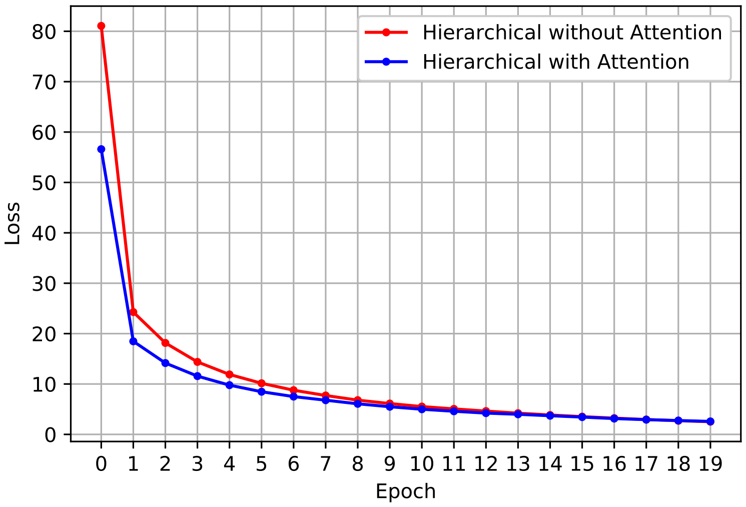

Figure 2: Learning curve (Training data) for Hierarchical Model with and without Attention Mechanism.

4. **Learning curve for Basic encoder-decoder model with attention:**

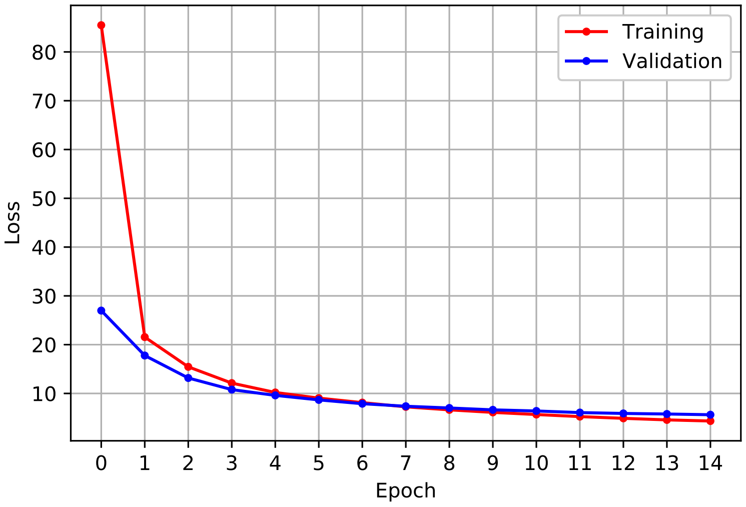

Figure 3: Learning curve for Basic Model with Attention Mechanism.

5. **BLEU scores on test data**

Table 1: BLEU scores of different models on test data

Model	**Without Attention**	**With Attention**
Basic	0.7066	0.7356
Heirarchical	0.836	0.7915

