Internship presentation

JAMES PAK

JULY 28TH, 2017

Content

- Data Preprocessing
  - Normalization
  - N-gram
  - Tokenization
  - Word Embedding
- DNNs
  - FCN
  - CNN
  - RNN
  - LSTM
  - LSTM + Attention Layer
- Applications
  - Sentiment Analysis
  - Movie Review Credibility Grader
  - Call-log Anomaly Detector
- Q&A
- Reference

Data Preprocessing

Normalization

- Trick1 - Lowercase Processing
  - Before: “A Pizza Can Get To Your House Faster Than An Ambulance.”
  - After: “a pizza can get to your house faster than an ambulance.”

- Trick2 - Special Character Processing
  - Before: “[lol] there’re handicap parking places in front of a skating rink.”
  - After: “there ’re handicap parking places in front of a skating rink . ”

- Trick3 - Regular expression Processing
  - Before: “My number is 213-999,1434 and email is mingunpa@usc.edu”
  - After: “My number is spt1 and email is spt2”

Normalization

- Trick4 - Num2word(Nico's script)
  - Before: "I went bobsleighing the other day, killed 250 bobs."
  - After: "I went bobsleighing the other day, killed two hundred fifty bobs."

- Trick5 - Stopword Processing
  - Before: "Time flies so fast, it's August"
  - After: "Time flies fast, August"

- Trick6 - Correcting typo
  - Before: "I like playing a guiter"
  - After: "I like playing a guitar"

Normalization

- Trick7 - Acronym Processing
  - Before: “I am from the U.S.”
  - After: “I am from the usa”

N-gram

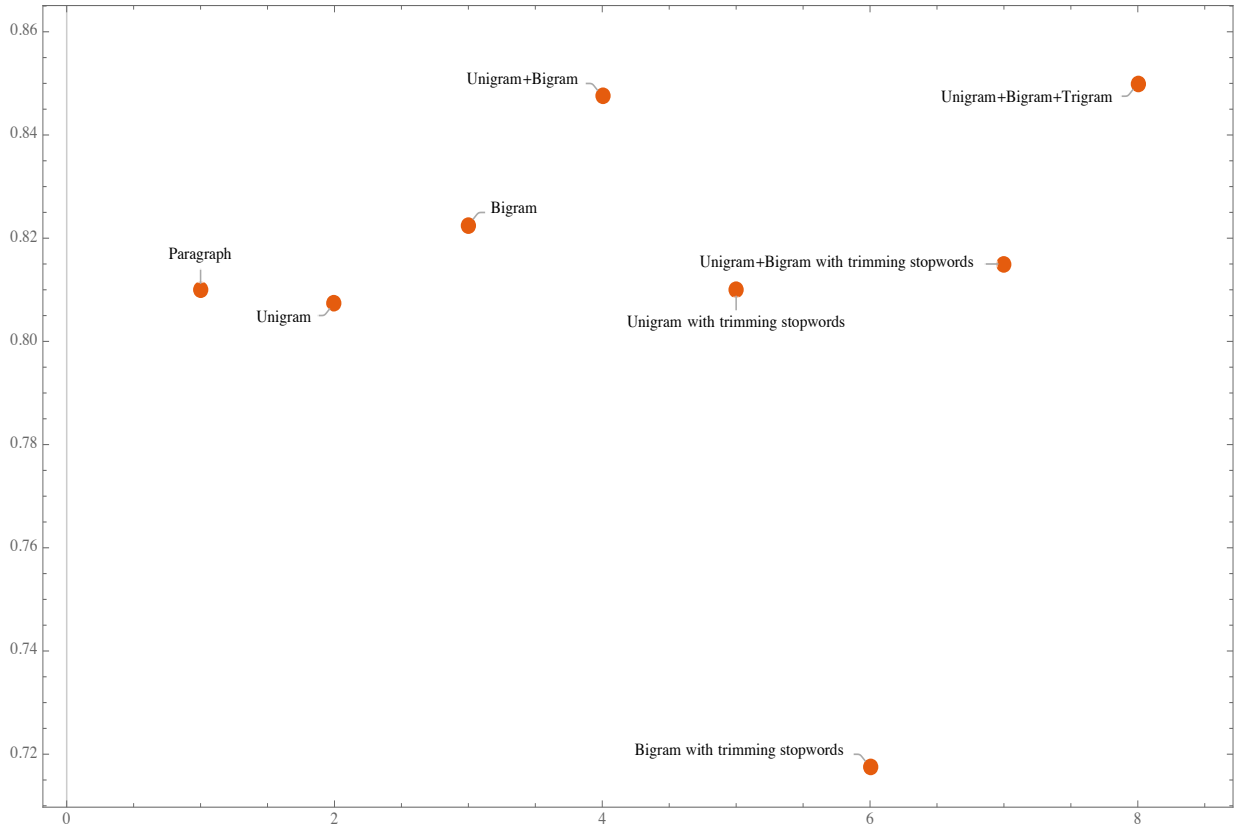
- Unigram
  - Before: “I am James from Korea”
  - After: [“I”, “am”, “James”, “from”, “Korea”]
- Bigram
  - Before: “I am James from Korea”
  - After: [“I am”, “am James”, “James from”, “from Korea”]
- Trigram
  - Before: “I am James from Korea”
  - After: [“I am James”, “am James from”, “James from Korea”]

N-gram

- Unigram + Bigram
  - Before: “I am James from Korea”
  - After: [“I”, “am”, “James”, “from”, “Korea”, “I am”, “am James”, “James from”, “from Korea”]
- Unigram + Bigram + Trigram
  - Before: “I am James from Korea”
  - After: [“I”, “am”, “James”, “from”, “Korea”, “I am”, “am James”, “James from”, “from Korea”, “I am James”, “am James from”, “James from Korea”]

Which one do you think is the best?

N-gram



N-gram test result based on Markov Classifier with 2,000 samples

Tokenization

- Definition
  - Tokenization is a way to split text into tokens. These tokens could be paragraphs, sentences, or individual words.

