Learning Distributed Representations of Sentences

Shuai Tang

Cognitive Science, UC San Diego

Brief Self-Introduction

- PhD student in Cognitive Science
- Learning representations of language by exploiting the context information

Learning Distributed Representations of Sentences

- Related work
  - (organized based on my understanding)
- My research

Why?

Learning Distributed Representations of Sentences

Local Representations ⟷ **Distributed Representations**

The simplest way to represent things with neural networks is to dedicate **one** neuron to **each** thing.

Each concept is represented by **many** neurons, and **each** neuron participates in the representation of **many** concepts.

One-hot Encoding

Clustering

Continuous Bag-of-words

Recurrent Neural Networks

Hinton, McClelland & Rumelhard, *Distributed Representations*, Parallel distributed processing 1986

Learning Distributed Representations of Sentences

Local Representations ⟷ **Distributed Representations**

Efficient usage of space.
Better at capturing componential structure in data.

Hinton, McClelland & Rumelhard, *Distributed Representations*, Parallel distributed processing 1986

Learning Distributed Representations of Sentences

Sentence → Vector

We communicate in sentences,
and they convey our thoughts.

Learning Distributed Representations of Sentences

Sentence $\longrightarrow$ Vector

If we convert a sentence into a vector that **captures the meaning** of the sentence, then Google can do much better searches; they can search based on what's being said in a document. (Hinton, 2015)

Natural Reasoning

How to evaluate?

Evaluating Representations of Sentences

- Supervised Evaluation
  - Sentiment Analysis (MR, CR, SUBJ, MPQA, SST, TREC)
  - Paraphrase Detection (MSRP)
  - Caption-Image Retrieval (COCO)
  - Semantic Relatedness (STSBenchmark, SICK)
  - Entailment/Natural Language Inference (**SNLI**, **MultiNLI**, SICK)

Conneau et al., *Supervised Learning of Universal Sentence Representations from Natural Language Inference Data*, EMNLP2017

Evaluating Representations of Sentences

- Supervised Evaluation
  - Sentiment Analysis (MR, CR, SUBJ, MPQA, SST, TREC)
  - Paraphrase Detection (MSRP)
  - Caption-Image Retrieval (COCO)
  - Semantic Relatedness (STSBenchmark, SICK)
  - Entailment/Natural Language Inference (**SNLI**, **MultiNLI**, SICK)
- Unsupervised Evaluation
  - Semantic Textual Similarity (STS14, STS15)

Facebook AI Research, *https://github.com/facebookresearch/SentEval* ,2017

Evaluating Representations of Sentences

- Supervised Evaluation
  - Sentiment Analysis (MR, CR, SUBJ, MPQA, SST, TREC)
  - Paraphrase Detection (MSRP)
  - Caption-Image Retrieval (COCO)
  - Semantic Relatedness (STSBenchmark, SICK)
  - Entailment/Natural Language Inference (**SNLI**, **MultiNLI**, SICK)
- Unsupervised Evaluation
  - Semantic Textual Similarity (STS14, STS15)
- Future…
  - Large-scale NLP tasks (Machine Translation, Amazon/Yelp Rating Prediction, etc.)
  - Human Evaluation

Facebook AI Research, *https://github.com/facebookresearch/SentEval* ,2017

Learning Distributed Representations of Sentences…

Learning Distributed Representations of Sentences

- … from unlabeled data (Context-based)

- … from labeled data

...from unlabeled data

Distributional Hypothesis

(Harris, 1954; Altmann & Steedman, 1988)

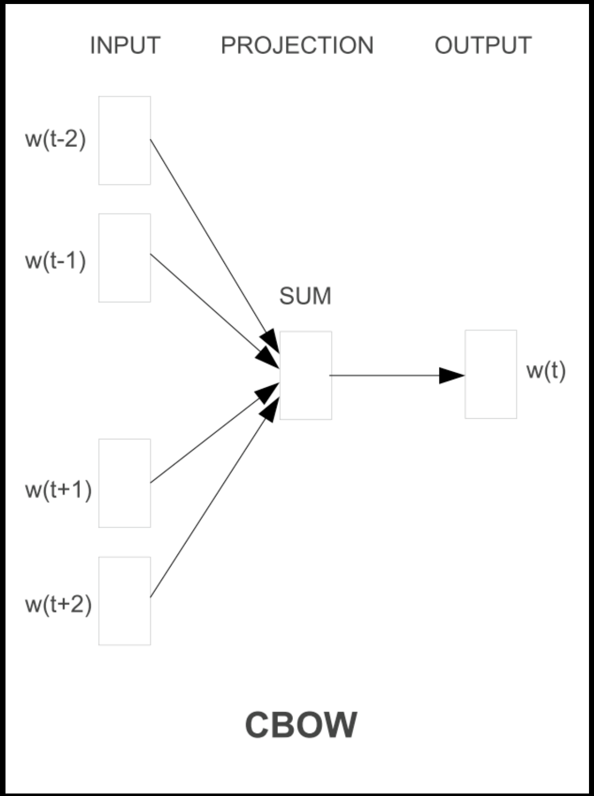
… from unlabeled data (Context-based)

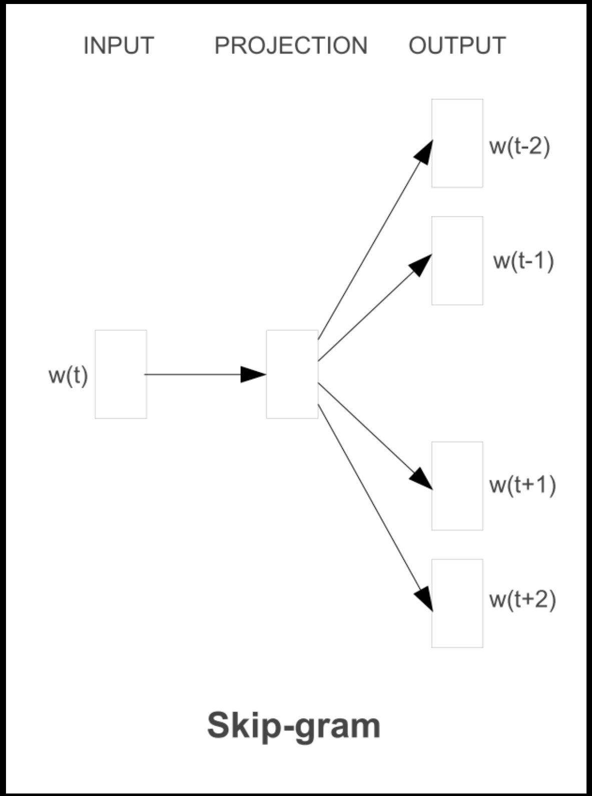
- Generative Objective
  - CBOW & Skip-gram → Paragraph Vectors
  - Skip-thought Vectors → FastSent → CNN-LSTM → Our RNN-CNN model
  - Seq2Seq → Sequential (Denoising) Auto-encoder
  - BTYE m-LSTM
- Discriminative Objective
  - Siamese CBOW → DiscSent → Quick-Thought Vectors

… from unlabeled data (Context-based)

- Generative Objective
  - CBOW & Skip-gram → Paragraph Vectors
  - Skip-thought Vectors → FastSent → CNN-LSTM → Our RNN-CNN model
  - Seq2Seq → Sequential (Denoising) Auto-encoder
  - BTYE m-LSTM
- Discriminative Objective
  - Siamese CBOW → DiscSent → Quick-Thought Vectors

