Abstract Text Summarization

Youngbin Kim

Outline

- Introduction
- Seq2Seq model
- Extensions & variants
- Experiment
- Future improvements

Automatic summarization

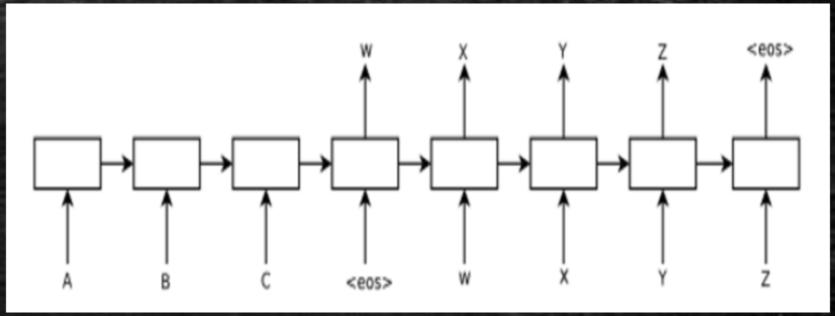
- def: *The process of shortening a text document with software, in order to create a summary with the major points of the original document*
- Application
  - Video summarization
  - Image Caption
  - Question answering system

Two ways to do text summarization

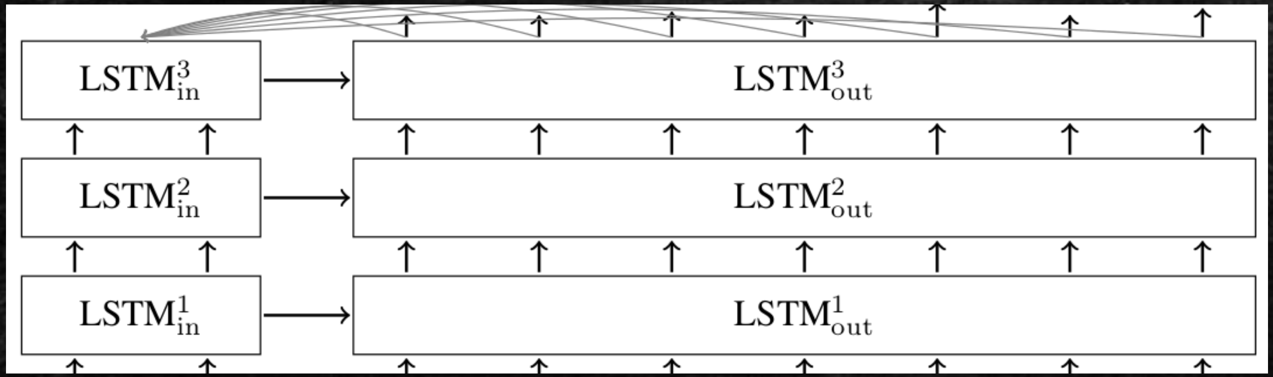
- Extractive summarization
  - Selecting subset of words from the source
  - Majority of text summarization
- Abstract summarization
  - Generate a summary based on semantic understanding of the text
  - Richer expressions, but more challenging (understanding of language model)

Related work

- Cho et al., 2014
  - Introduction of Sequence to Sequence model



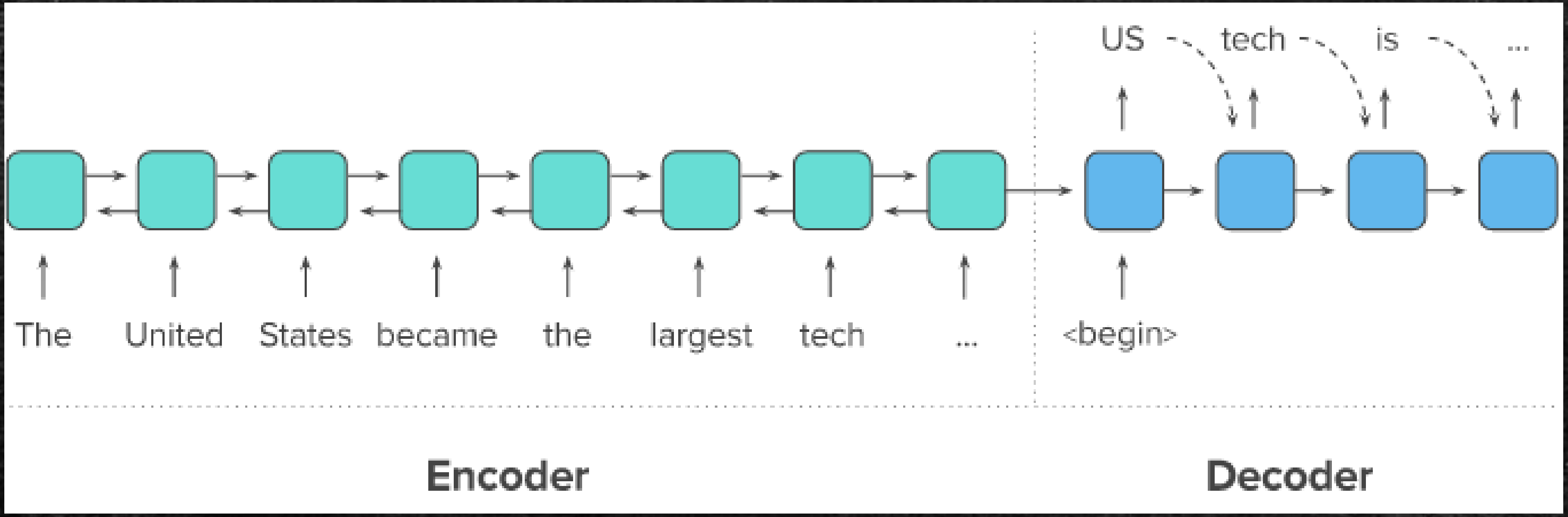
- Bahdanau et al., 2014
  - Attention mechanism



Related work

- Rush et al., 2015
  - Applied Seq2Seq to summarization
- Nallapati et al., 2016
  - Extended model with bidirectional encoder and generator-pointer decoder to deal with Out-Of-Vocabulary words

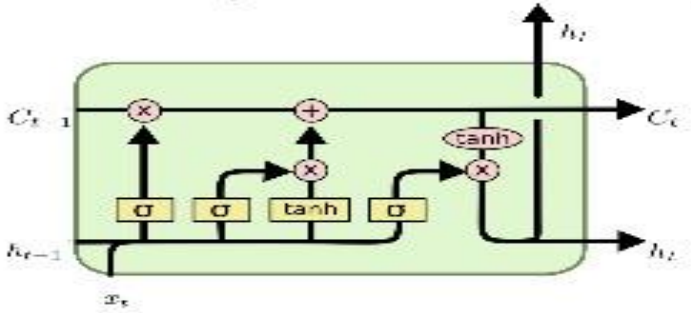
Basic Sequence to sequence model



LSTM / GRU

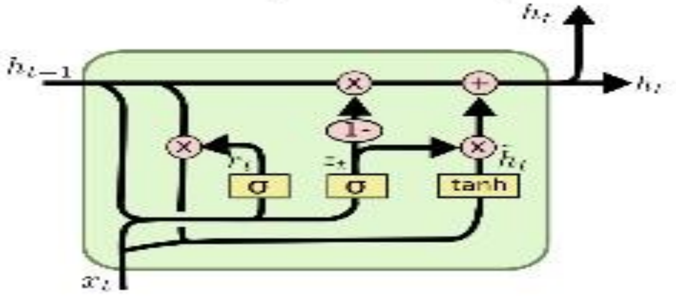
LSTM and GRU

- LSTM [Hochreiter&Schmidhuber97]