Table 2: Summaries generated by different models

Input	**Basic Model**	**Basic Model w/attention**	**Hierarchical Model**	**Hierarchical Model w/attention**
Sequence 16	Mostly cloudy , with a low around 44 . West wind around 6 mph becoming calm .	Mostly cloudy , with a low around 46 . South wind between 3 and 6 mph .	Mostly cloudy , with a low around 45 . South wind around 6 mph becoming calm .	Mostly cloudy , with a low around 44 . East wind around 5 mph becoming calm .
Sequence 7	Snow . Low around 39 . South southwest wind between 7 and 10 mph . Chance of precipitation is 90 % . New snow accumulation of 2 to 4 inches possible .	Flurries . Mostly cloudy , with a low around -21 . Wind chill values as low as -40 . West southwest wind between 6 and 8 mph .	Snow . Low around 38 . West southwest wind 5 to 8 mph becoming east . Chance of precipitation is 100 % . New snow accumulation of 1 to 3 inches possible .	Flurries . Cloudy , with a low around -21 . Wind chill values as low as -40 . West wind between 5 and 8 mph .
Sequence 502	A chance of rain , mainly after noon . Cloudy , with a high near 54 . Calm wind becoming south between 4 and 7 mph . Chance of precipitation is 40 % .	A chance of showers , mainly after 1pm . Cloudy , with a high near 54 . Calm wind becoming south around 6 mph . Chance of precipitation is 30 % . amounts of less than a tenth of an inch possible . amounts than a tenth and quarter of an inch possible .	A 50 percent chance of showers . Mostly cloudy , with a high near 53 . East wind around 6 mph becoming calm .	A chance of showers , mainly after 1pm . Cloudy , with a high near 53 . Calm wind becoming south around 6 mph . Chance of precipitation is 30 % .

6. **Best Results - Parameter Setting:** The model which gave the best results for us was a hierarchical encoder + decoder model. The architecture was,

   (a) Encoder embedding size: 256

   (b) Level 1 bidirectional LSTM encoder: 512 units

   (c) Dropout with 0.5 probability

   (d) Level 2 bidirectional LSTM encoder: 512 units

   (e) LSTM Decoder : 1024 units

   (f) Encoder embedding size: 256

   (g) Dense Layer : Decoder Vocabulary Size

   (h) Batch Size: 250

   (i) Optimizer: Adam

   (j) Learning Rate: 0.001

   (k) Epochs: 35

7. **Dimensions at each layer:**

Table 3: Dimensions at each layer

Layer	**Input Dimension**	**Output Dimension**
INEMBED	batchSize x maxSequenceLength	batchSize x maxSequenceLength x 256
ENCODER	batchSize x maxSequenceLength x 256	batchSize x 1024
ATTENTION	batchSize x maxSequence x 1024	batchSize x maxDecoderSequence x 1024
OUTEMBED	batchSize x maxDecoderSeqLength	batchSize x maxDecoderSeqLength x 256
DECODER	s0: batchSize x 1024 s1: batchSize x maxDecoderSeqLength x 256	batchSize x maxDecoderSeqLength x 1024
SOFTMAX	batchSize x maxDecoderSeqLength x 1024	batchSize x maxDecoderSeqLength x decoderVocabSize

8. **Unidirectional v/s Bidirectional LSTM- Better?** As can be seen from the learning curve, and also the BLEU scores, bidirectional LSTM was better for the task.

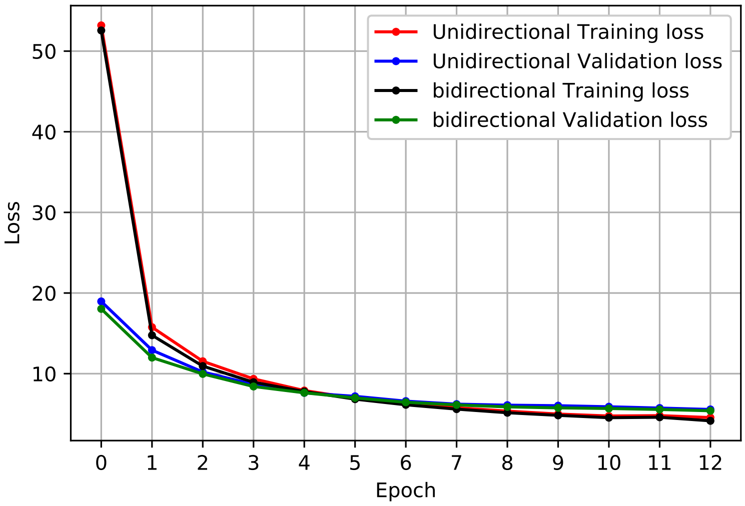

Figure 4: Learning curve for Unidirectional vs Bidirectional LSTM Encoder model.