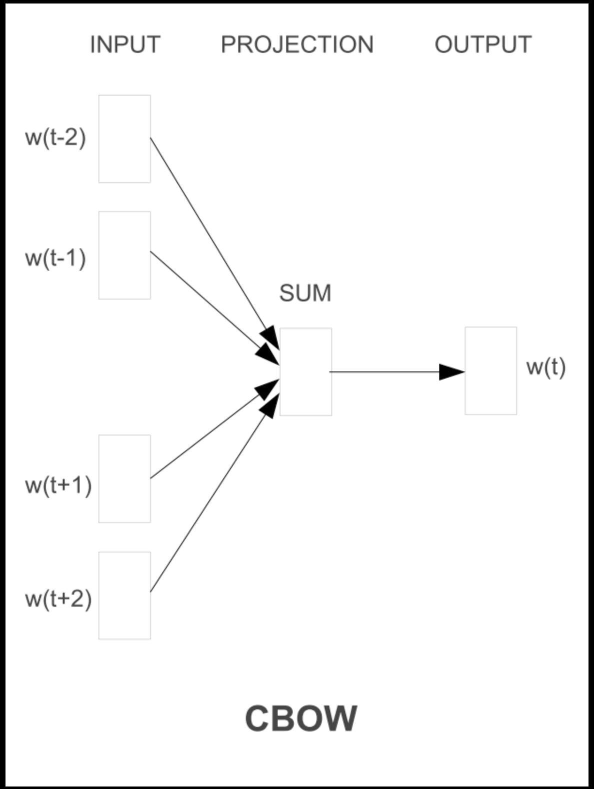
CBOW & Skip-gram

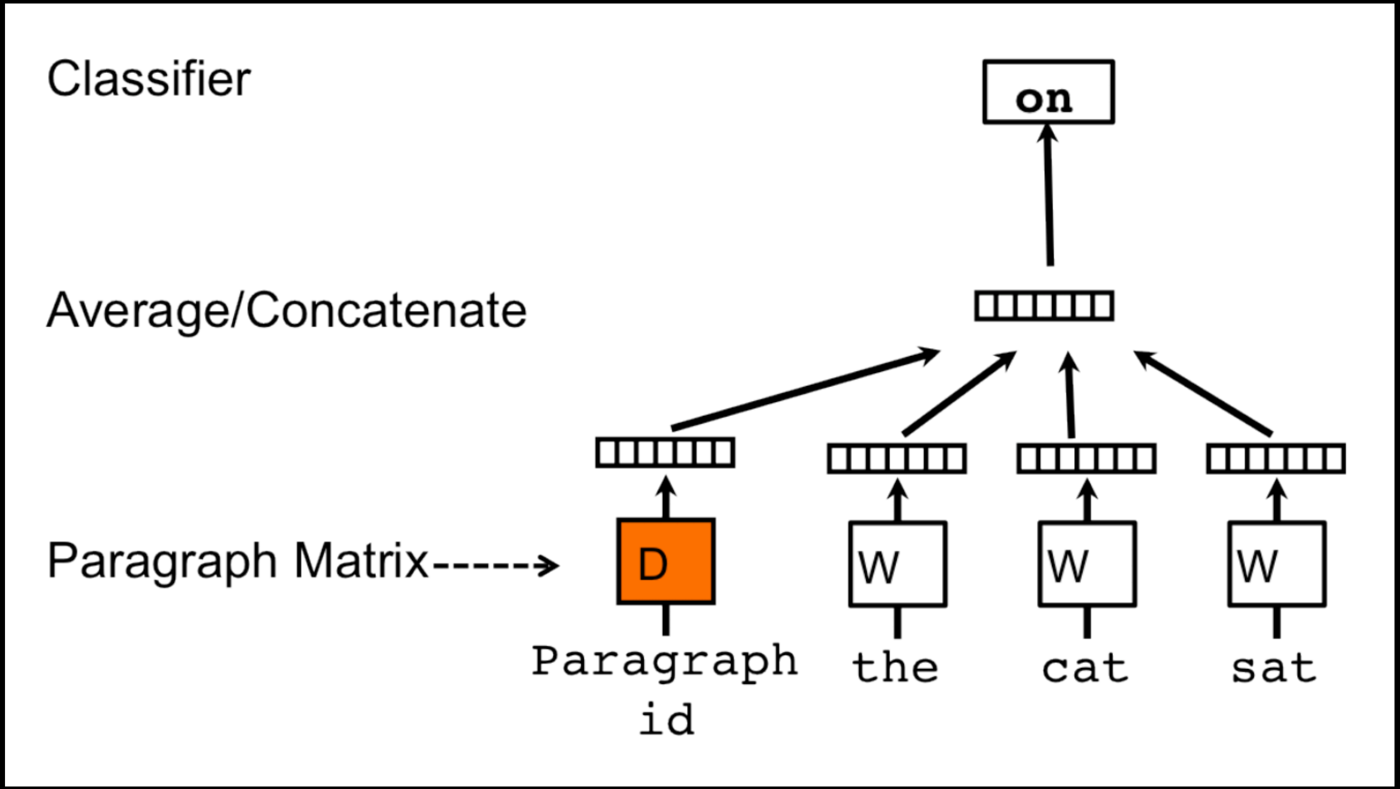




Mikolov et al., *Efficient Estimation of Word Representations in Vector Space*, NIPS2013

CBOW & Skip-gram → Paragraph Vectors



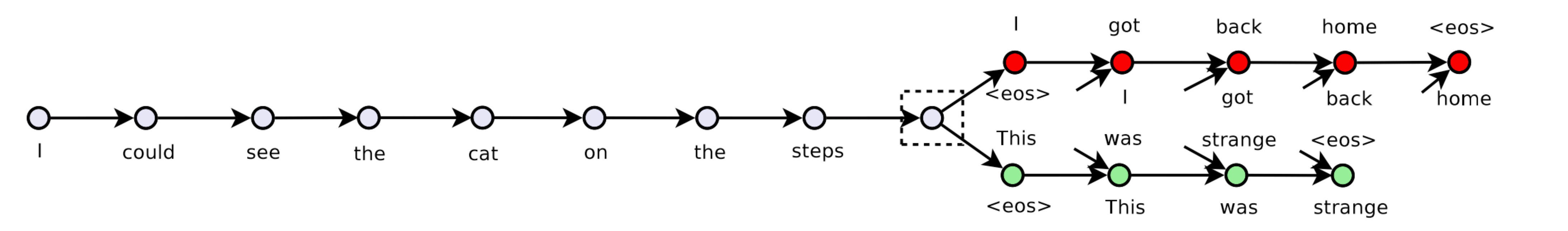


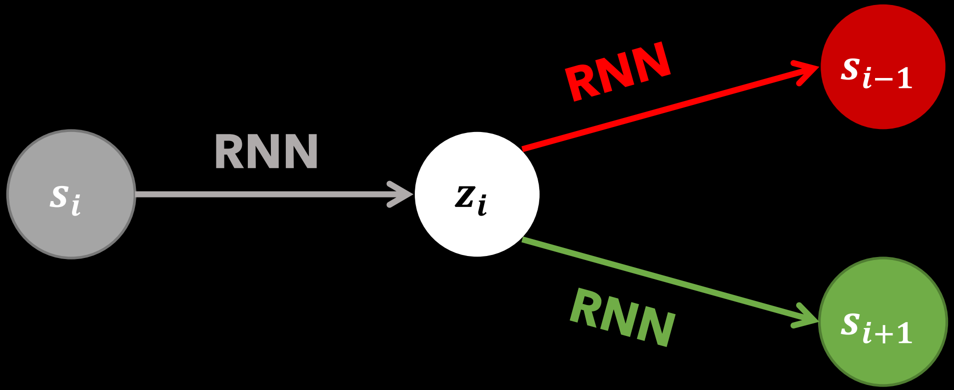
Le et al., *Distributed Representations of Sentences and Documents*, ICML2014

… from unlabeled data (Context-based)

- Generative Objective
  - CBOW & Skip-gram → Paragraph Vectors
  - Skip-thought Vectors → FastSent → CNN-LSTM → Our RNN-CNN model
  - Seq2Seq → Sequential (Denoising) Auto-encoder
  - BTYE m-LSTM
- Discriminative Objective
  - Siamese CBOW → DiscSent → Quick-Thought Vectors