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right)$$
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$
$$o_t = \sigma\left(W_o \, [h_{t-1}, x_t] + b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

- GRU [Cho+14]



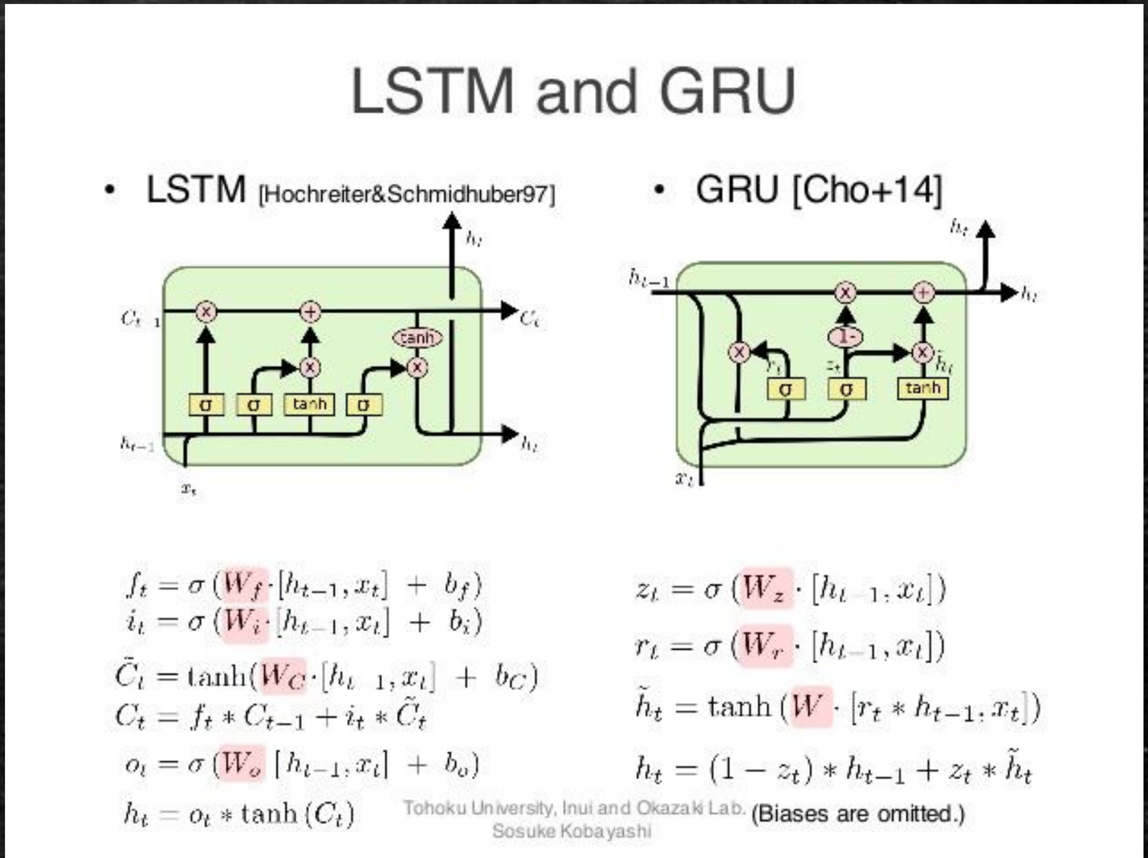
$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

(Biases are omitted.)

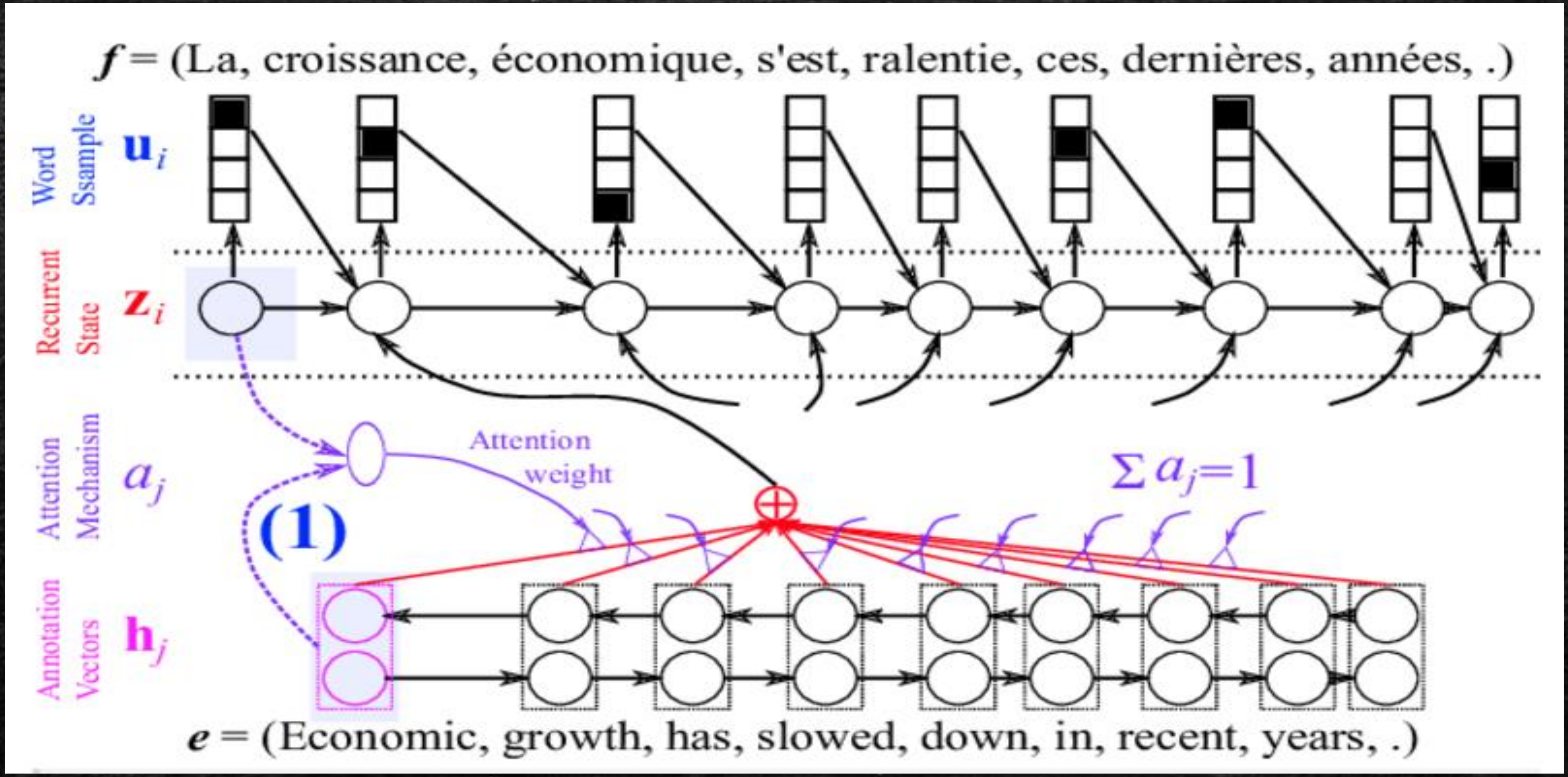
Tohoku University, Inui and Okazaki Lab.
Sosuke Kobayashi

LSTM / GRU

- Both prevent vanishing gradient problem
- GRUs train faster
- LSTMs outperform in tasks requiring modeling long-distance relations.