- Base Tokens
  - SOS: Start of Sentence
  - EOS: End of Sentence
  - UNK: Unknown
  - PAD: Padding

- Unigram + Bigram
  - Before: [“I”, “am”, “James”, “from”, “Korea”, “I am”, “am James”, “James from”, “from Korea”]
  - After: [11, 13, 2341, 54, 156, 34245, 34135, 23412, 24534]

Word Embedding

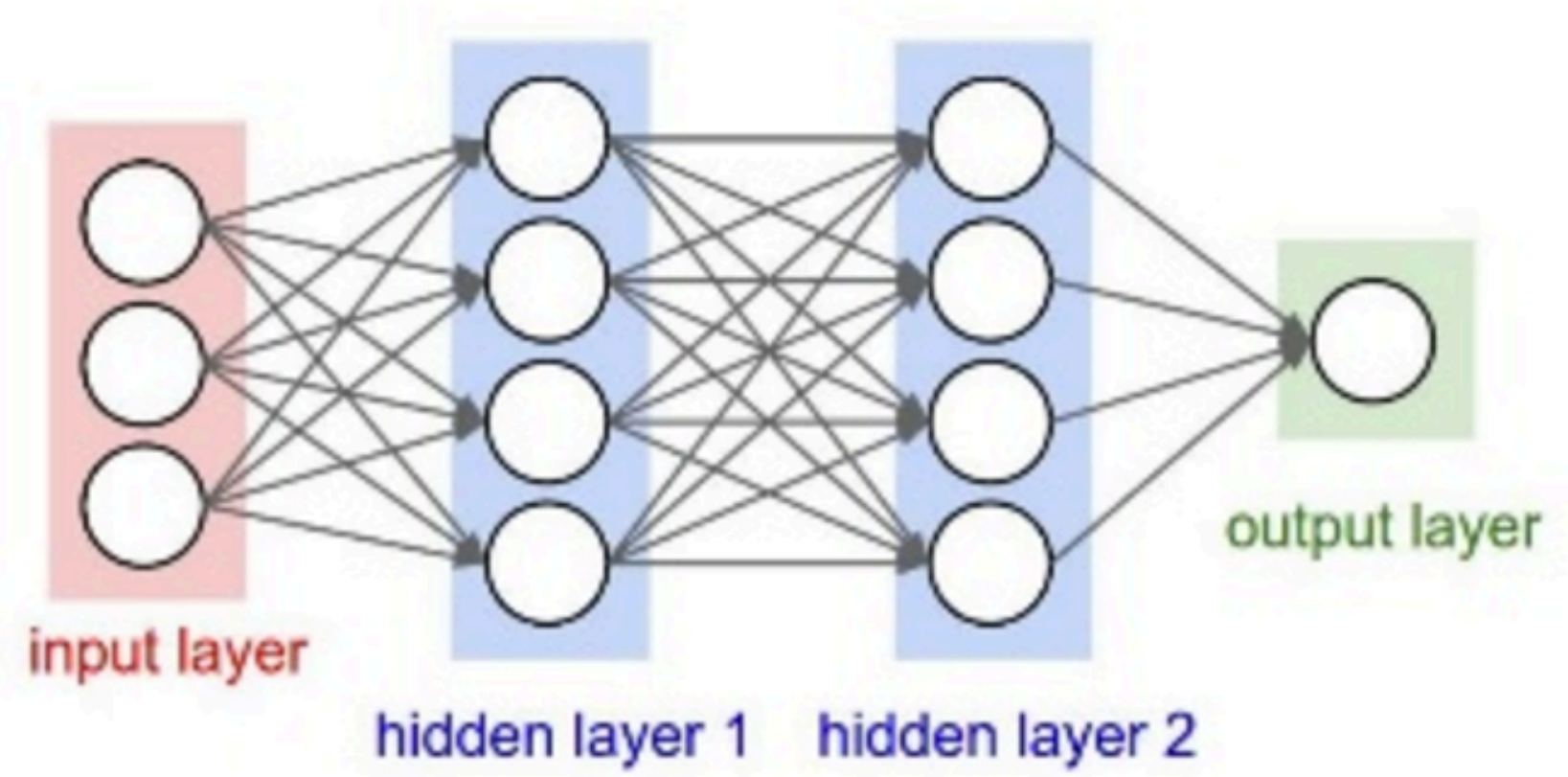
- Definition
  - Word embedding is the feature learning techniques in NLP where words or phrases from the vocabulary are mapped to vectors of real numbers.

- Unigram with 5 dimension
  - Before: [“I”, “am”, “James”]
  - After: [[“0.4927”, “-0.3927”, “0.4827”, “-0.2721”, “0.4827”,], [“0.0927”, “0.1927”, “0.7227”, “-0.5926”, “-0.9927”,], [“-0.3827”, “0.3422”, “0.1927”, “-0.3452”, “0.3495”,],]

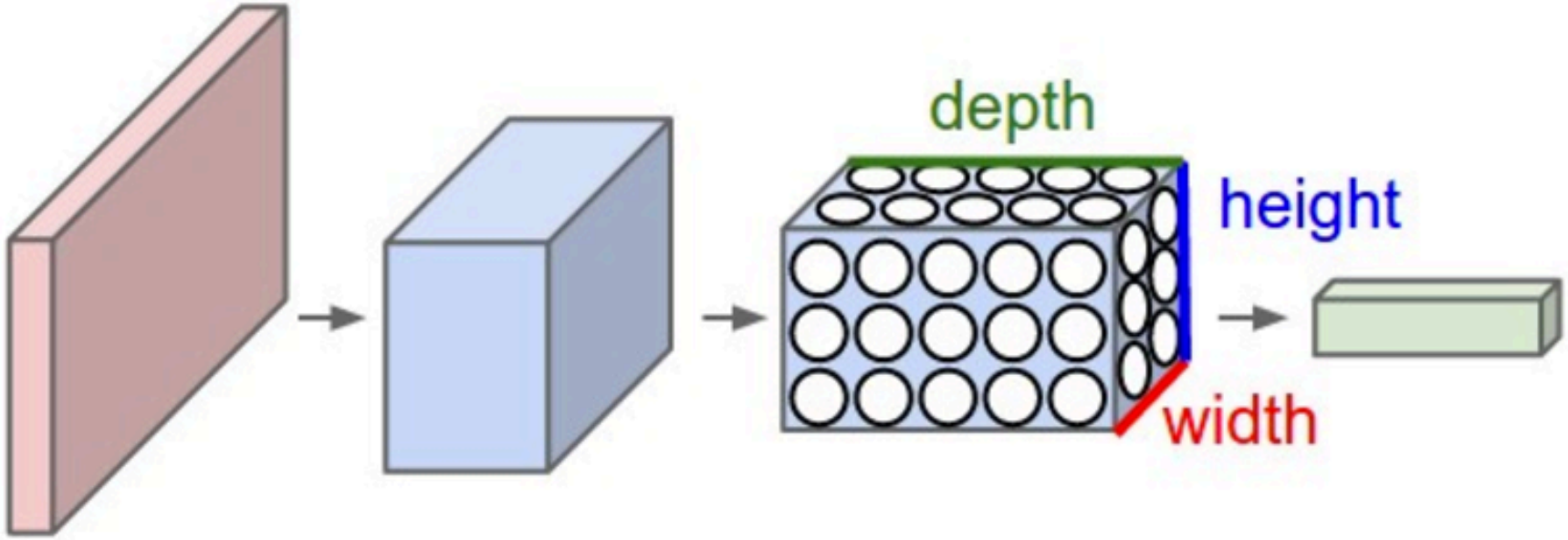
- Models
  - C&W model
  - Word2Vec
  - GloVe