Skip-thought Vectors

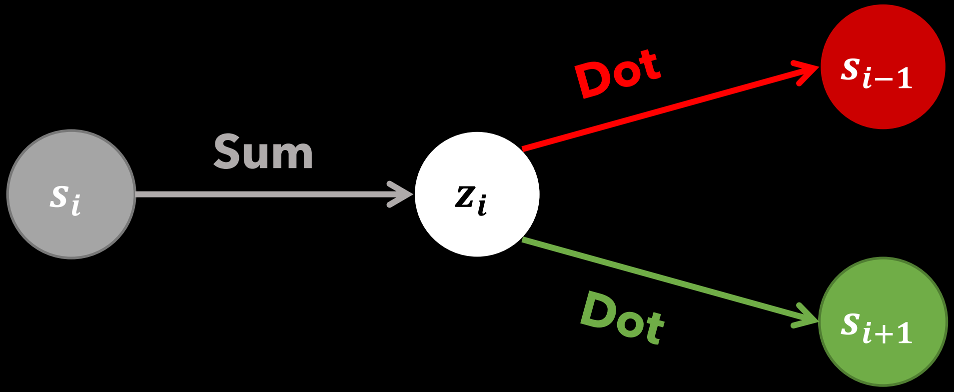


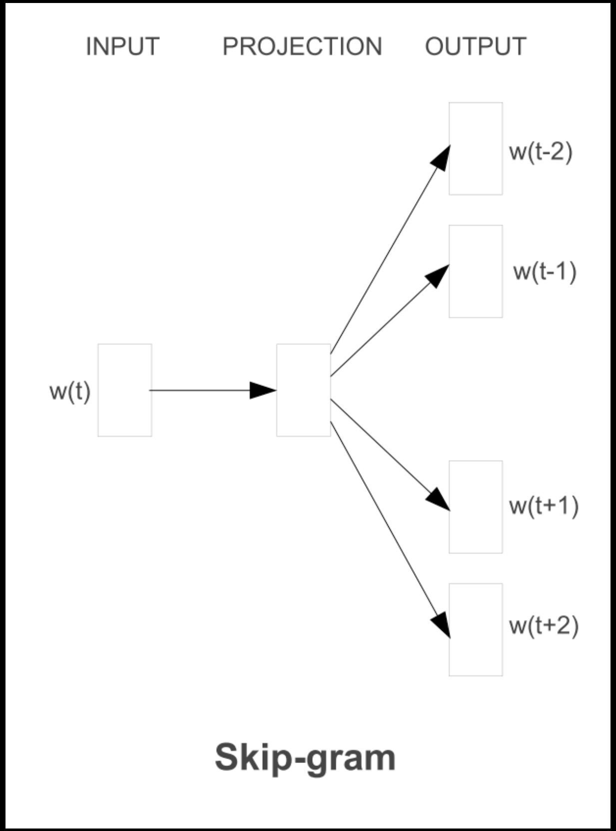


Kiros et al., *Skip-thought Vectors*, NIPS2015

Skip-thought→FastSent (Log-bilinear)

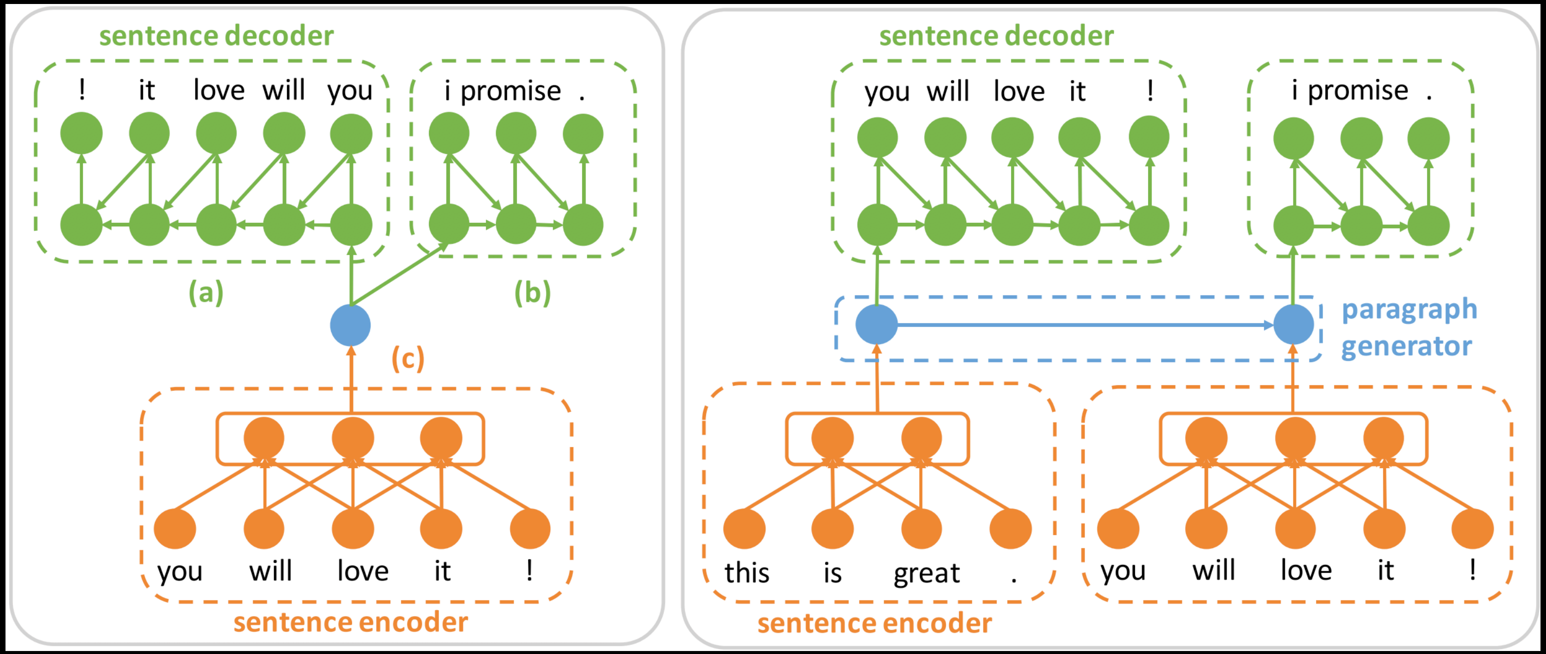
$$\sum_{w \in S_{i-1} \cup S_{i+1}} \phi(\mathbf{s_i}, v_w)$$





Hill et al., *Learning Distributed Representations of Sentences from Unlabelled Data*, HLT-NAACL2016

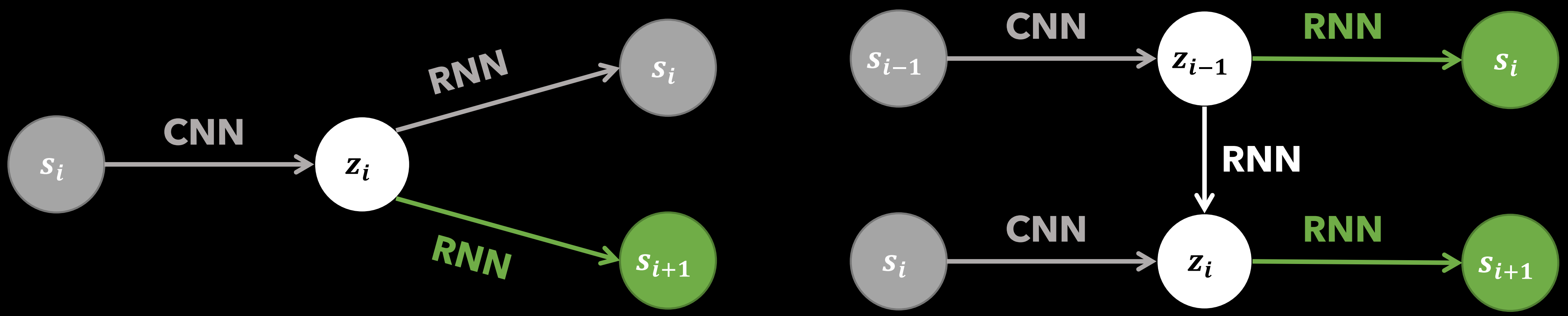
Skip-thought→FastSent→CNN-LSTM



Composite Model **Hierarchical Model**

Gan et al., *Learning Generic Sentence Representations Using Convolutional Neural Networks*, EMNLP2017

Skip-thought→FastSent→CNN-LSTM

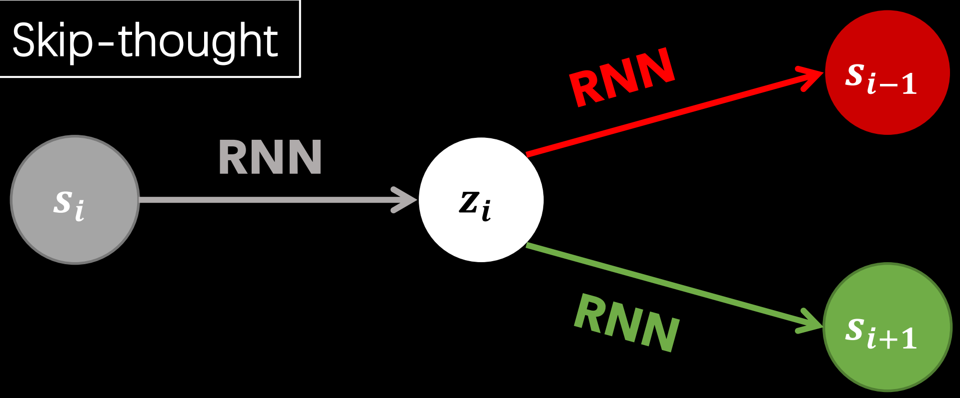


Composite Model **Hierarchical Model**

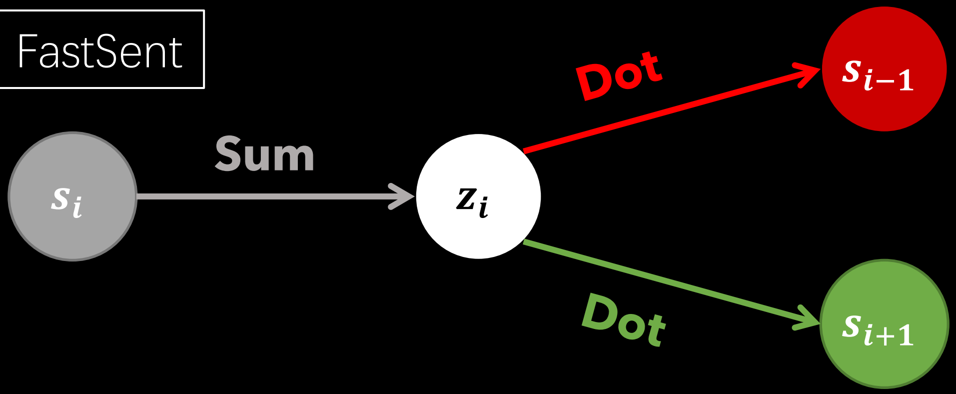
Gan et al., *Learning Generic Sentence Representations Using Convolutional Neural Networks*, EMNLP2017