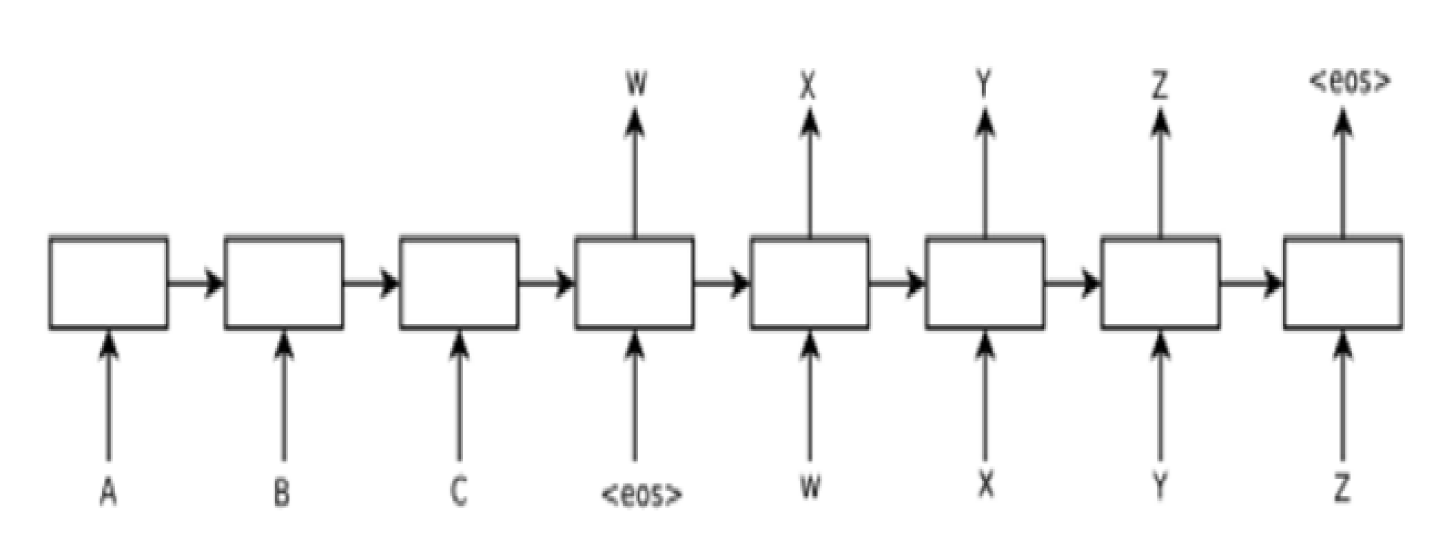
Attention



f = (La, croissance, économique, s'est, ralentie, ces, dernières, années, .)
Word Ssample
u_i
Recurrent State
z_i
Attention Mechanism
a_j
Attention weight
Σ a_j=1
(1)
Annotation Vectors
h_j
e = (Economic, growth, has, slowed, down, in, recent, years, .)

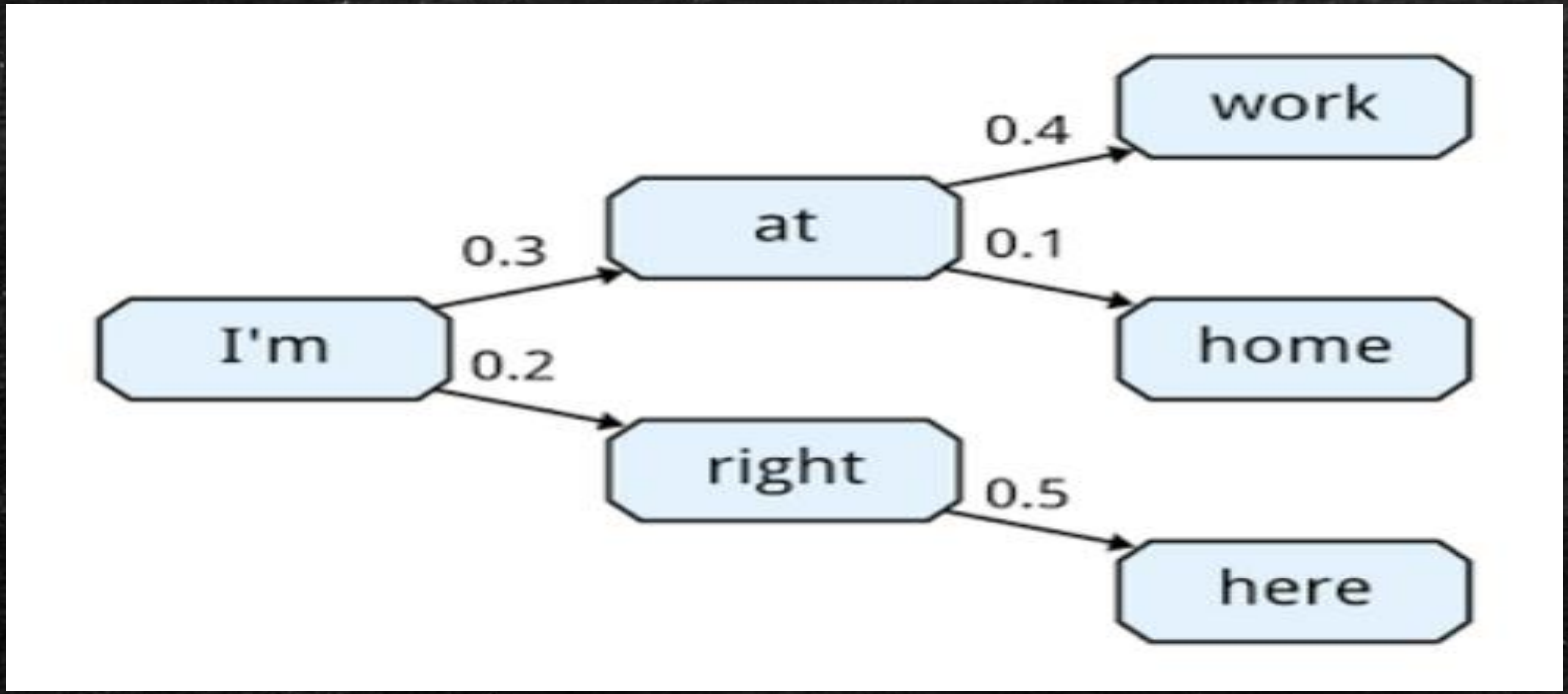
Decoder

- Cross Entropy Loss for each generated word $-\sum_{i=1}^{n}\sum_{i=c}^{C} t_{ic} \cdot log(y_{ic})$
- During training, each word in an actual summary is fed in
- Multiplied by weight vectors (0 if </S> else 1)