9. **Basic v/s Hierarchical - Better?** Hierarchical model as described in (4), gave the best results.

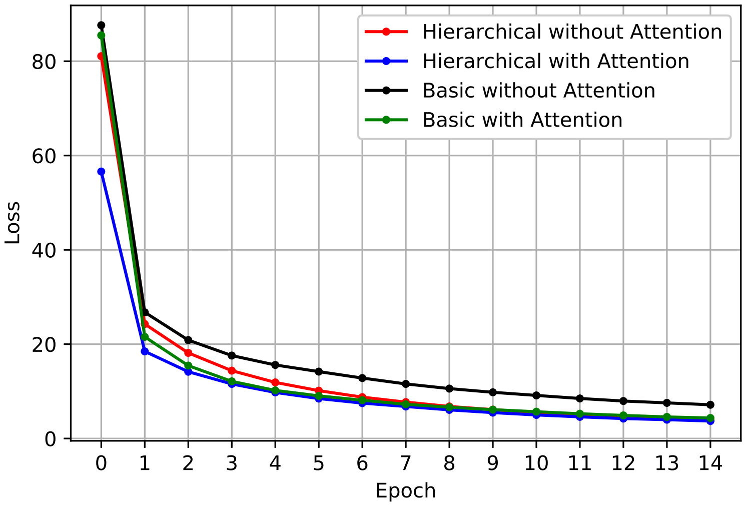

Figure 5: Learning curve for Basic Model vs Hierarchical Model

10. **Effect of attention** In case of basic model, the addition of attention provided to be beneficial. Whereas in the case of adding attention only on the last layer of the hierarchical encoder was not useful.

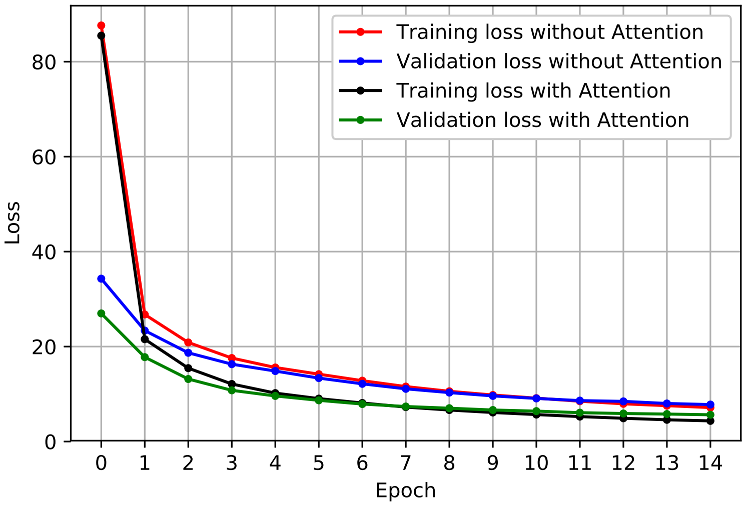

Figure 6: Learning curve Attention vs No Attention on basic model.

11. **Formulation of Attention Mechanism - Hierarchical:** Attention can be formulated in two stages, with the first level attention focusing on important words in the fields and the second level focusing on important fields in the sentence.
Encoder:
Layer 1:
for j = 1 to len(fieldWords)

    1) $x_{it} = W_e w_{it}, t \in [1, T_f]$
    2) $\overrightarrow{h_{it}} = \overrightarrow{LSTM}(x_{it}), t \in [1, T_f]$
    3) $\overleftarrow{h_{it}} = \overleftarrow{LSTM}(x_{it}), t \in [1, T_f]$
    4) $h_{it} = [\overrightarrow{h_{it}}, \overleftarrow{h_{it}}]$
    5) $emb_{it} = [h_{it}, W_e w_{it}^{j}]$

$u_{it} = \tanh(W_w emb_{it} + b_w)$
$\alpha_{it} = \frac{\exp(u_{it} u_w)}{\sum_t exp(u_{it} u_w)}$
$v_{it} = \sum_t \alpha_{it} h_{it}$