DNNs

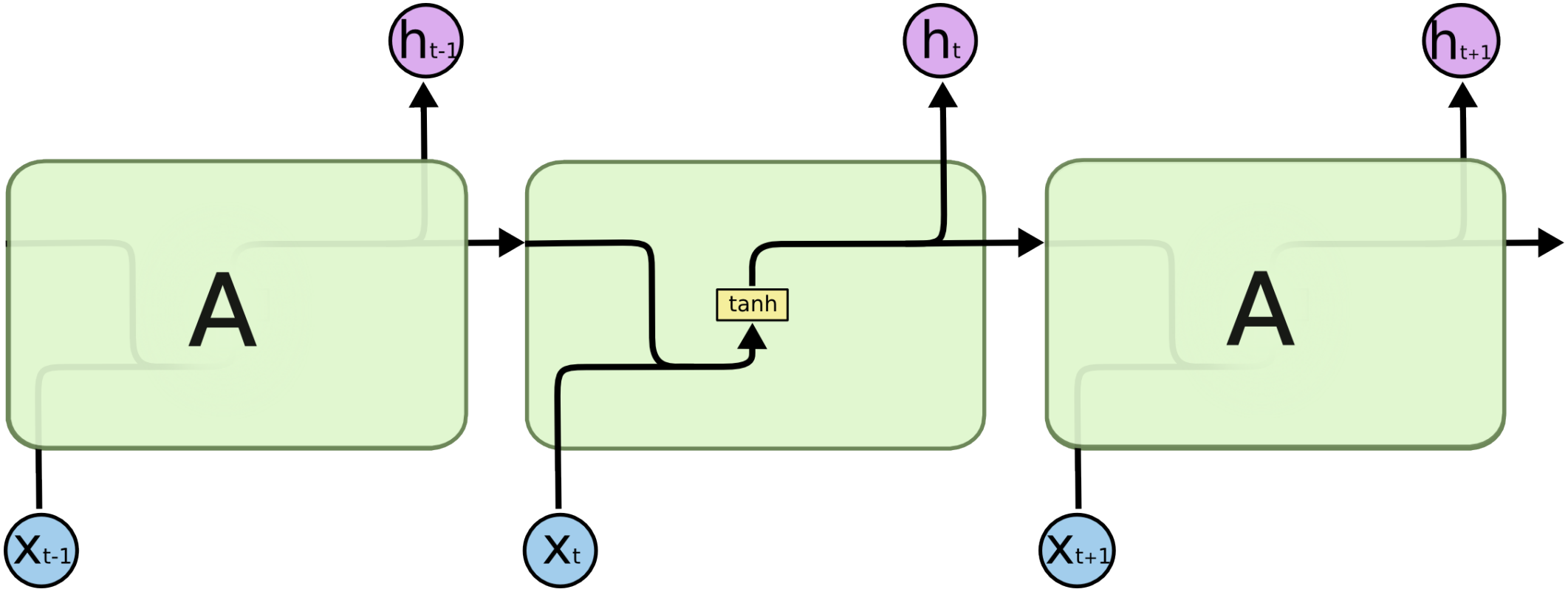
Fully Convolutional Network(FCN)



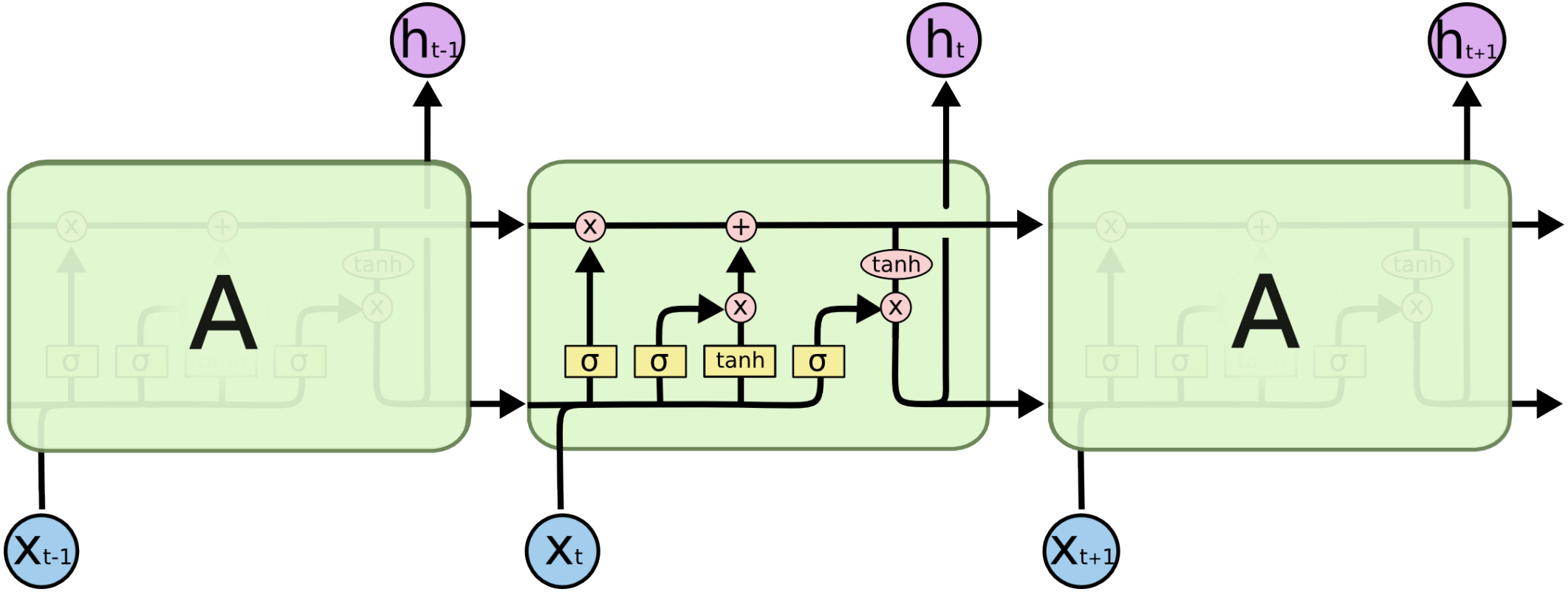
Convolutional Neural Network(CNN)