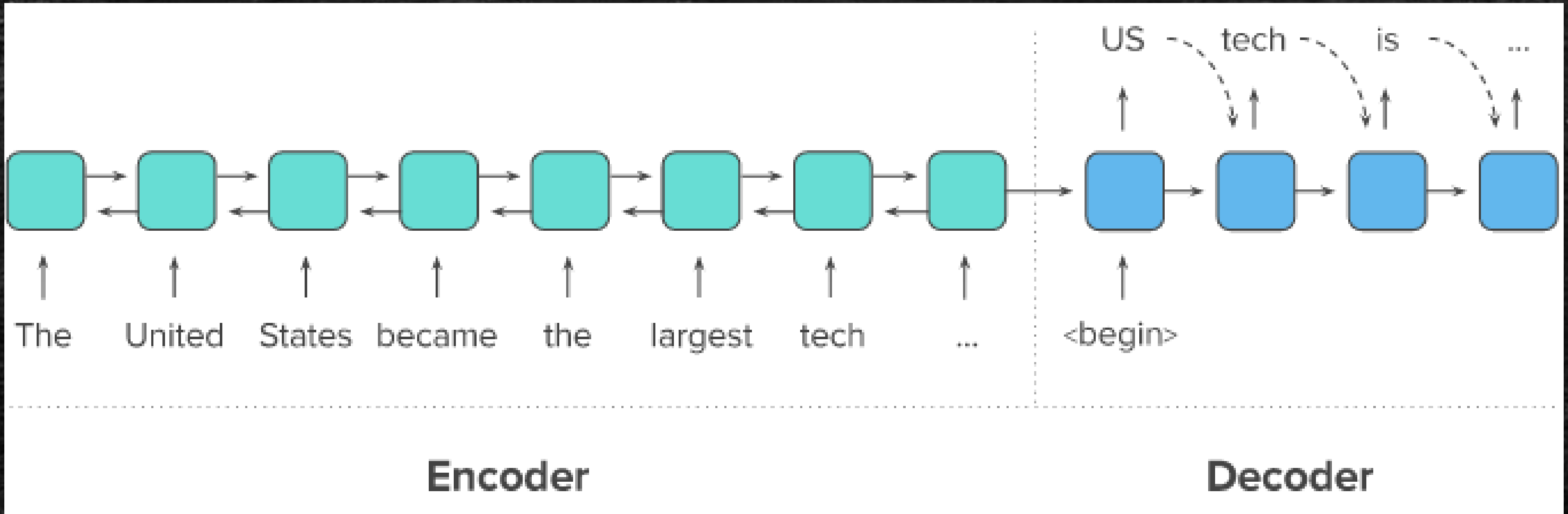
Decoder

- Beam search used for decoding (e.g beam size = 4)