Layer 2:

$g_{it} = LSTM_2(v_{it}, S_{it-1}), t \in [1, T_{fw}]$
$c_{it} = \tanh(W1_w g_{it} + b1_w), t \in [1, T_{fw}]$

$\beta_{it} = \frac{\exp(u_{it}u_w)}{\sum_t exp(u_{it}u_w)}, t \in [1, T_{fw}]$
$d_{it} = \sum_t \beta_{it} h_{it}, t \in [1, T_{fw}]$

Decoder:
$z_{it} = W_{e2} d_{it}, t = 0$
$z_{it} = W_{e2} y_{it-1}, t \in [1, T']$
$(y_i, s_i) = softmax(LSTM(z_{it-1}, s_{it-1})), t \in [1, T']$

12. **Visualisation of Attention Layer** : After visualizing the attention weights,it was observed that the alignments were "one to many". We also see non-contiguous alignments .

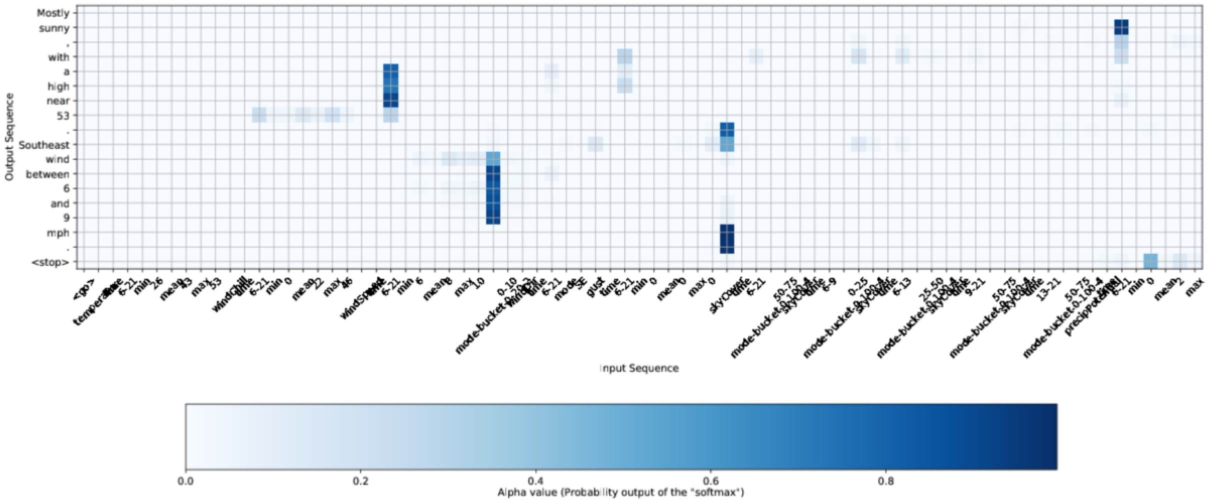