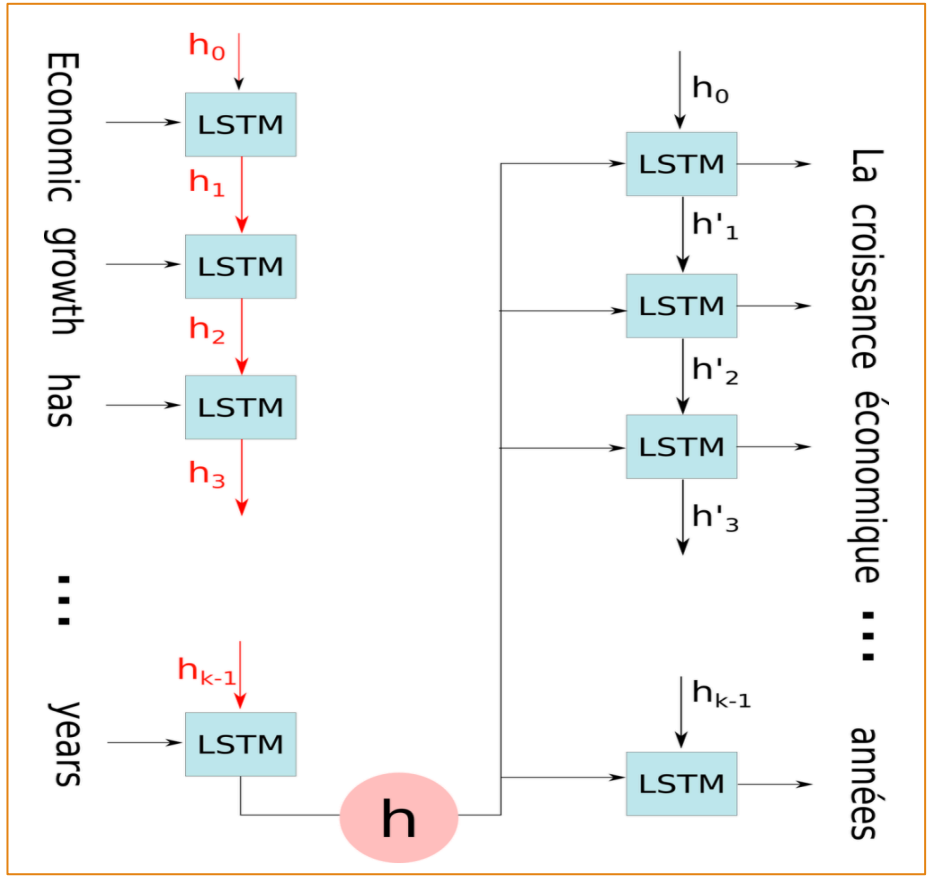
Recurrent Neural Network(RNN)

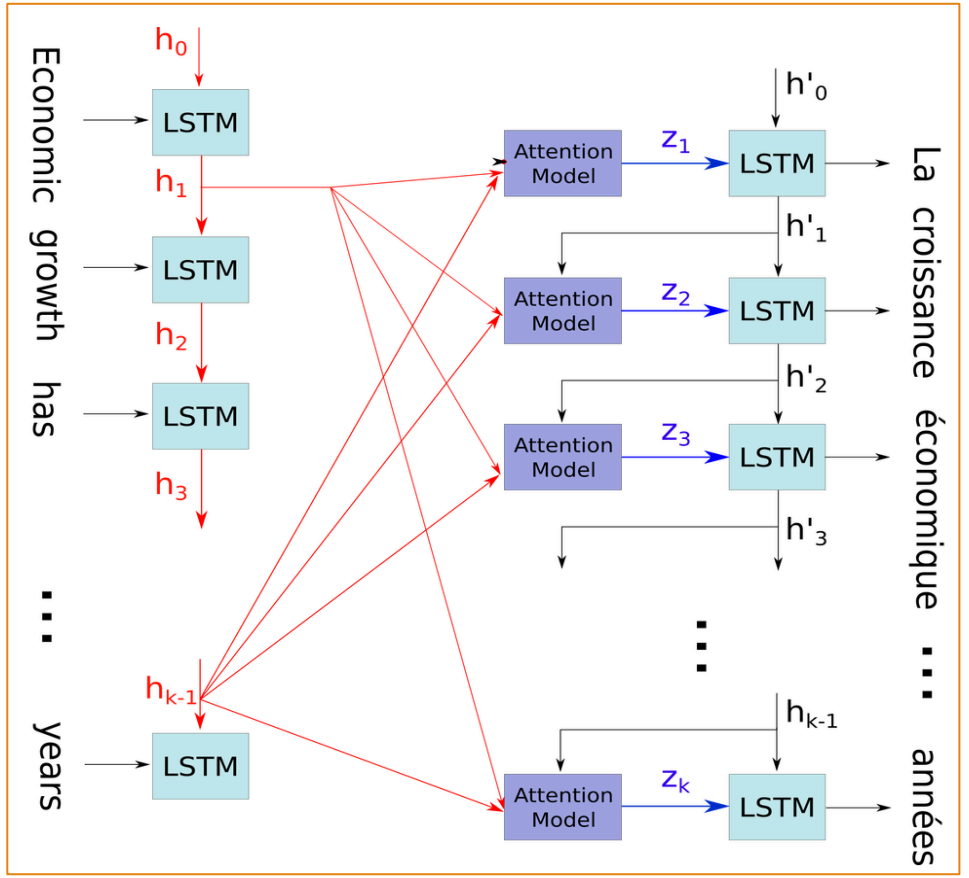


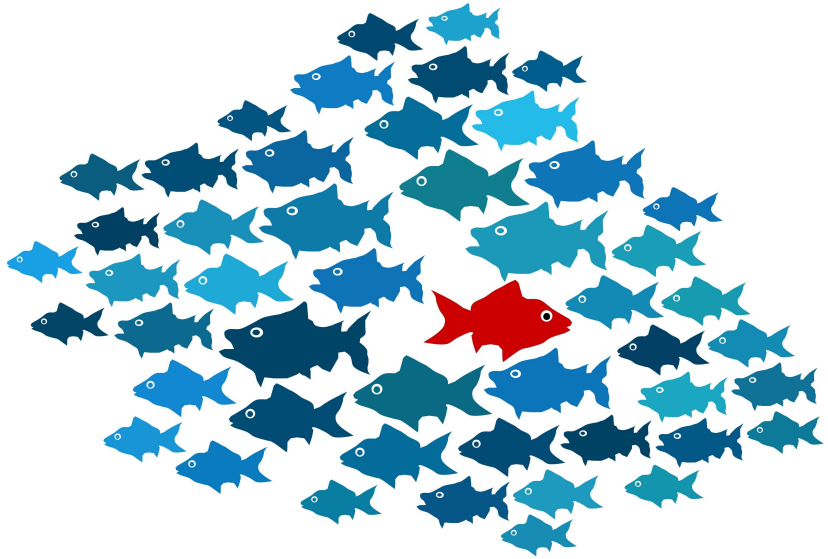
Long Short-Term Memory(LSTM)



Attention Trick







Applications

Sentiment Analysizer

- Data Set Information
  - IMDB Dataset
  - 25,000 Train set(12,500 positive, 12,500 negative)
  - 25,000 Test set(12,500 positive, 12,500 negative)

Sentiment Analysizer

- Data Preprocessing

	Use
Lowercase	O
Special Character	O
Regular expression Processing	X
Num2word	X
Stopword	X
Correcting typo	X
Acronym	X

Sentiment Analysizer

- Training Environment

	Use
Max length	500
Max Feature	20000
Pre-trained Model	GloVe
Word Embedding Dimension	100
Loss	binary crossentropy
Optimizer	adam