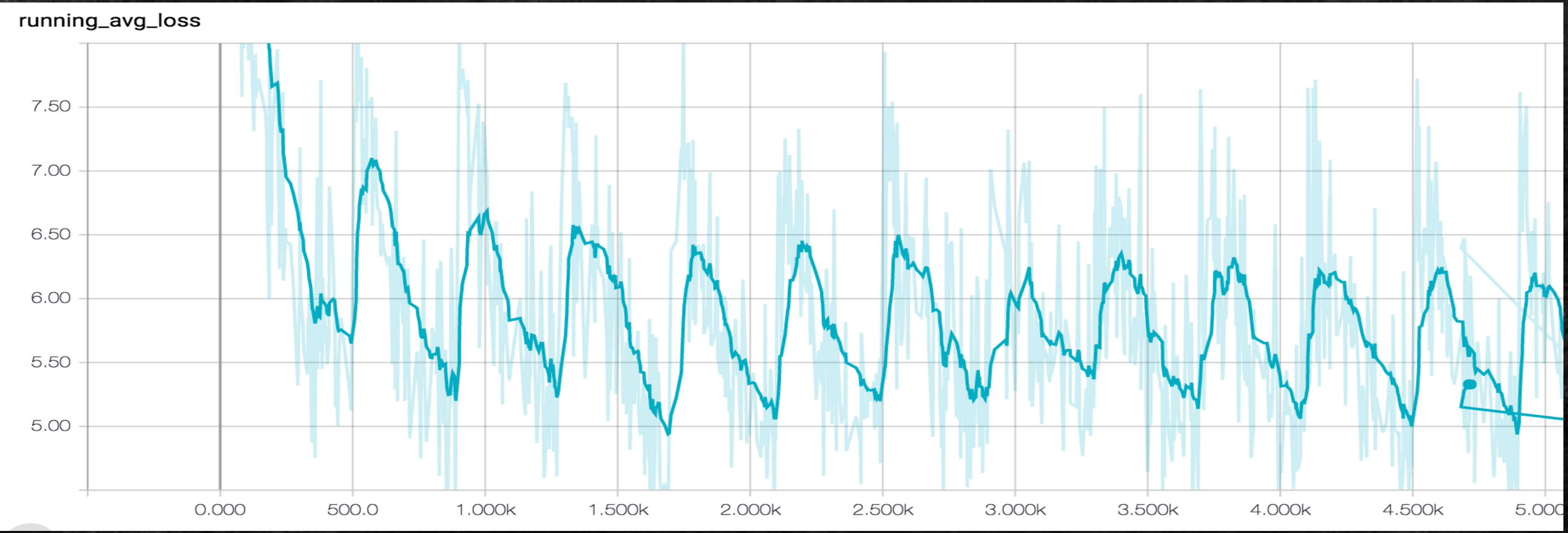
Decoder

- Greedy search during eval



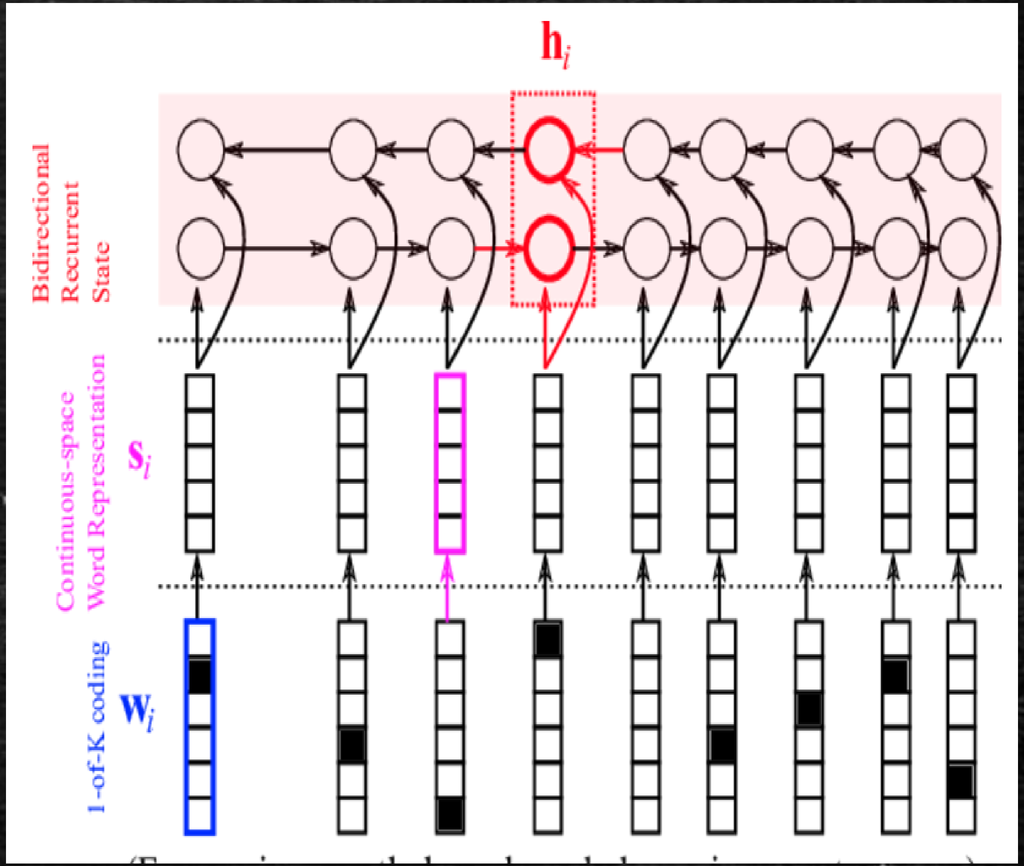
Initial result

- Training loss – problem?

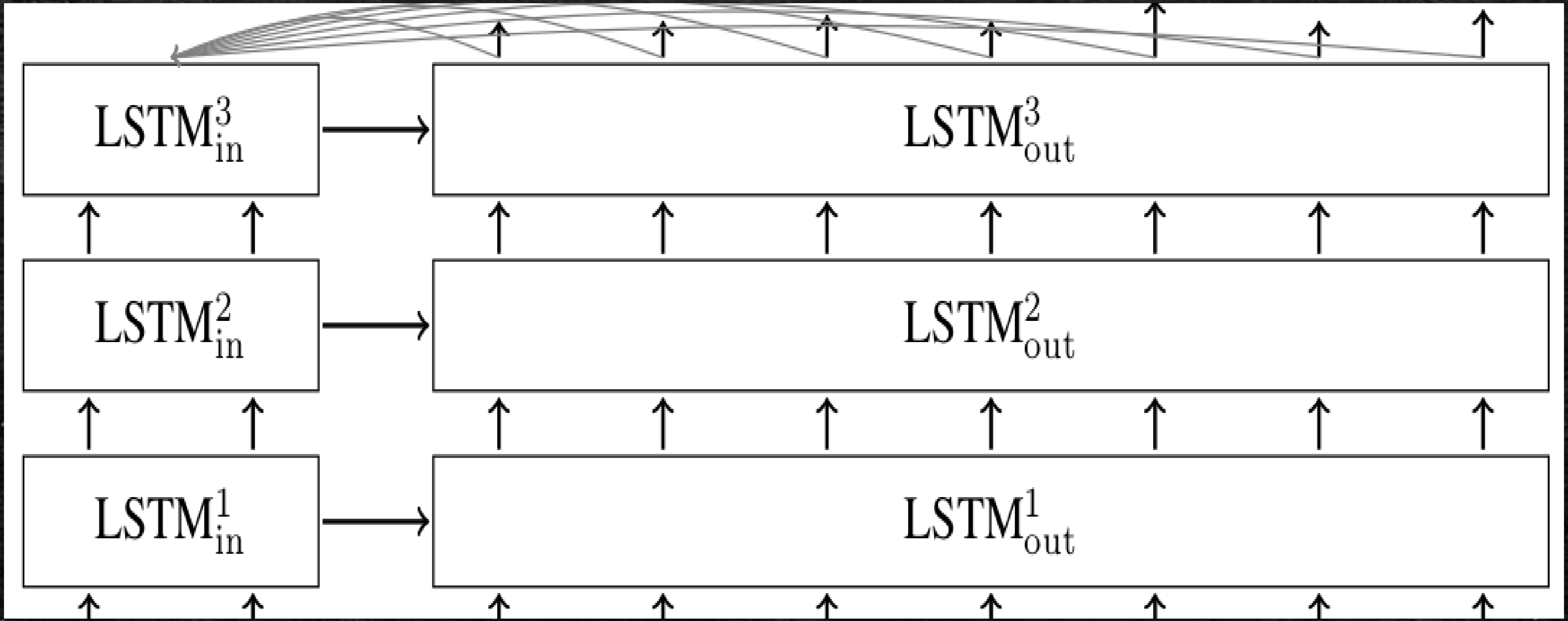


Bidirectional encoder

- Make predictions based on future words by having the RNN model read through the corpus backwards



Go deeper!

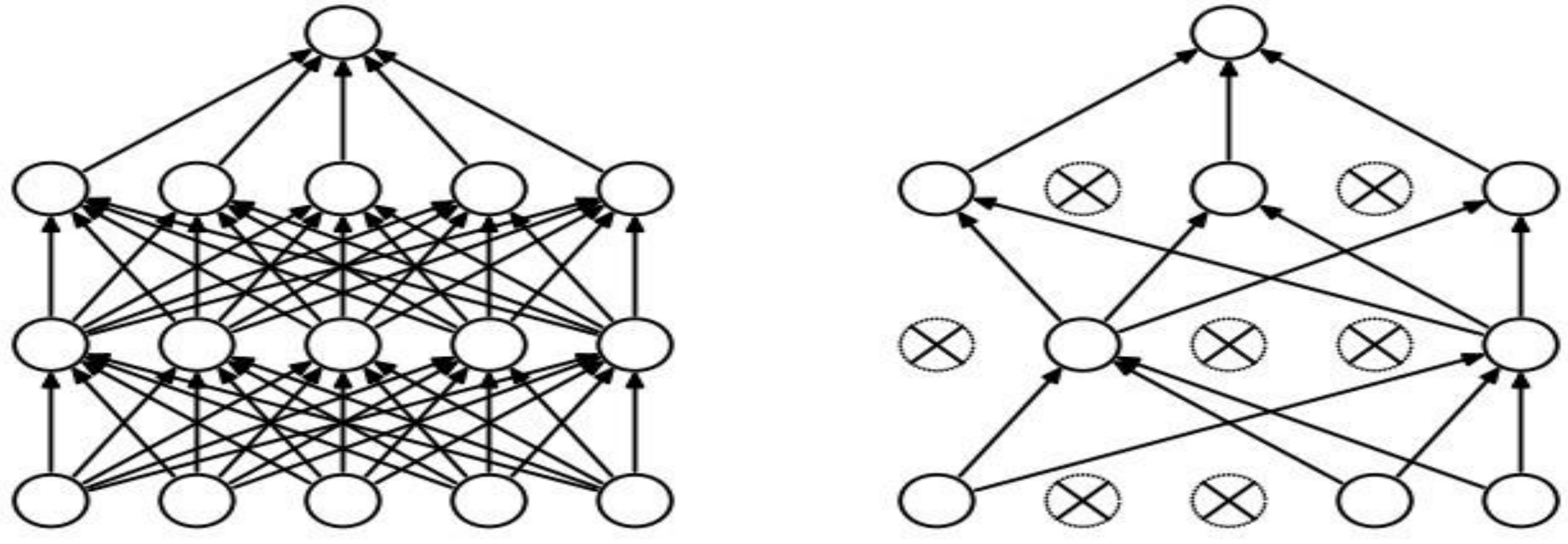


Adam Optimizer

- Adaptive Learning rate
- Faster convergence
- Learns much better !