Figure 7: Visualisation of Attention layer for a training sequence

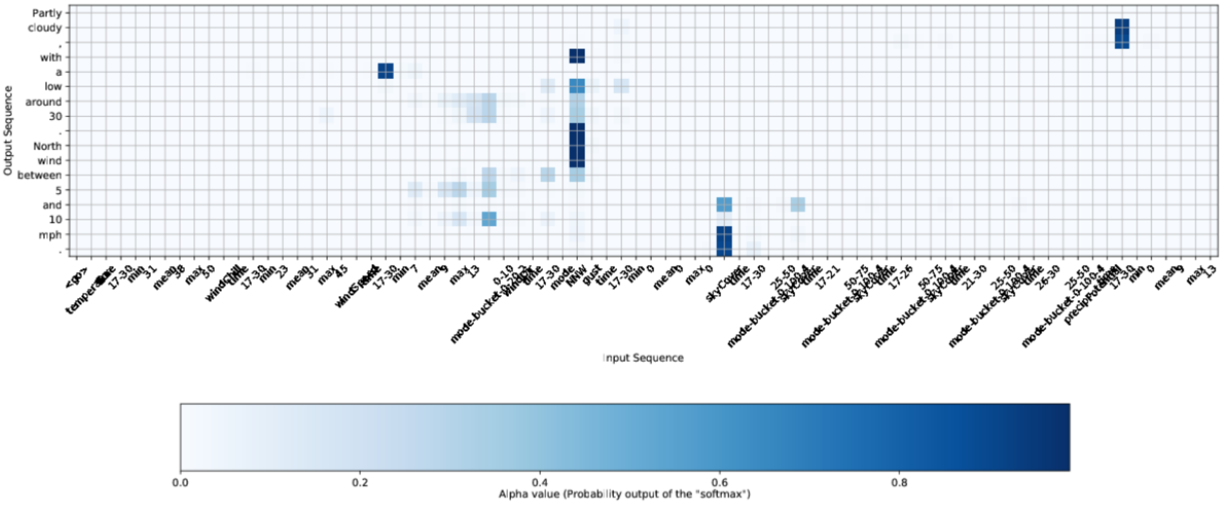

Figure 8: Visualisation of Attention layer for a validation sequence

13. **Effect of drop-out**

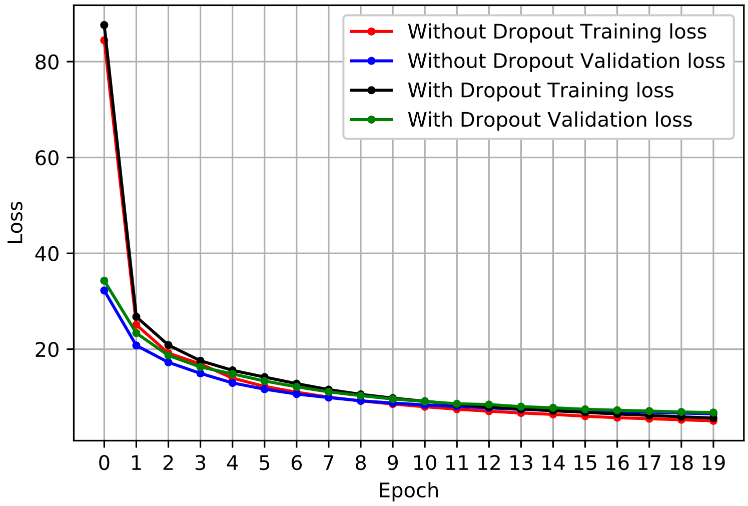

Figure 9: Learning curve Dropout vs No Dropout on basic model.

14. **Early stopping** : As could be seen from the learning curves, there was a steady decrease in the validation losses and early stopping was not necessary in many cases.

15. **Validation BLEU - Early stopping:** We implemented early stopping using Validation BLEU as well, but the scores kept on increasing at regular intervals as could be seen in the plot.