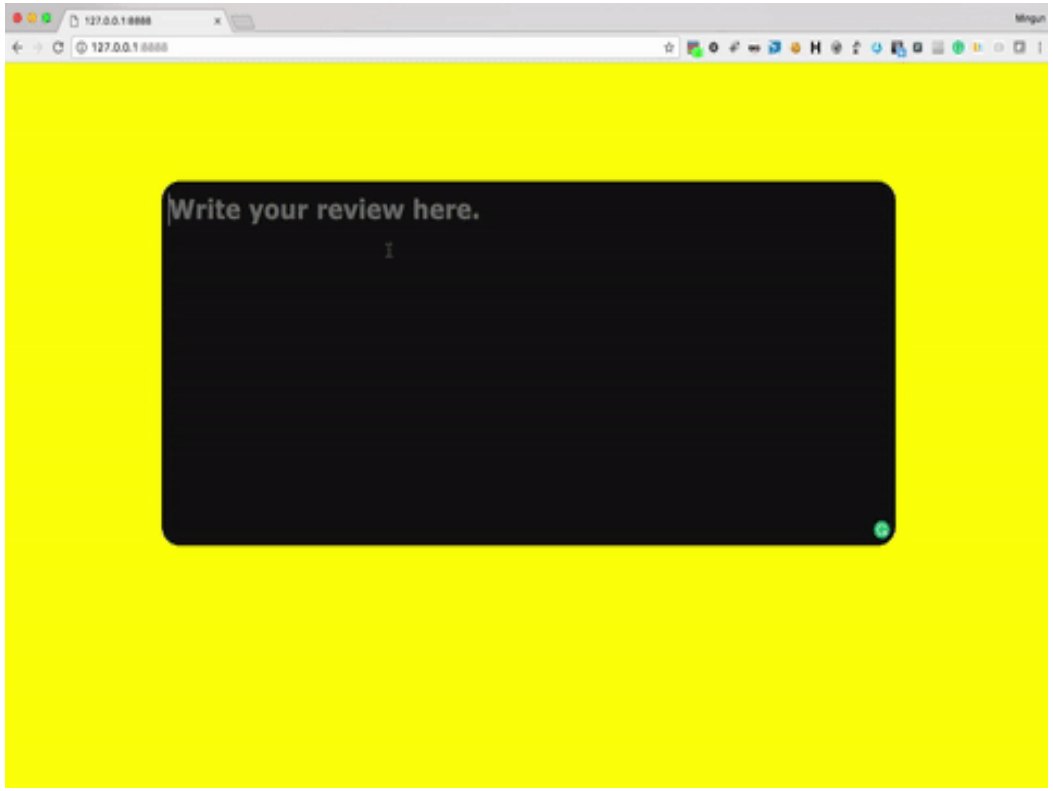
Sentiment Analysizer

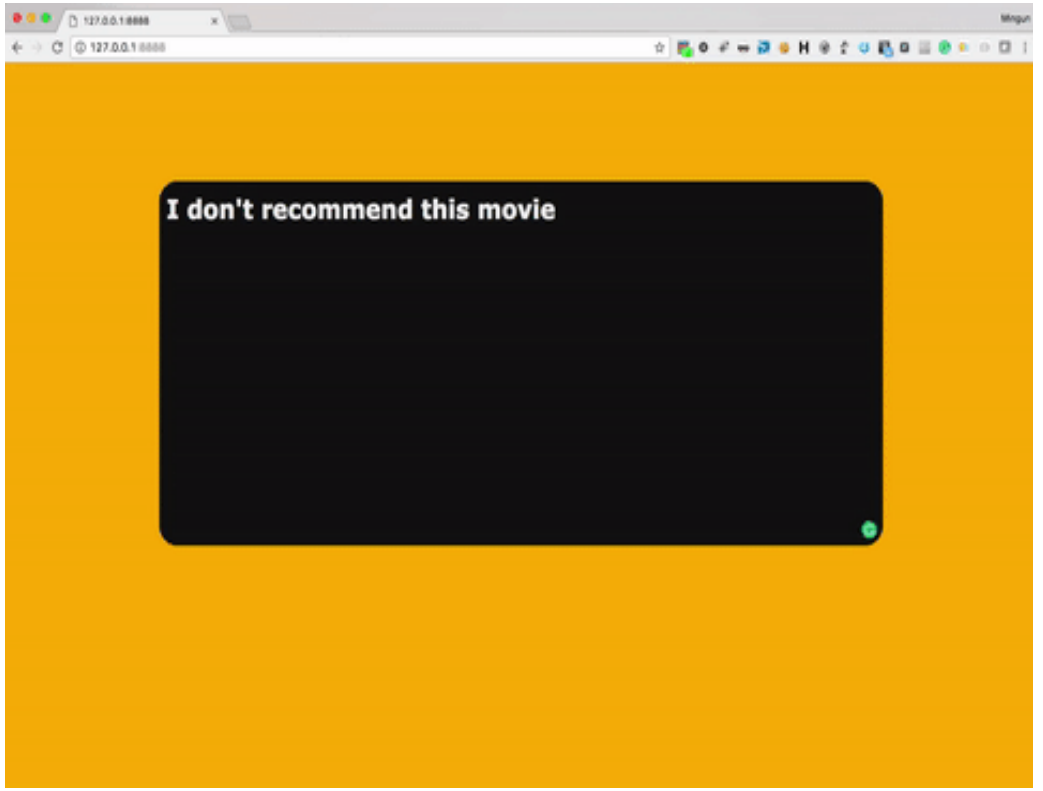
- Model Comparison

	DenseNet	DenseNet	CNN	CNN	LSTM	LSTM+CNN
n-gram	Unigram	Unigram+Bigram	Unigram	Unigram+Bigram	Unigram	Unigram
Hidden Size	250	250	250	250	128	128, 250
filter	-	-	250	250	-	250
Kernal Size	-	-	3	3	-	3
Pooling Layer	-	-	Avg	Avg	-	Avg
Dropout	-	-	0.2	0.2	0.2	0.2
Accuracy	84.5	87.8	88.5	90.8	80.4	83.9

Sentiment Analysizer

- Application Demo(http://localhost:8888)





Movie Review Credibility Grader

- Data Set Information
  - IMDB Dataset
  - 100,000 unlabeled sentences related to movie review

Movie Review Credibility Grader

- Data Preprocessing

	Use
Lowercase	O
Special Character	O
Regular expression Processing	X
Num2word	X
Stopword	X
Correcting typo	O
Acronym	X

Movie Review Credibility Grader

- Training Environment

	Use
Max length	16
Min length	3
Pre-trained Model	GloVe
Teacher forcing ratio	0.5
Maximum Norm	2.0
Word Embedding Dimension	200
Loss	negative log likelihood
Optimizer	Stochastic gradient descent

Movie Review Credibility Grader

- Model Information

	Use
Encoder	GRU
Number of Encoder Layer	1
Decoder	GRU+Attention Layer
Number of Decoder Layer	1