Overfitting - Dropout

- To prevent Network from overfitting..
- While training, dropout is implemented by only keeping a neuron active with some probability pp (a hyperparameter), or setting it to zero otherwise



(a) Standard Neural Net (b) After applying dropout.

Overfitting - Batch Normalization

- Provide any layer in a Neural Network with inputs that are zero mean/unit variance
- Slower and ineffective? - need more investigation

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma\widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

Overfitting - L2 Regularization

- Use weights, not biases
- Add to the training loss

$$\text{L2:} \quad \frac{\lambda}{2}\|\mathbf{w}\|^2 = \frac{\lambda}{2}\sum_{j=1}^{m} w_j^2$$

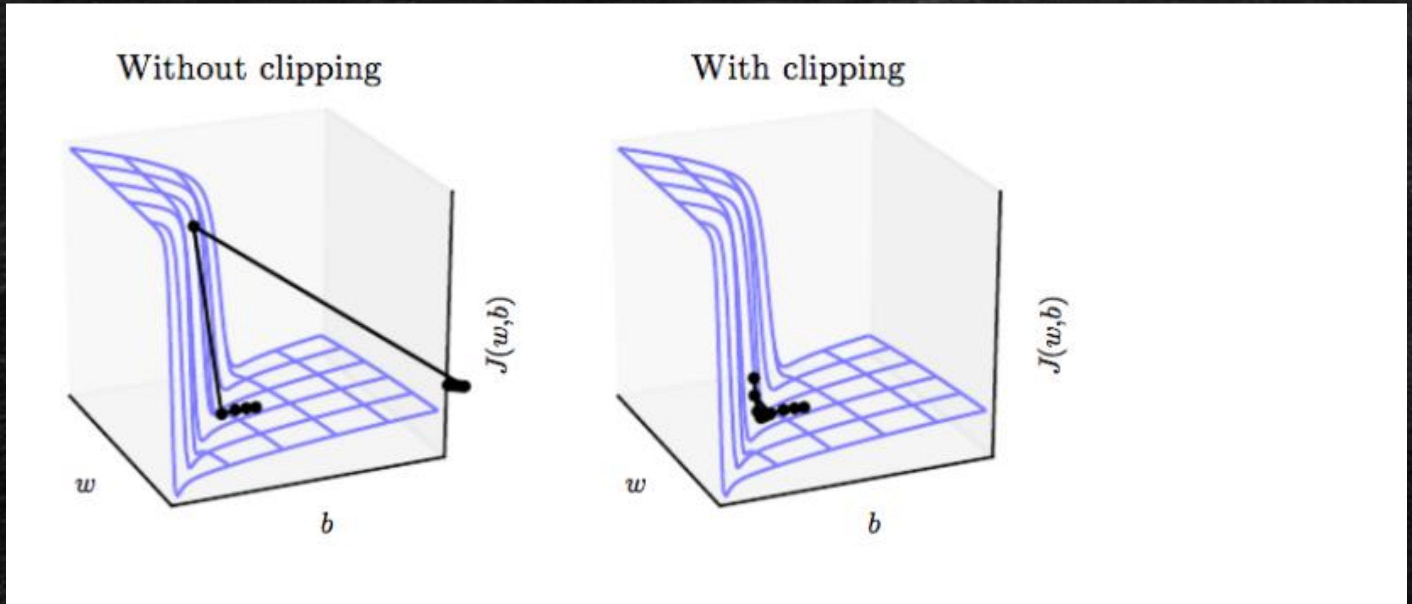
Gradient Clipping

- Deal with exploding gradients
- Clipping with global norm

```
t_list[i] * clip_norm / max(global_norm, clip_norm)
```

where:

```
global_norm = sqrt(sum([l2norm(t)**2 for t in t_list]))
```



— Goodfellow et al., *Deep Learning*

Sampled softmax and output projection

- Batch-size x num_decoder_symbols
- Out of memory error
- To handle large output vocabulary
- To decode from it, we need to keep track of the output projection

Hyperparameter Search

- # of layers
- # of hidden units in rnn cells
- Learning rate
- Epsilon (for AdamOptimizer)
- Embedding dimension
- Lambda for L2 regularization
- BiRNN vs RNN for encoder
- Attention for decoder

Extension 1 - Pretrained GloVe

- unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus
- Pre-trained with Wikipedia 2014 + Gigaword 5 (6B tokens, 400K vocab, uncased, 50d, 100d, 200d, & 300d vectors

$$J(\theta) = \frac{1}{2} \sum_{i,j=1}^{W} f(P_{ij})(u_i^T v_j - \log P_{ij})^2$$

Extension 2 – tf-idf

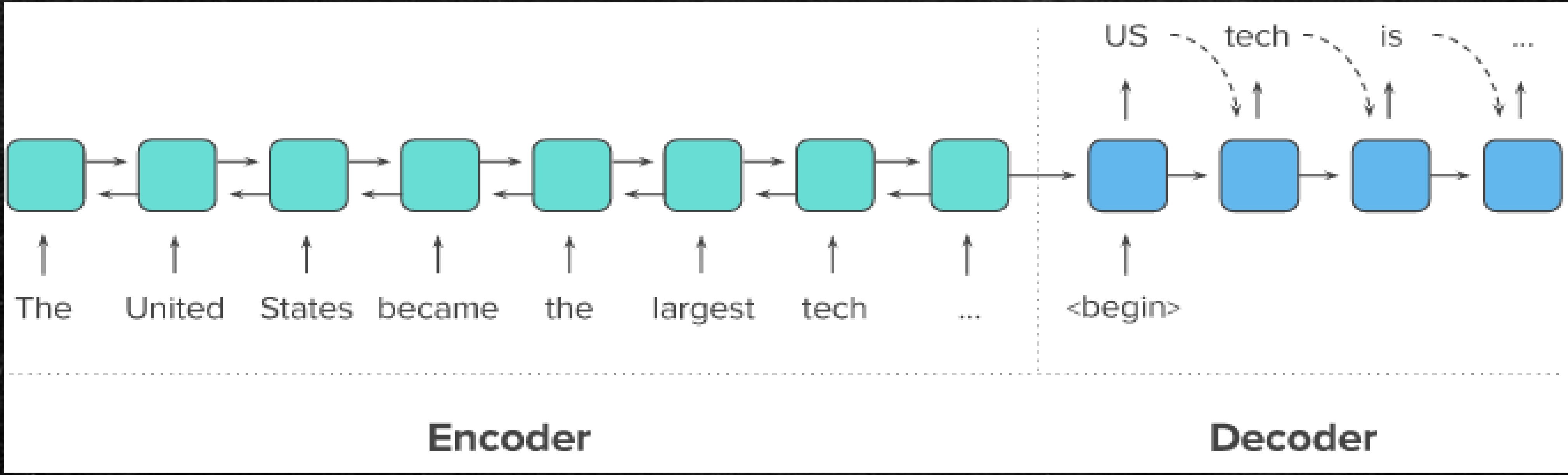
- During preprocessing, compute idf scores
- tf-idf(d, t) = tf(t) * idf(d, t)
- idf(d, t) = log [n / df(d, t)] + 1
- During training, compute tf and get tf-idf score
- Concatenate this to word repretentation

Extension 3 - Pos tagging

- Part-of-speech tagging (nouns, verbs, adjectives, adverbs)
- Can help in summarization (Pronoun)
- Each tagging converted to the vector of some dimension (e.g 5)

Reversing the input

- Reduce the short-term dependency
- Deals with exploding gradients problem
- Also effective in this task!