Skip-thought→FastSent→CNN-LSTM

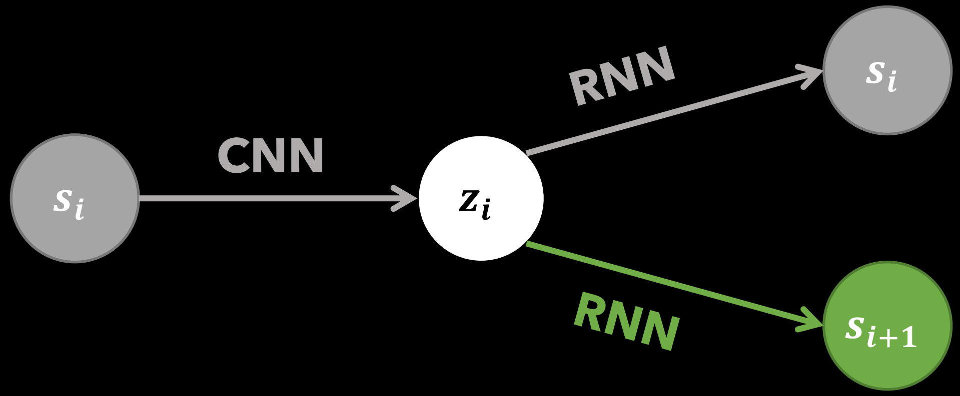
Skip-thought



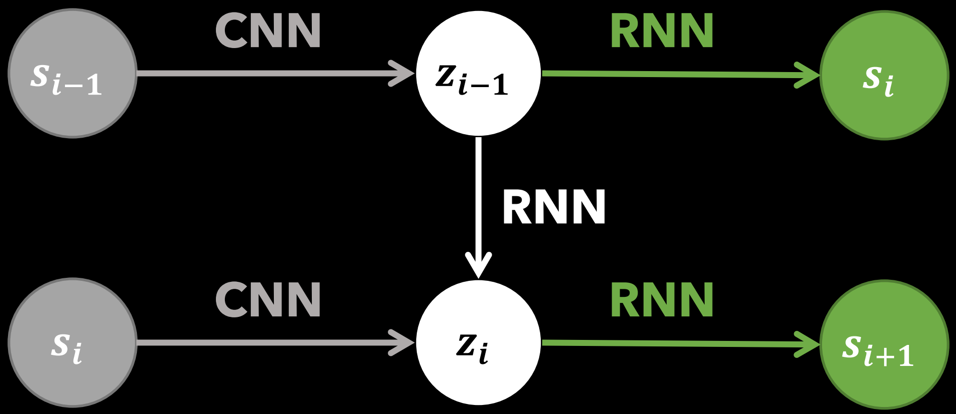
FastSent



Composite CNN-LSTM

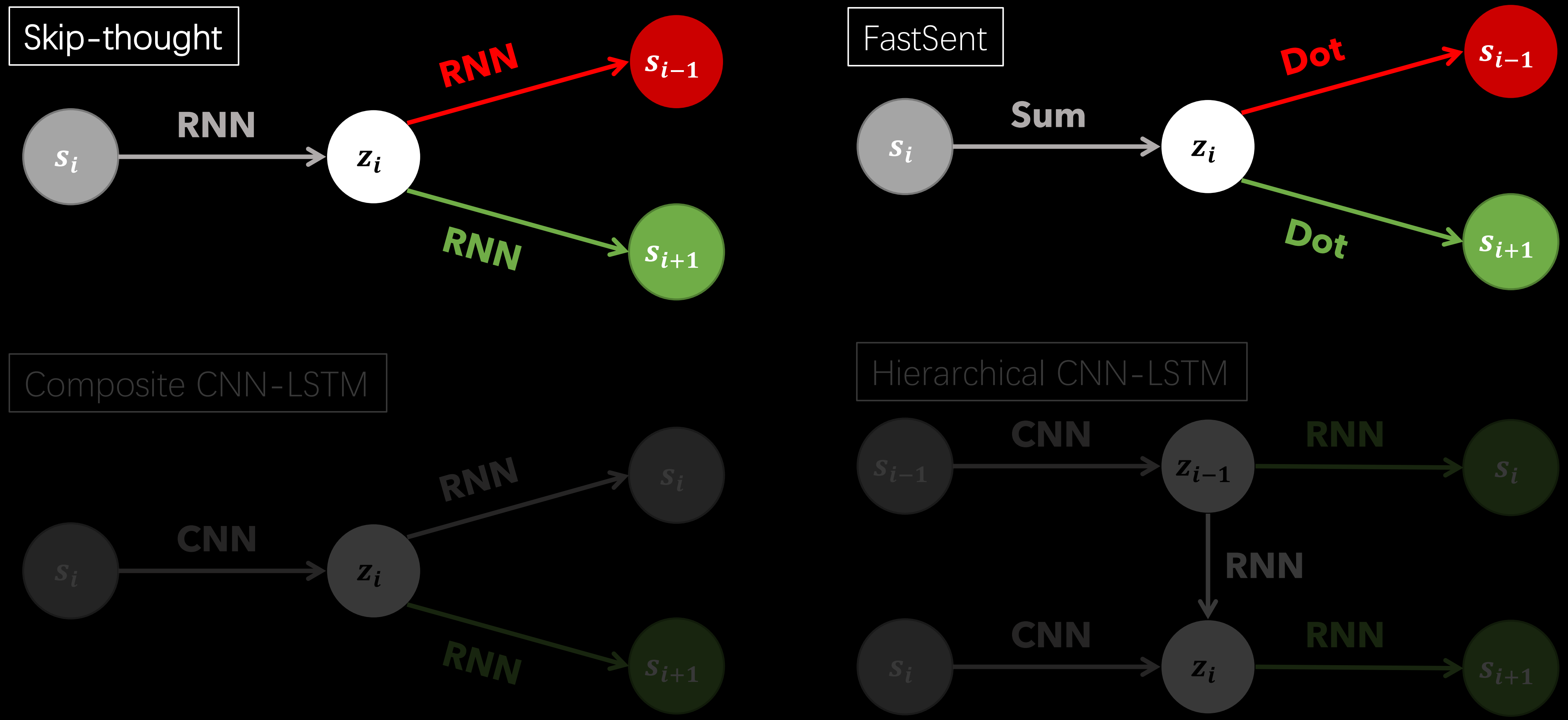


Hierarchical CNN-LSTM



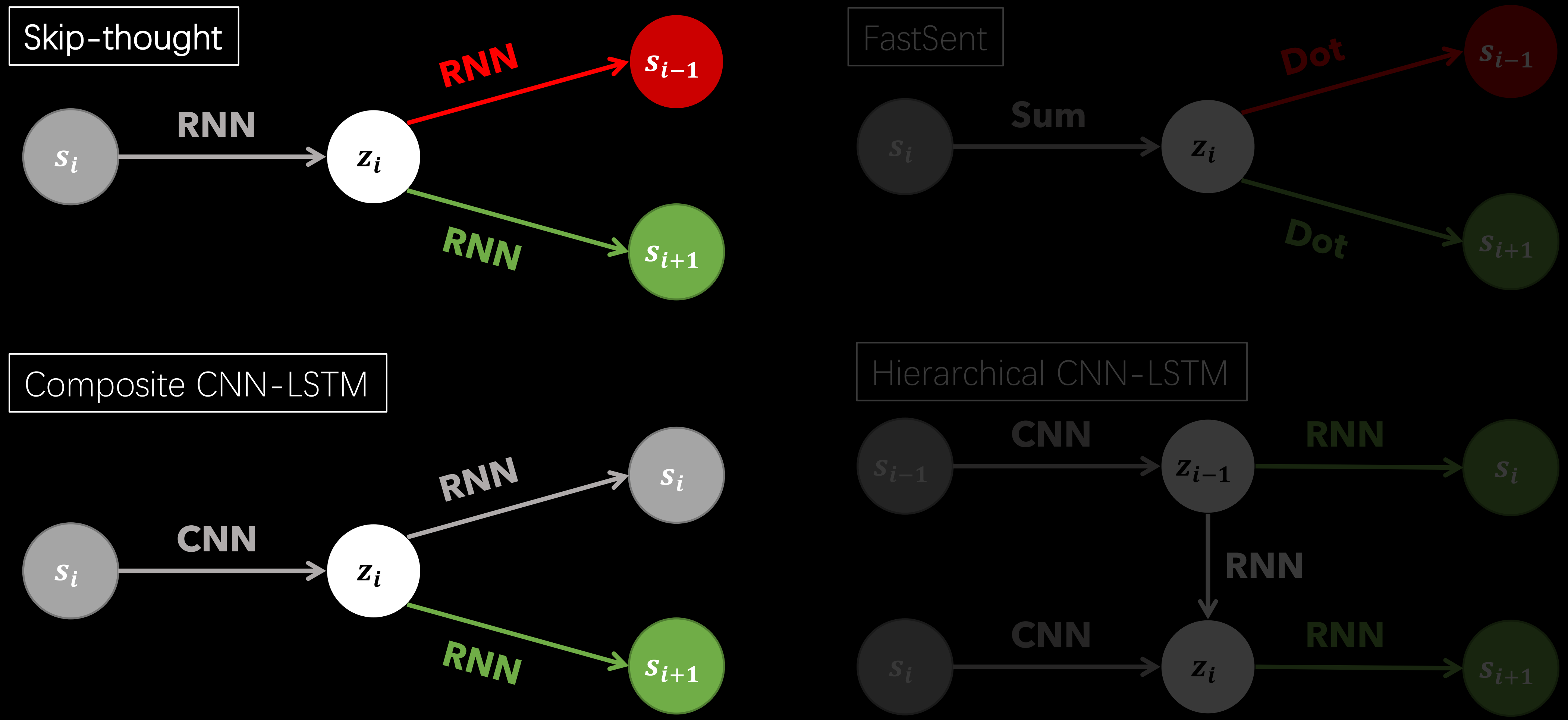
Tang et al., *Trimming and Improving Skip-thought Vectors*, ArXiv2017