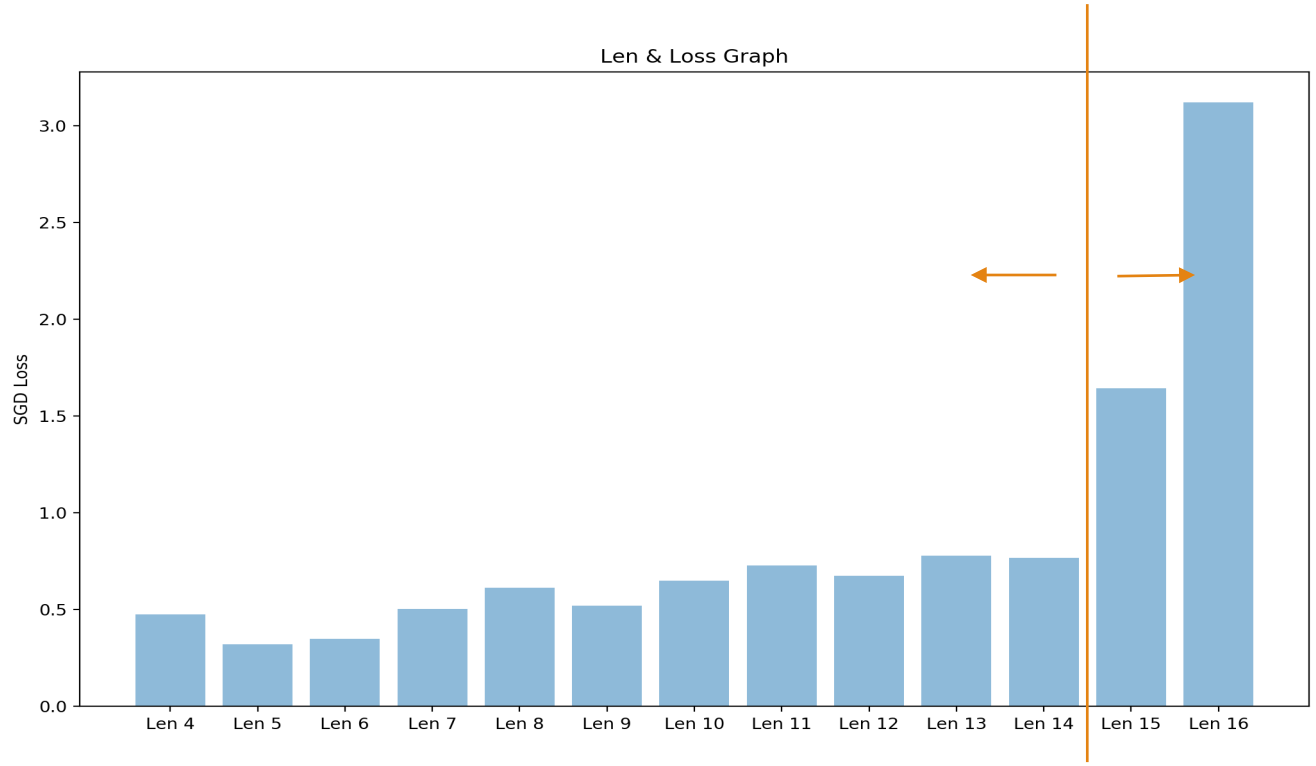
Movie Review Credibility Grader

- If Avg Loss < 0.5 : good model, it can reconstruct any sentences that are in the same domain very well.
- My model has 0.33 Avg loss!!

```
> ahmad uses the eye to see the princess .          <- Input
= ahmad uses the eye to see the princess .          <- Expected Output
< ahmad uses the eye to see the princess . <EOS>    <- Output

> there are a lot of laughs in the process .
= there are a lot of laughs in the process .
< there are a lot of laughs in the process . <EOS>

> he understands what to do without having to think about the matter .
= he understands what to do without having to think about the matter .
< he understands what to do without having to think about the matter . <EOS>

> a typical romantic comedy with its moments i guess .
= a typical romantic comedy with its moments i guess .
< a typical romantic comedy with its moments i guess . <EOS>
```