Neural Bag-Of-Word Encoder

- Each word -> word vector (Embedding matrix)
- Sentence -> Average of word vectors
- Much faster but ..

Dataset

- CNN/ DailyMail dataset
  - ~300k (90k CNN, 200k DailyMail)
  - 4 hand-crafted summaries
  - Split : Training - 0.9, Dev - 0.05, Test - 0.05
  - Problem ?
- DUC 2004
  - 500 docs
  - 4 summaries to compare
  - Frequently used for testing for summarization task
- Signal Media One-Million News Articles
  - 1M news articles with headlines

Experiment - Preprocessing

- Compute and store Idf scores
- Create binary files
  - Extract text and title
  - Lowercasing, Clean, tokenize
  - Each number to #
  - Convert to serialized tf.train.Example Protobuf
- Create Vocabulary
  - 200K most frequent + { <S>, {/s}, <PAD>, <UNK> }
  - During training, load this vocab and create embedding matrix
    - Unknown randomly initialized [-0.25, 0.25]
  - Percentage of words in GloVe
    - Tokenize 25%
    - Lower casing 65%
    - Cleaning string 68% (e.g Special Characters. Quotes)
    - <UNK>

Experiment

- Shuffled mini-batch
- Hyperparameters
  - Batch: 32
  - # of layers: 4
  - Embedding dimension: 256
  - Learning rate: 0.1
  - Epsilon: 0.01
  - Lambda: 0.0001
- Decoder – Beam 4
- AWS EC2 P2, Tesla K80 GPU + Nvidia GTX1080

Evaluation

- Loss
- ROUGE-1, ROUGE-2, ROUGE-L
- Qualitative Analysis

ROUGE

- Recall-Oriented Understudy for Gisting Evaluation

ROUGE-N

$$= \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

- ROUGE-L: Longest Common Subsequence (LCS)[4] based statistics.
- Good metric for summarization ?

Result

	LOSS	ROUGE-1
Baseline 1	8.652	4.268
Baseline 2	5.864	8.654
Extension 1 - Pretrained	5.396	11.035
Variant 1 – Average Encoder	7.095	6.132

Qualitative Analysis – Early Stage

Text	langley , arkansas (cnn) -- one person remained missing monday from last week 's flash flood at an arkansas campground that left ## dead , and `` there 's still a possibility there could be others , '' gov . mike beebe told cnn . rescuers found a ##th body over the weekend about half to three-quarters of a mile downstream from the campground , arkansas state police capt .
Headline	new : `` there could be others '' as search for flood victims goes on , governor says
Generated summary	new : the the of the the

Qualitative Analysis - 1

Text	(cnn) -- americans should n't expect to see the ##,### u.s. troops in afghanistan come home any time soon , no matter who is declared the victor in the country 's presidential election . u.s. marines patrol near herat , afghanistan , in july . in fact , the pentagon is planning to add #,### troops by the end of the year .
Headline	new : president obama says u.s. goal remains defeating al qaeda , its allies
Generated summary	opcw : rick obama president obama reveals goal on obama qaeda through report says

Qualitative Analysis - 2

Text	north korea held a huge rally friday in the center of its capital , pyongyang , to celebrate the launch of a long-range rocket this week that put a satellite in orbit and provoked international condemnation . a special broadcast on state-run television showed crowds of soldiers and civilians standing in neat ranks , clapping and cheering as officials made congratulatory speeches praising the regime 's ruling dynasty ...
Headline	new : north korean state media say satellite is to monitor weather
Generated summary	the koreans officer says officer the rocket for service

Qualitative Analysis - 3

Text	if you 're a big `` the hunger games '' fan like i am , you were probably crazy excited to watch the movie 's first trailer , which recently debuted . for those of you who are n't as familiar with the popular trilogy of books turned major motion pictures , here 's a quick synopsis . the story takes place during an unidentified time in the future in a post-apocalyptic nation called panem .
Headline	learn how to survive a stressful work environment from 'the hunger games '
Generated summary	adam sandler plays character donny berger with conviction

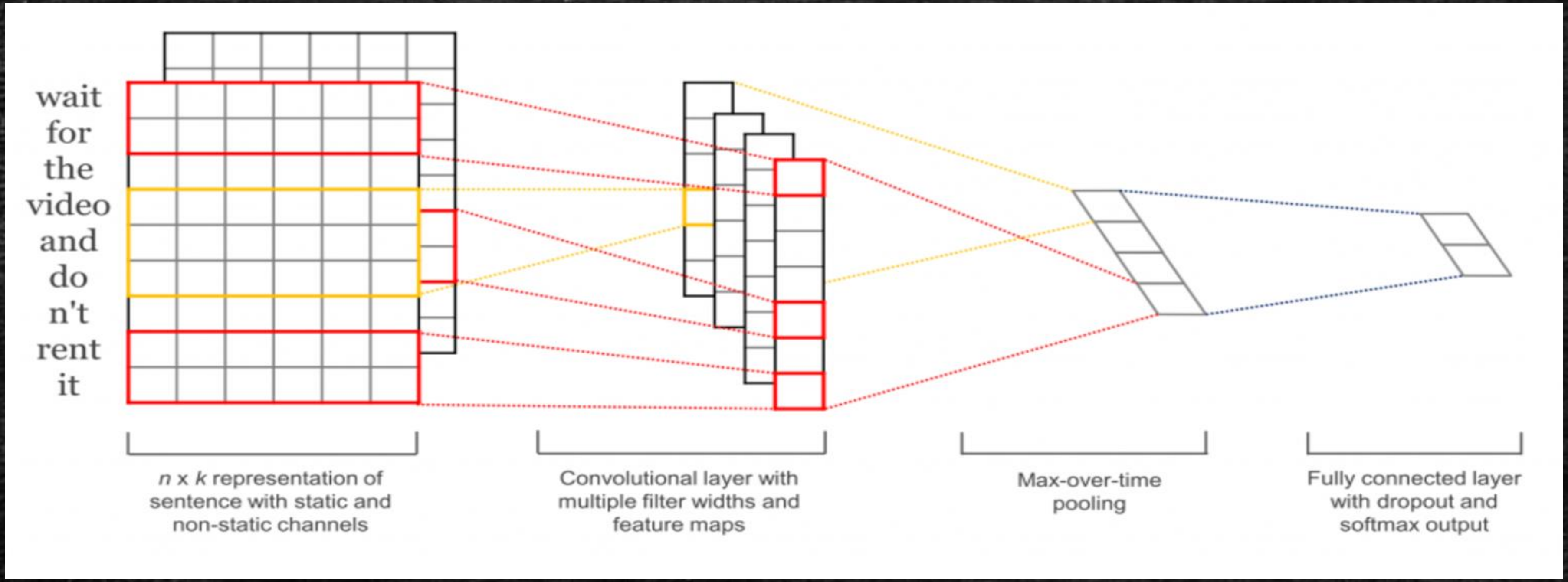
Qualitative Analysis

- Problem - still overfitting
  - More data ?