Skip-thought→FastSent→CNN-LSTM



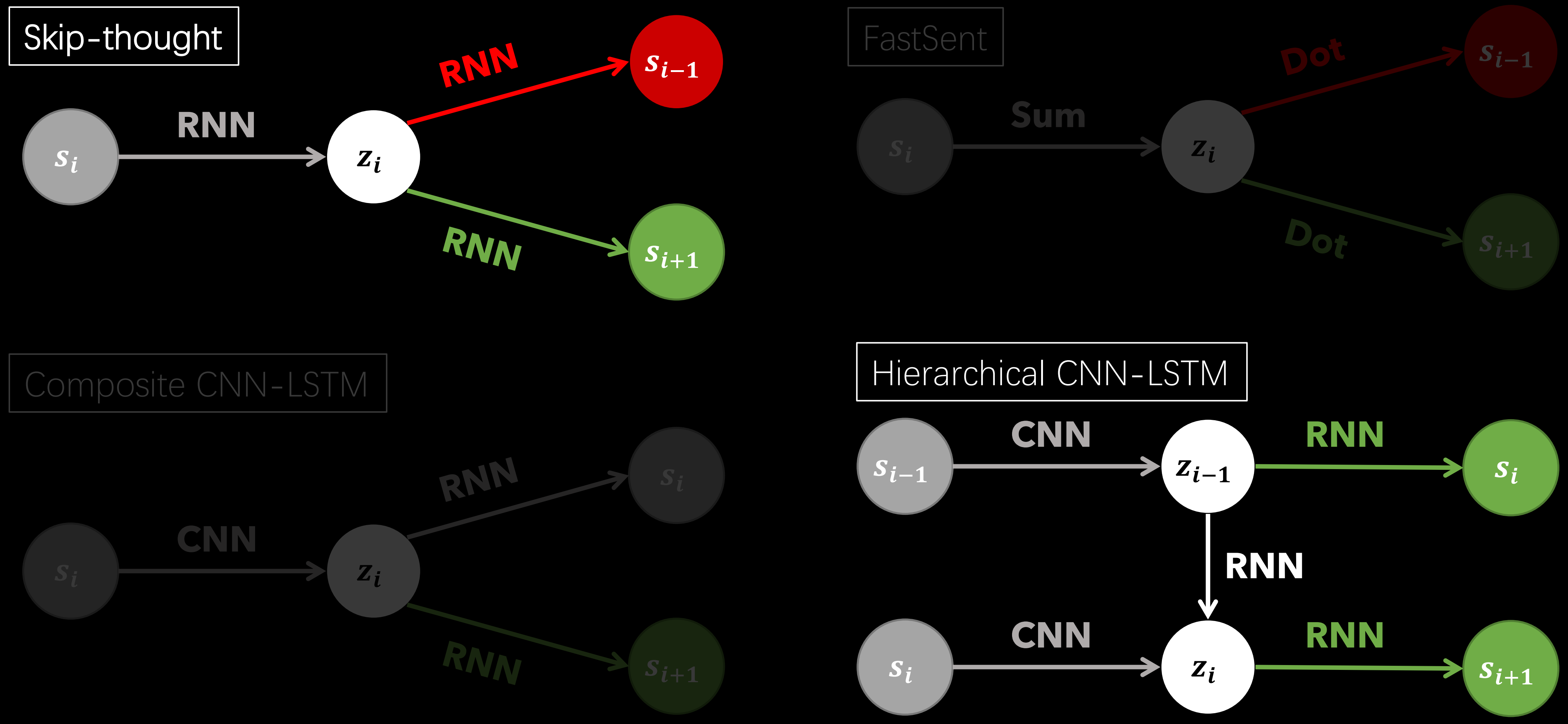
Tang et al., *Trimming and Improving Skip-thought Vectors*, ArXiv2017

Skip-thought→FastSent→CNN-LSTM



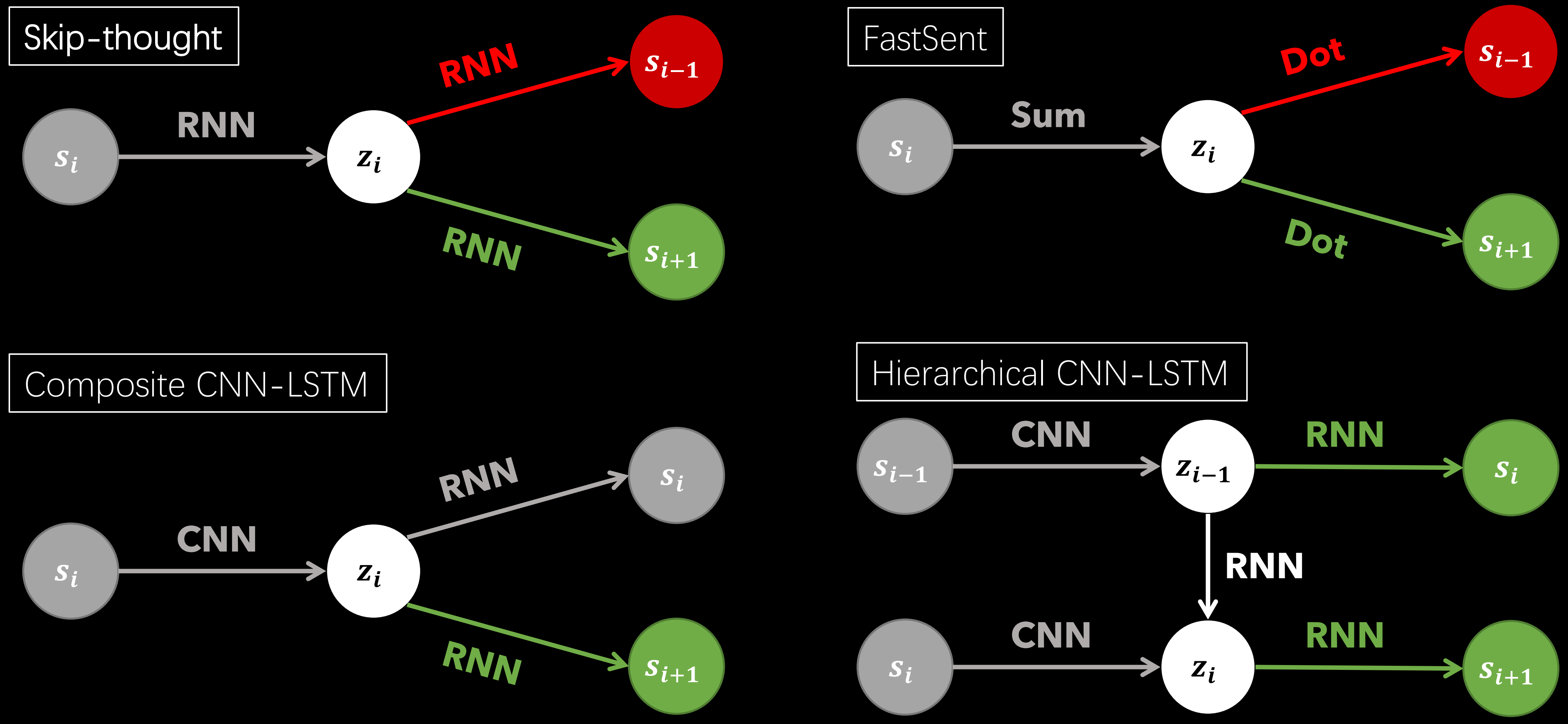
Tang et al., *Trimming and Improving Skip-thought Vectors*, ArXiv2017