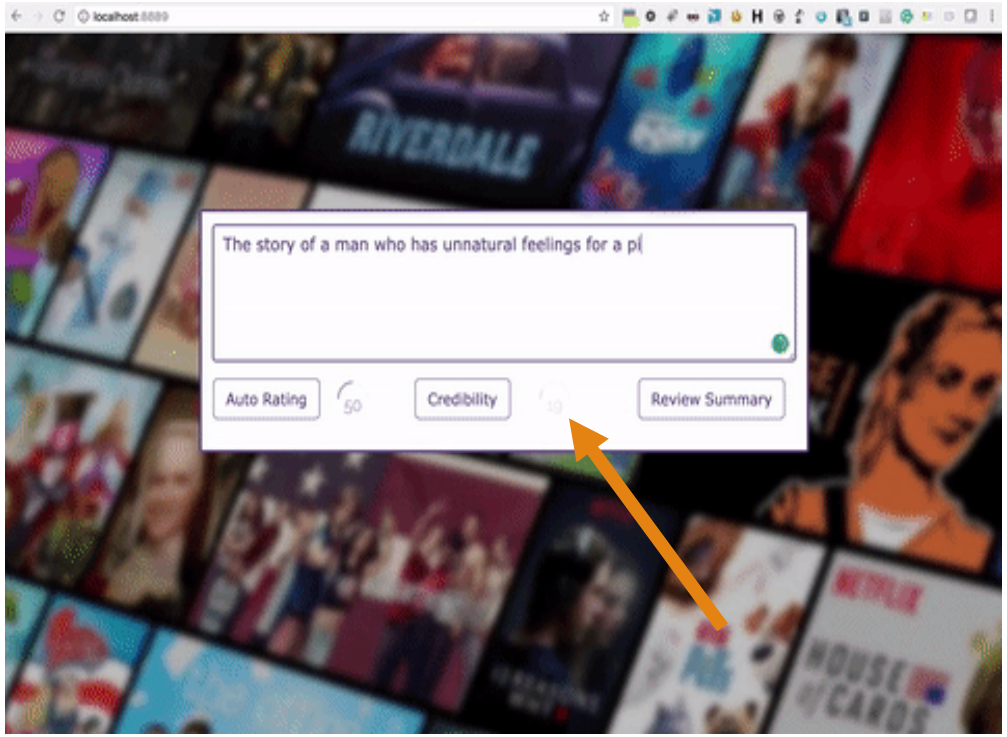
Movie Review Credibility Grader

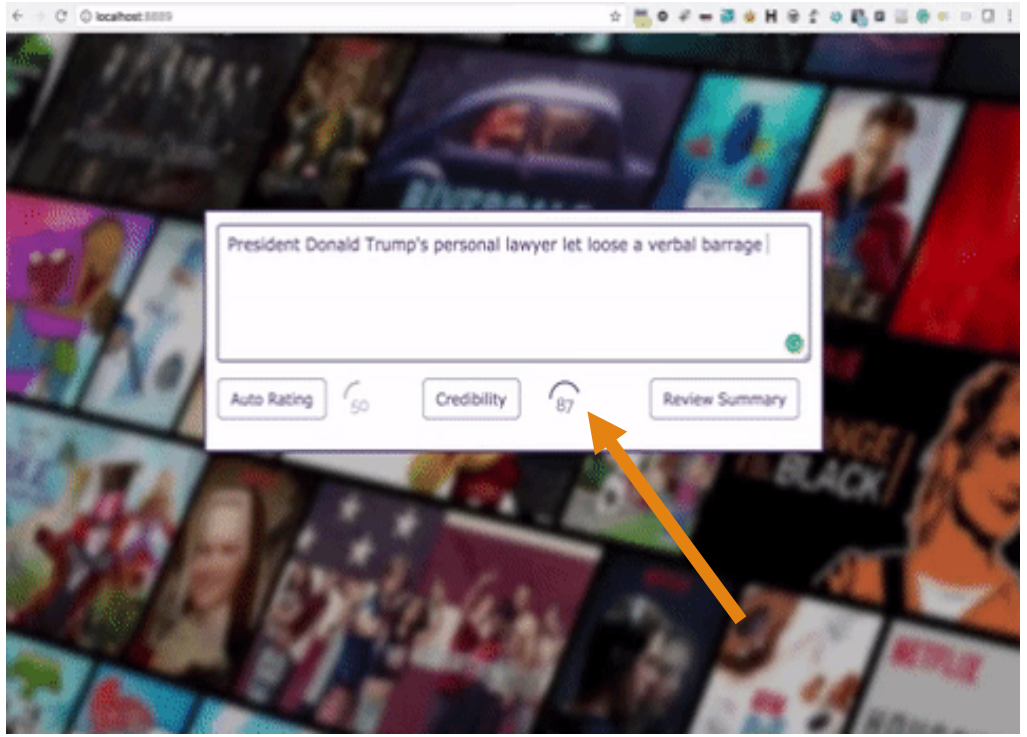
- Limit length test



Movie Review Credibility Grader

- Application Demo(http://localhost:8889)





Call-log Anomaly Detector

- Data Set Information
  - Earning Call log(not public data)
  - 1,500,000 unlabeled sentences related to earning conference call

Call-log Anomaly Detector

- Data Preprocessing

	Use
Lowercase	O
Special Character	O
Regular expression Processing	O
Num2word	O
Stopword	X
Correcting typo	X
Acronym	O

Call-log Anomaly Detector

- One more Trick: Since the loss of sentences that has more than 14 words increases exponentially, I put the threshold as the length 14.
- If I have a sentence that has more than 14 words, I randomly pick the index of it and cut off the sentence from the index up to index+14-1
- Let's assume random index would be 3

Sick	people	must	walk	to	the	back	of
the	drugstore	to	get	their	prescription	while	healthy
people	can	buy	cigarettes	in	the	front	.

Call-log Anomaly Detector

- One more Trick: Since the loss of sentences that has more than 14 words increases exponentially, I put the threshold as the length 14.
- If I have a sentence that has more than 14 words, I randomly pick the index of it and cut off the sentence from the index up to index+14-1
- Let's assume random index would be 3

			walk	to	the	back	of
the	drugstore	to	get	their	prescription	while	healthy
people							

Call-log Anomaly Detector

- Training Environment

	Use
Max length	14
Min length	3
Pre-trained Model	GloVe
Teacher forcing ratio	0.5
Maximum Norm	2.0
Word Embedding Dimension	200
Loss	negative log likelihood
Optimizer	Stochastic gradient descent

Call-log Anomaly Detector

- Model Information

	Use
Encoder	GRU
Number of Encoder Layer	1
Decoder	GRU+Attention Layer
Number of Decoder Layer	1

Call-log Anomaly Detector

- Model Performance
  - Average Loss: 0.31

  - Training Time: about 150 hours based on digits machine

1/5 Random Sample

- **Input**: good day ladies and gentlemen and welcome to the agilent technologies fourth quarter two thousand sixteen earnings conference call
- **Trimmed text**: good day ladies and gentlemen and welcome to the agilent technologies fourth quarter two thousand sixteen earnings conference call
- **Text of 14 length by Random Index Picker**: day ladies and gentlemen and welcome to the agilent technologies fourth quarter two thousand
- **Original text sequence**: ['day', 'ladies', 'and', 'gentlemen', 'and', 'welcome', 'to', 'the', 'agilent', 'technologies', 'fourth', 'quarter', 'two', 'thousand']
- **Decoded text sequence**: ['day', 'ladies', 'and', 'gentlemen', 'and', 'welcome', 'to', 'fiebig', 'agilent', 'technologies', 'fourth', 'quarter', 'two', 'thousand', '<EOS>']
- **Decoded text avg loss**: [0.28271052]
- **Probability of genuine earning call**: [96.13263702]%

2/5 Random Sample

- **Input**: thank you karen and welcome everyone to agilent 's fourth quarter conference call for fiscal year two thousand sixteen
- **Trimmed text**: thank you karen and welcome everyone to agilent 's fourth quarter conference call for fiscal year two thousand sixteen
- **Text of 14 length by Random Index Picker**: and welcome everyone to agilent 's fourth quarter conference call for fiscal year two
- **Original text sequence**: ['and', 'welcome', 'everyone', 'to', 'agilent', "'s", 'fourth', 'quarter', 'conference', 'call', 'for', 'fiscal', 'year', 'two']
- **Decoded text sequence**: ['and', 'welcome', 'everyone', 'to', 'agilent', "'s", 'fourth', 'quarter', 'conference', 'call', 'for', 'fiscal', 'year', 'two', '<EOS>']
- **Decoded text avg loss**: [0.00107612]
- **Probability of genuine earning call**: [99.99993896]%

3/5 Random Sample

- **Input**: i 'm very pleased to announce that our agilent team ended two thousand sixteen with another strong quarter of excellent results
- **Trimmed text**: i 'm very pleased to announce that our agilent team ended two thousand sixteen with another strong quarter of excellent results
- **Text of 14 length by Random Index Picker**: announce that our agilent team ended two thousand sixteen with another strong quarter of
- **Original text sequence**: ['announce', 'that', 'our', 'agilent', 'team', 'ended', 'two', 'thousand', 'sixteen', 'with', 'another', 'strong', 'quarter', 'of']
- **Decoded text sequence**: ['announce', 'that', 'bbc', 'baxter', 'team', 'ended', 'two', 'thousand', 'sixteen', 'with', 'another', 'strong', 'quarter', 'of', '<EOS>']
- **Decoded text avg loss**: [0.54387116]
- **Probability of genuine earning call**: [86.8372345]%

4/5 Random Sample

- **Input**: i will start by looking at our key numbers from the quarter
- **Trimmed text**: i will start by looking at our key numbers from the quarter
- **Text of 14 length by Random Index Picker**: i will start by looking at our key numbers from the quarter
- **Original text sequence**: ['i', 'will', 'start', 'by', 'looking', 'at', 'our', 'key', 'numbers', 'from', 'the', 'quarter']
- **Decoded text sequence**: ['i', 'knocking', 'view', 'thirtyth', 'looking', 'at', 'our', 'key', 'numbers', 'from', 'fiebig', 'quarter', '<EOS>']
- **Decoded text avg loss**: [1.80416393]
- **Probability of genuine earning call**: [32.05383682]%