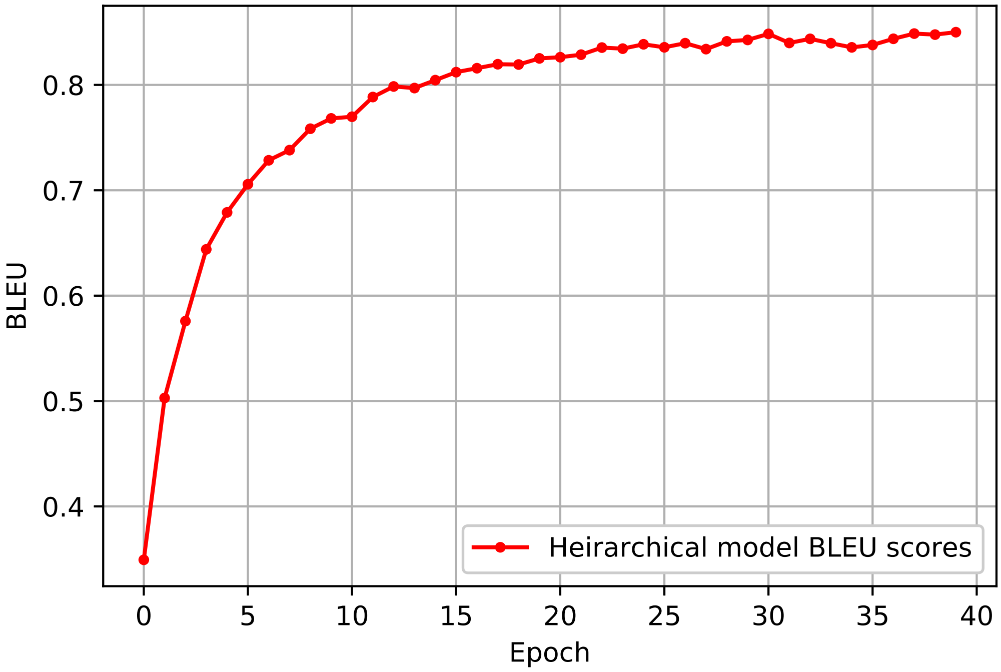

Figure 10: BLEU score vs Epochs

16. **Gradient Clipping :** We also tried clipping the gradients by a maximum gradient norm of 5. We were able to get a BLEU score of 0.827 on the test data.

17. **MultiLayer LSTM** : We tried creating a multilayer LSTM encoder with 2 layers and 256 units per layer. The encoder outputs were concatenated and fed to the decoder with architecture of the basic model. We were able to get a BLEU score of 0.6297 on the validation data.

18. **Beam Search**

Table 4: BLEU Scores of Greedy Decoder vs Beam Search Decoder on Validation data.

Model	**BLEU Scores**
Greedy Decoder	0.7037
Beam Search Decoder , N=5	0.7269

19. **Effect of Beam Width - Beam Search**

Table 5: Effect of Different Beam Width on a basic model with 128 units LSTM Cell.

Beam width	**BLEU Scores**
N=3	0.4339
N=5	0.4439
N=8	0.4725

Table 6: Effect of Different Beam Width on a Hierarchical model

Beam width	**BLEU Scores**
N=1	0.8299
N=3	0.8326
N=5	0.8315
N=9	0.8315

Table 7: Summaries generated by different beam widths

Input	**Beam Width = 1**	**Beam Width = 3**	**Beam Width = 5**
Sequence 16	Mostly cloudy , with a low around 45 . South wind between 3 and 7 mph .	Mostly cloudy , with a low around 45 . South wind between 3 and 7 mph .	Mostly cloudy , with a low around 47 . Calm wind becoming south southeast around 6 mph .
Sequence 7	Snow and areas before 10pm , then snow showers . Low around 8 . South southwest wind between 5 and 10 mph . Chance of precipitation is 80 % . New snow accumulation of 1 to 3 inches possible .	Flurries . Low around 7 . West wind 5 to 10 mph becoming south .	Snow , mainly before 10pm . Low around 49 . West wind around 7 mph . Chance of precipitation is 90 % . New rainfall amounts between a quarter and half of an inch possible .
Sequence 502	A chance of showers , mainly after 1pm . Cloudy , with a high near 53 . Calm wind becoming south around 6 mph . Chance of precipitation is 30 % .	A chance of showers , mainly after 1pm . Cloudy , with a high near 53 . Calm wind becoming south around 6 mph . Chance of precipitation is 30 % .	A chance of showers , mainly after 1pm . Cloudy , with a high near 53 . Calm wind becoming south around 6 mph . Chance of precipitation is 30 % .

EXTRA WORK DONE

- Hierarchical Encoder
- Beam Search
- MultiLayer LSTM
- Gradient Clipping
- Hierarchical Encoder + Attention