Skip-thought→FastSent→CNN-LSTM



Tang et al., *Trimming and Improving Skip-thought Vectors*, ArXiv2017

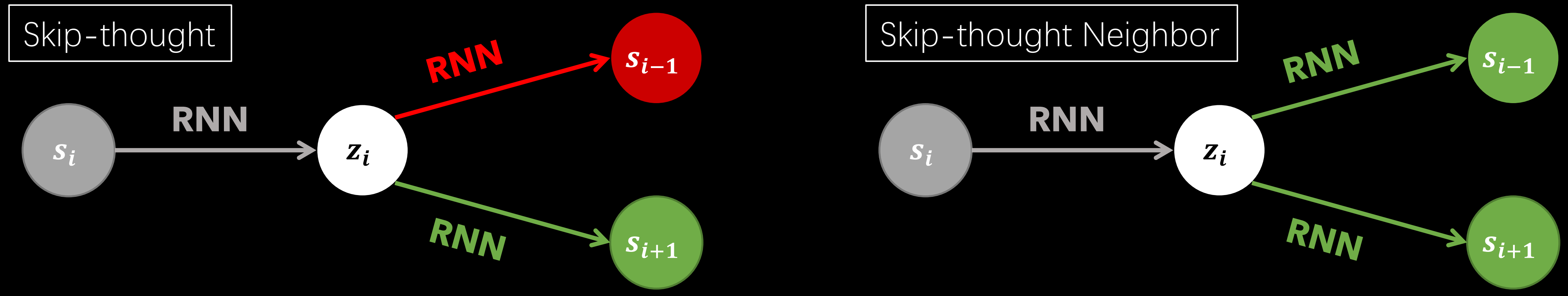
Skip-thought→FastSent→CNN-LSTM



Tang et al., *Trimming and Improving Skip-thought Vectors*, ArXiv2017

Skip-thought→Our Skip-thought Neighbor

- Neighborhood Hypothesis
- Given the current sentence, inferring the previous sentence and inferring the next sentence both provide **same** supervision power.

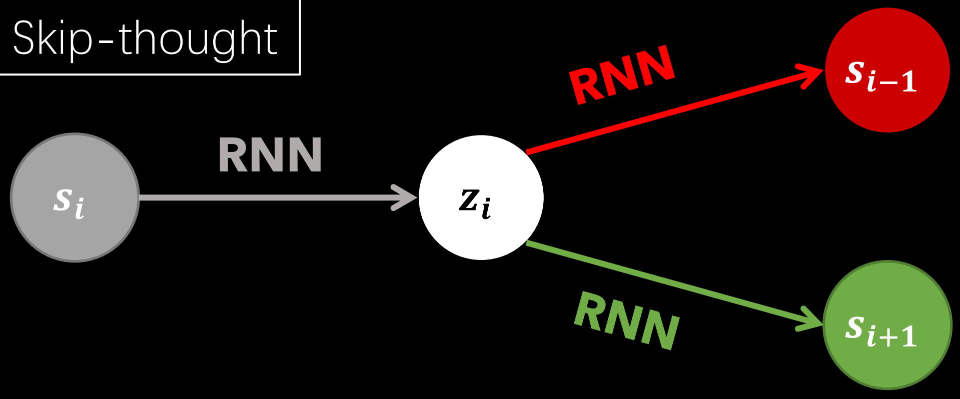


Tang et al., *Rethinking Skip-thought: A Neighborhood based Approach*, RepL4NLP@ACL2017

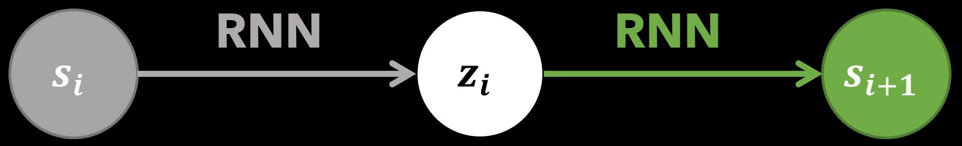
Skip-thought→Our Trimmed Skip-thought

- Neighborhood Hypothesis
- Given the current sentence, inferring the previous sentence and inferring the next sentence both provide **same** supervision power.

Skip-thought



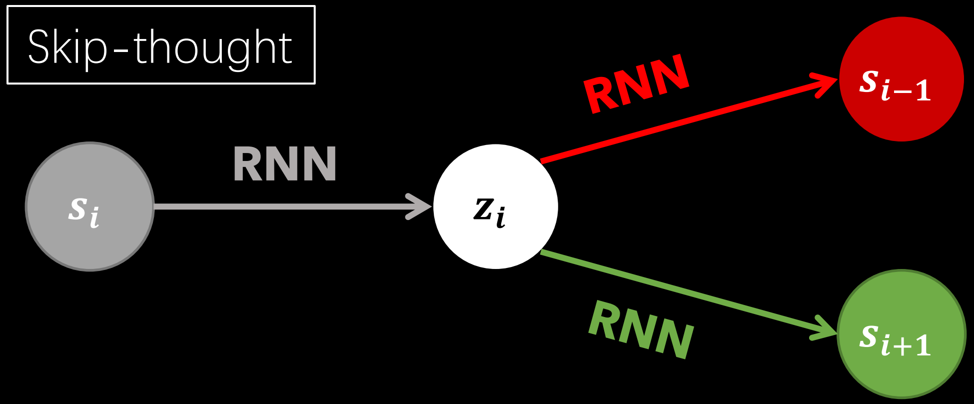
Trimmed Skip-thought

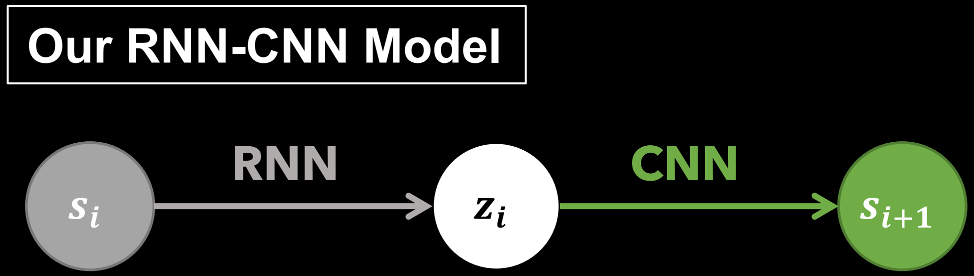


Tang et al., *Trimming and Improving Skip-thought Vectors*, ArXiv2017

Skip-thought→Our RNN-CNN Model

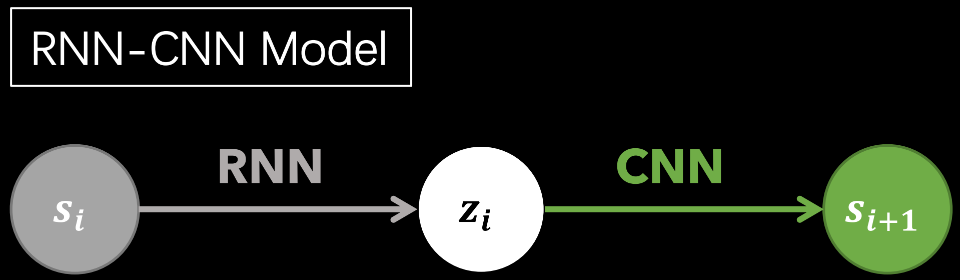
- Neighborhood Hypothesis
- Given the current sentence, inferring the previous sentence and inferring the next sentence both provide **same** supervision power.