5/5 Random Sample

- **Input**: first we continued to deliver above market growth
- **Trimmed text**: first we continued to deliver above market growth
- **Text of 14 length by Random Index Picker**: first we continued to deliver above market growth
- **Original text sequence**: ['first', 'we', 'continued', 'to', 'deliver', 'above', 'market', 'growth']
- **Decoded text sequence**: ['first', 'we', 'continued', 'to', 'deliver', 'above', 'market', 'growth', '<EOS>']
- **Decoded text avg loss**: [0.00109439]
- **Probability of genuine earning call**: [99.99993896]%

1/5 Random IMDB sample

- **Input:** This movie was the most disheartening cinematic experience I have ever had
- **Trimmed text:** this movie was the most disheartening cinematic experience i have ever had
- **Text of 14 length by Random Index Picker:** this movie was the most disheartening cinematic experience i have ever had
- **Original text sequence:** ['this', 'movie', 'was', 'the', 'most', 'disheartening', 'cinematic', 'experience', 'i', 'have', 'ever', 'had']
- **Decoded text sequence:** ['this', 'creative', 'alicia', 'winning', 'most', 'gentlemen', 'location', 'experience', 'i', 'have', 'flavors', 'had', '<EOS>']
- **Decoded text avg loss:** [4.35477161]
- **Probability of genuine earning call:** [2.56865072]%

2/5 Random IMDB sample

- **Input:** The only thing I can admire is the acting of some characters.
- **Trimmed text:** the only thing i can admire is the acting of some characters .
- **Text of 14 length by Random Index Picker:** the only thing i can admire is the acting of some characters .
- **Original text sequence:** ['the', 'only', 'thing', 'i', 'can', 'admire', 'is', 'the', 'acting', 'of', 'some', 'characters', '.']
- **Decoded text sequence:** ['gmo', 'only', 'thing', 'due', 'revpar', 'one887', 'is', 'insourced', 'revpar', 'of', 'some', 'characters', 'gentlemen', '<EOS>']
- **Decoded text avg loss:** [4.36413717]
- **Probability of genuine earning call:** [2.54471421]%

3/5 Random IMDB sample

- **Input:** The film could make a fortune being sold as a sleep aide.
- **Trimmed text:** the film could make a fortune being sold as a sleep aide .
- **Text of 14 length by Random Index Picker:** the film could make a fortune being sold as a sleep aide .
- **Original text sequence:** ['the', 'film', 'could', 'make', 'a', 'fortune', 'being', 'sold', 'as', 'a', 'sleep', 'aide', '.']
- **Decoded text sequence:** ['freshman', 'conviction', 'could', 'make', 'a', 'macnevin', 'being', 'sold', 'as', 'a', 'sleep', 'gentlemen', 'gentlemen', '<EOS>']
- **Decoded text avg loss:** [2.04181981]
- **Probability of genuine earning call:** [25.52840805]%

4/5 Random IMDB sample

- **Input:** This movie does not need a doting Jewish mother for comic relief.
- **Trimmed text:** this movie does not need a doting jewish mother for comic relief .
- **Text of 14 length by Random Index Picker:** this movie does not need a doting jewish mother for comic relief .
- **Original text sequence:** ['this', 'movie', 'does', 'not', 'need', 'a', 'doting', 'jewish', 'mother', 'for', 'comic', 'relief', '.']
- **Decoded text sequence:** ['this', 'rotterdam', 'ingots', 'disc', 'fiveg', 'a', 'gentlemen', 'exe', 'revenue', 'for', 'performed', 'relief', 'gentlemen', '<EOS>']
- **Decoded text avg loss:** [4.98894739]
- **Probability of genuine earning call:** [1.36250317]%

5/5 Random IMDB sample

- **Input:** The main characters experience inner struggles and cope with extremely hard decisions.
- **Trimmed text:** the main characters experience inner struggles and cope with extremely hard decisions .
- **Text of 14 length by Random Index Picker:** the main characters experience inner struggles and cope with extremely hard decisions .
- **Original text sequence:** ['the', 'main', 'characters', 'experience', 'inner', 'struggles', 'and', 'cope', 'with', 'extremely', 'hard', 'decisions', '.']
- **Decoded text sequence:** ['fiebig', 'main', 'focusing', 'healthpost', 'sia', 'glasscock', 'and', 'investor', 'with', 'extremely', 'hard', 'decisions', '<EOS>']
- **Decoded text avg loss:** [4.42105055]
- **Probability of genuine earning call:** [2.40397191]%

Anthony's Curiosity

Top 1/5 Anomaly from 1, 000 samples

- **Original Text:** i 'm joined by klaus kleinfeld chairman and chief executive officer and william oplinger executive vice president and chief financial officer
- **Decoded Text:** ['glasco', 'vigilant', 'believing', 'and', 'revpar', 'executive', 'revpar', 'and', 'william', 'ends', 'executive', 'counteracting', 'president', 'president', '<EOS>']
- **Avg Loss:** [5.84113741]

Anthony's Curiosity

Top 2/5 Anomaly from 1, 000 samples

- **Original Text:** this new test can identify pd lone expression levels on the surface of non small cell lung cancer tumor cells and provide information on the survival benefit with opdivo for patients with non squamous non small cell lung cancer
- **Decoded Text:** ['location', 'lung', 'execution', 'office', 'segments', 'and', 'provide', 'information', 'on', 'fiebig', 'exe', 'exe', 'believing', 'believing', '<EOS>']
- **Avg Loss:** [5.26097631]

Anthony's Curiosity

Top 3/5 Anomaly from 1, 000 samples

- **Original Text:** growth in genomics reflected strong market performance in the us and china across our array cgh target enrichment and sureselect products
- **Decoded Text:**['market', 'performance', 'fifth', 'fiebig', 'word', 'and', 'china', 'welcome', 'our', 'array', 'dollarseighty', 'definitions', 'ninety', 'ninety', '<EOS>']
- **Avg Loss:**[5.04620838]

Anthony's Curiosity

Top 4/5 Anomaly from 1, 000 samples

- **Original Text:** and we also launched several targeted solutions such as our gc/q tof pesticide analysis solution and our lc/qtof water analysis system
- **Decoded Text:**['several', 'targeted', 'drive', 'threeeight', 'as', 'our', 'location', 'vigilant', 'pesticide', 'analysis', 'solution', 'and', 'our', 'our', '<EOS>']
- **Avg Loss:**[4.84434652]

Anthony's Curiosity

Top 5/5 Anomaly from 1, 000 samples

- **Original Text:** after comments by klaus and bill we will take your questions
- **Decoded Text:**['certain', 'investor', 'vigilant', 'antonio', 'and', 'bill', 'we', 'rehabilitate', 'take', 'your', 'believing', '<EOS>']
- **Avg Loss:**[4.77075529]


Question

Reference

- http://blog.aylien.com/overview-word-embeddings-history-word2vec-cbow-glove/
- http://colah.github.io/posts/2015-08-Understanding-LSTMs/
- https://blog.heuritech.com/2016/01/20/attention-mechanism/
- http://jokes.cc.com/funny-nationality/f3aybf/signs-you-re-in-america