Future work

- Experiment on larger dataset (E.g Signal Media One-Million News Articles)
- Improve the quality of generated summary
  - Only use stopwords for summary ?
  - Other models ?

Future work – CNN encoder



Or Facebook's conv seq2seq?

Future work – Skip connections

- Even deeper network with Residual Learning
- Google NMT 4 -> 8 layers

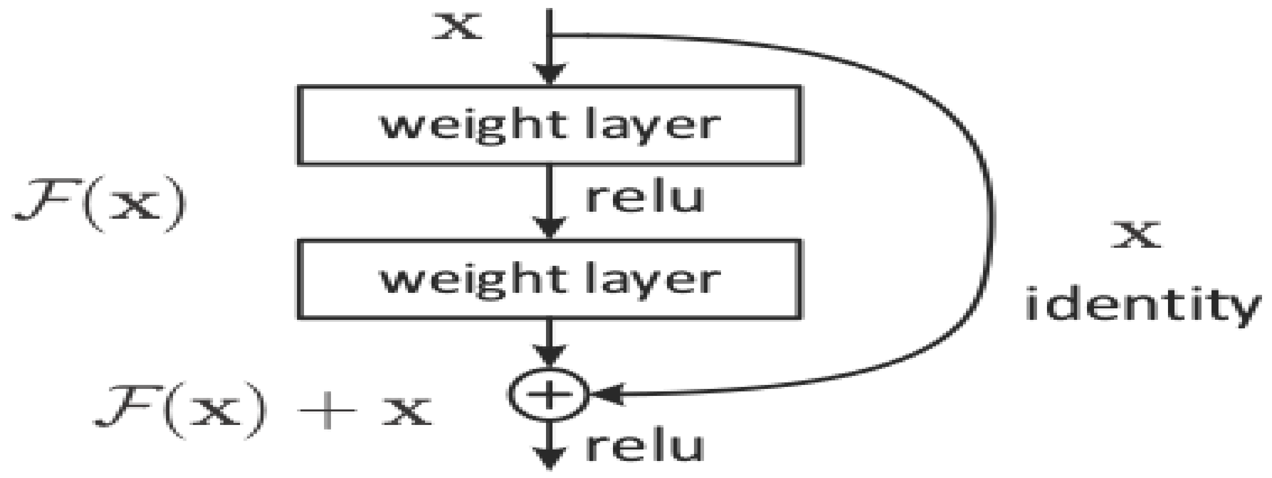
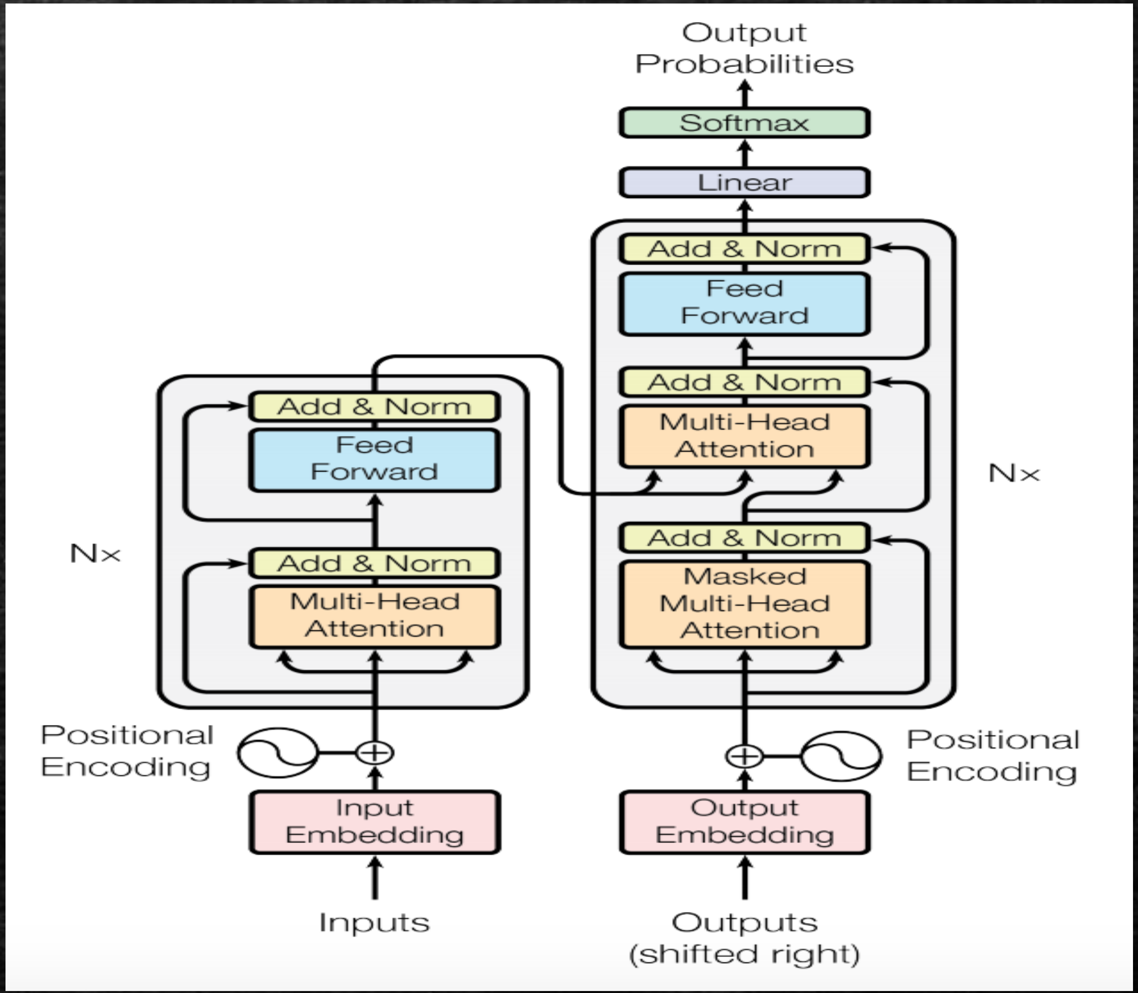


Figure 2. Residual learning: a building block.

Future work – Attention is All You Need

- State-of-the-art MT (June, 2017)
- Very fast



Future work – RL with Different Metric ?

- ROGUE score is likely not to be the best metric for Summarization
- Sentence Similarity networks to compare?

Reference

- https://arxiv.org/pdf/1706.03762.pdf
- http://www.wildml.com/2015/12/implementing-a-cnn-for-text-classification-in-tensorflow/
- https://cs224d.stanford.edu/lecture_notes/LectureNotes4.pdf
- http://peterroelants.github.io/posts/neural_network_implementation_intermezzo02/
- http://83.212.103.151/~mkalochristianakis/techNotes/ipromo/rougen5.pdf
- https://www.tensorflow.org/tutorials/seq2seq
- https://kratzert.github.io/2016/02/12/understanding-the-gradient-flow-through-the-batch-normalization-layer.html
- https://devblogs.nvidia.com/parallelforall/introduction-neural-machine-translation-gpus-part-3/
- https://en.wikipedia.org/wiki/Automatic_summarization
- http://web.stanford.edu/class/cs224n/lectures/cs224n-2017-lecture3.pdf
- https://nlp.stanford.edu/projects/glove/