Tang et al., *Exploring Asymmetric Encoder-Decoder Structure for Context-based Sentence Representation Learning*, ArXiv2017

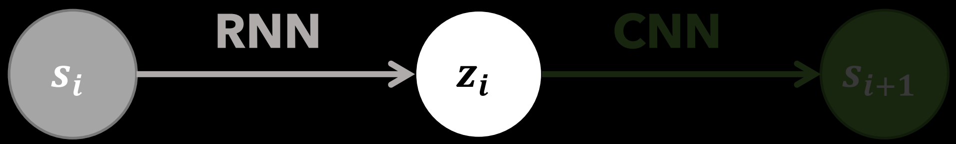
Our RNN-CNN model



Tang et al., *Exploring Asymmetric Encoder-Decoder Structure for Context-based Sentence Representation Learning*, ArXiv2017

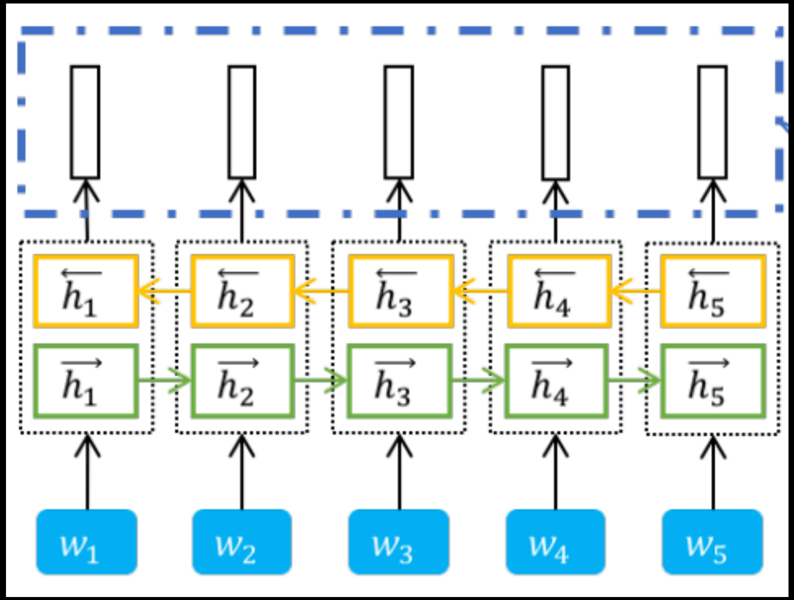
Our RNN-CNN model

RNN-CNN Model



Encoder: Bi-directional GRU

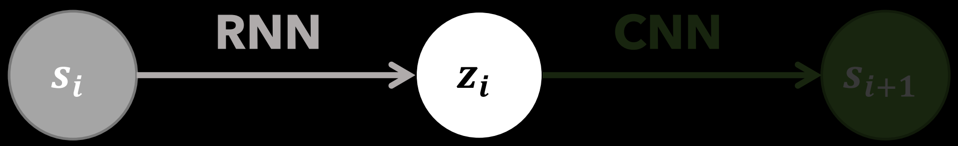
Explicit usage of word order information



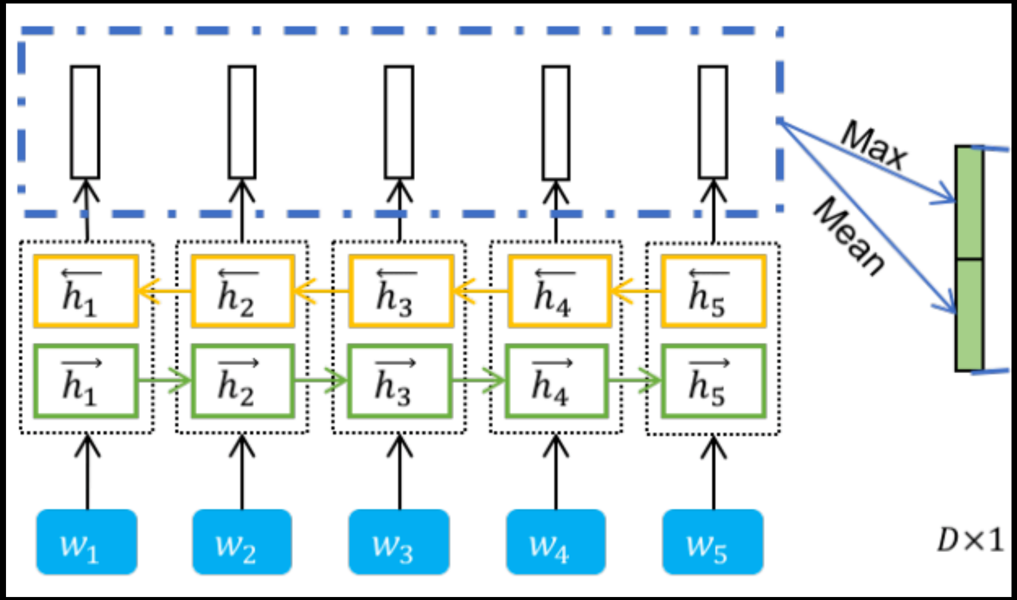
Tang et al., *Exploring Asymmetric Encoder-Decoder Structure for Context-based Sentence Representation Learning*, ArXiv2017

Our RNN-CNN model

RNN-CNN Model

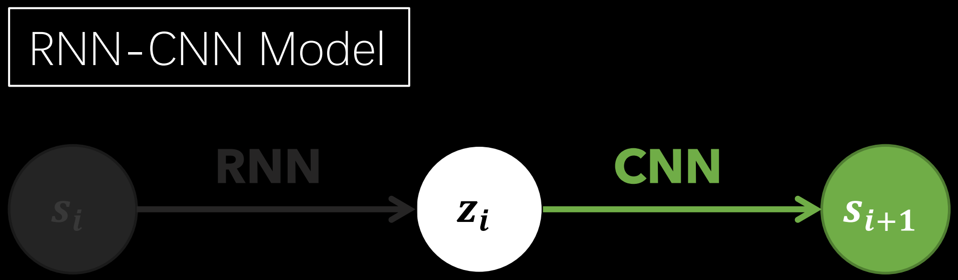


Mean+Max Pooling

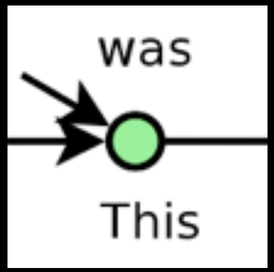


Tang et al., *Exploring Asymmetric Encoder-Decoder Structure for Context-based Sentence Representation Learning*, ArXiv2017

Our RNN-CNN model



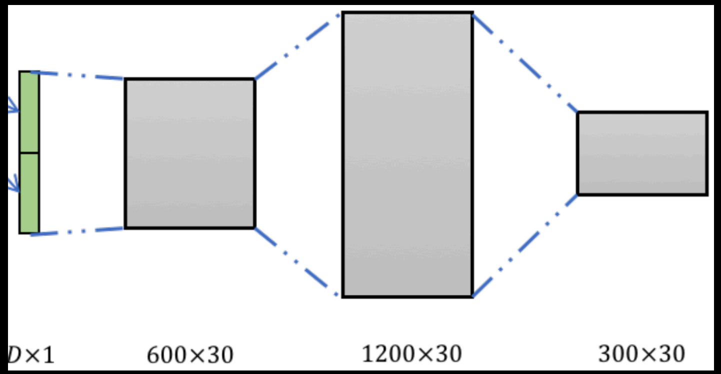
Decoder: 3-layer ConvNet



Higher Training Efficiency

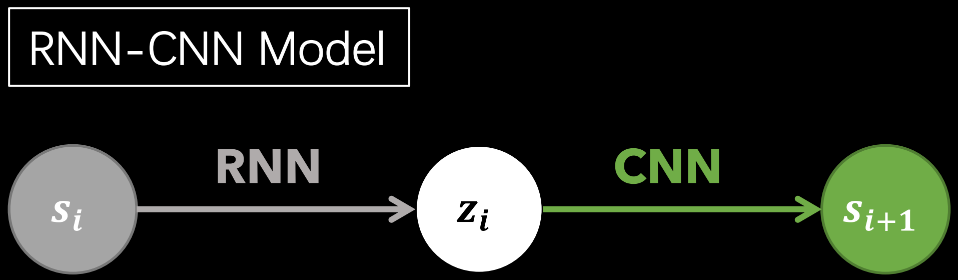
Less Constraints in decoding

Application



Tang et al., *Exploring Asymmetric Encoder-Decoder Structure for Context-based Sentence Representation Learning*, ArXiv2017

Our RNN-CNN model

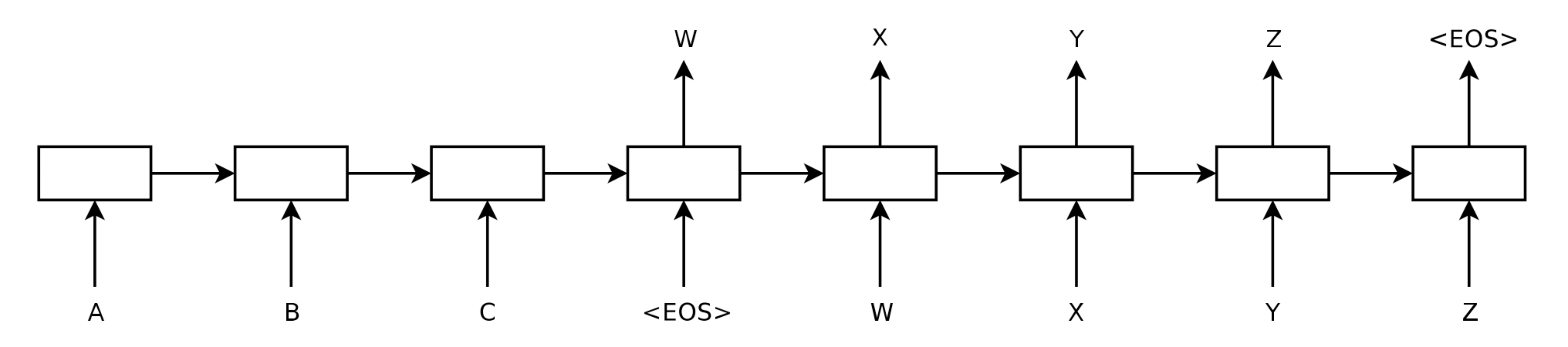


Tang et al., *Exploring Asymmetric Encoder-Decoder Structure for Context-based Sentence Representation Learning*, ArXiv2017

… from unlabeled data (Context-based)

- Generative Objective
  - CBOW & Skip-gram → Paragraph Vectors
  - Skip-thought Vectors → FastSent → CNN-LSTM → Our RNN-CNN model
  - Seq2Seq → Sequential (Denoising) Auto-encoder
  - BTYE m-LSTM
- Discriminative Objective
  - Siamese CBOW → DiscSent → Quick-Thought Vectors

Seq2Seq



Sutskever et al., *Sequence to Sequence Learning with Neural Networks*, NIPS2014

Seq2Seq → Sequential (D) Auto-Encoder

SAE ↔ **SDAE**

- Noise is applied on the source sentences.
- In a randomly selected training sentence,
  - Each word has 10% probability of being deleted.
  - Each word has 10% probability of being swapped with the next one.

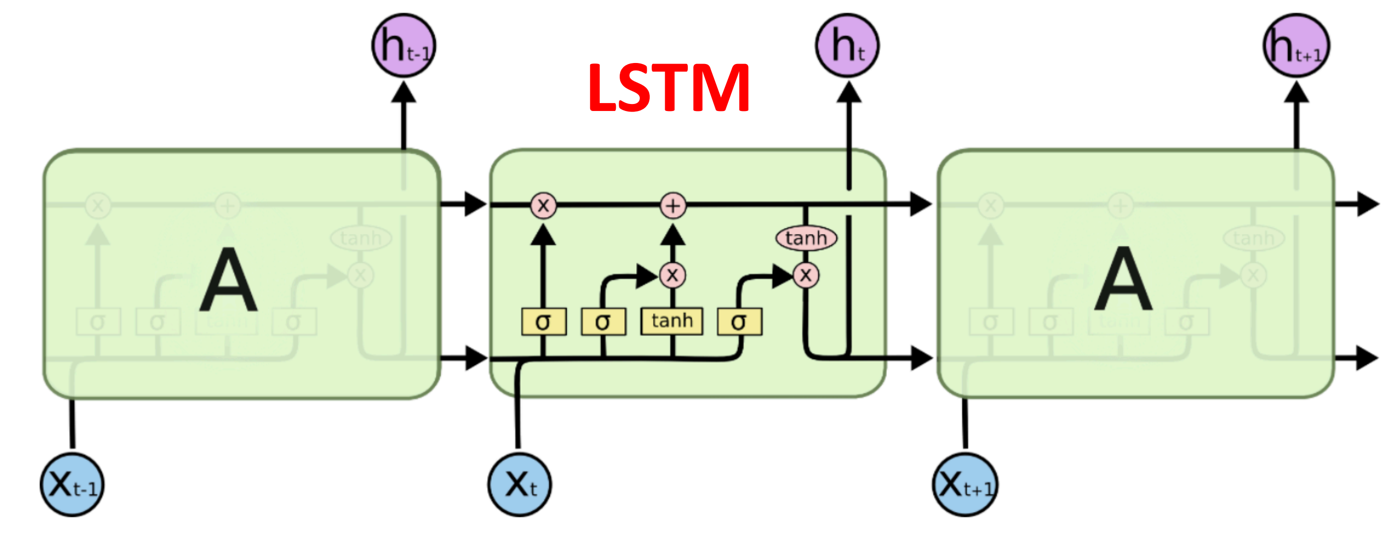
Hill et al., *Learning Distributed Representations of Sentences from Unlabelled Data*, HLT-NAACL2016

… from unlabeled data (Context-based)

- Generative Objective
  - CBOW & Skip-gram → Paragraph Vectors
  - Skip-thought Vectors → FastSent → CNN-LSTM → Our RNN-CNN model
  - Seq2Seq → Sequential (Denoising) Auto-encoder
  - BTYE m-LSTM
- Discriminative Objective
  - Siamese CBOW → DiscSent → Quick-Thought Vectors

BYTE multiplicative-LSTM

- Character-level Language Modeling
- Multiplicative-LSTM



http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Radford et al., *Learning to Generate Reviews and Discovering Sentiment*, ArXiv2017

… from unlabeled data (Context-based)

- Generative Objective
  - CBOW & Skip-gram → Paragraph Vectors
  - Skip-thought Vectors → FastSent → CNN-LSTM → Our RNN-CNN model
  - Seq2Seq → Sequential (Denoising) Auto-encoder
  - BTYE m-LSTM
- Discriminative Objective
  - Siamese CBOW → DiscSent → Quick-Thought Vectors

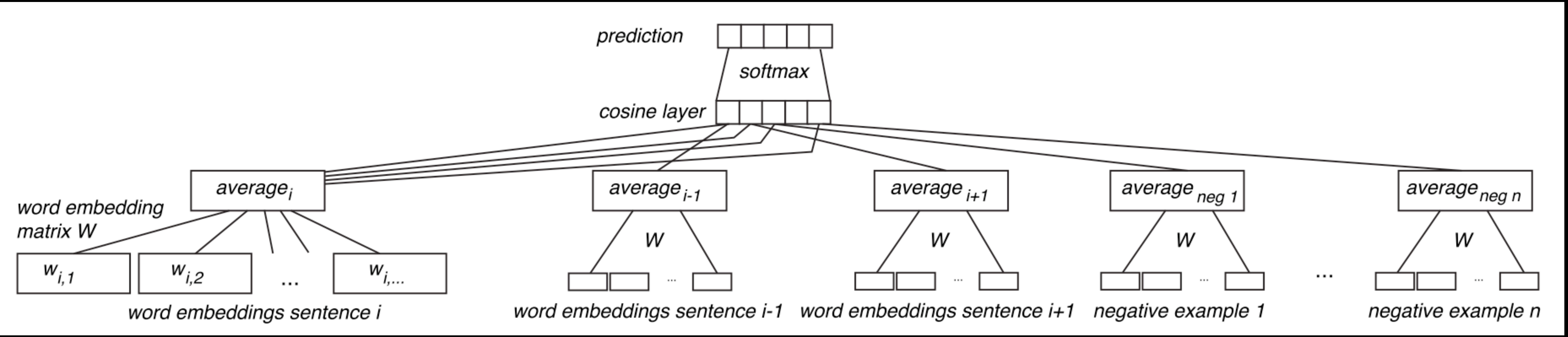
Distributional Hypothesis

(Harris, 1954; Altmann & Steedman, 1988)

… from unlabeled data (Context-based)

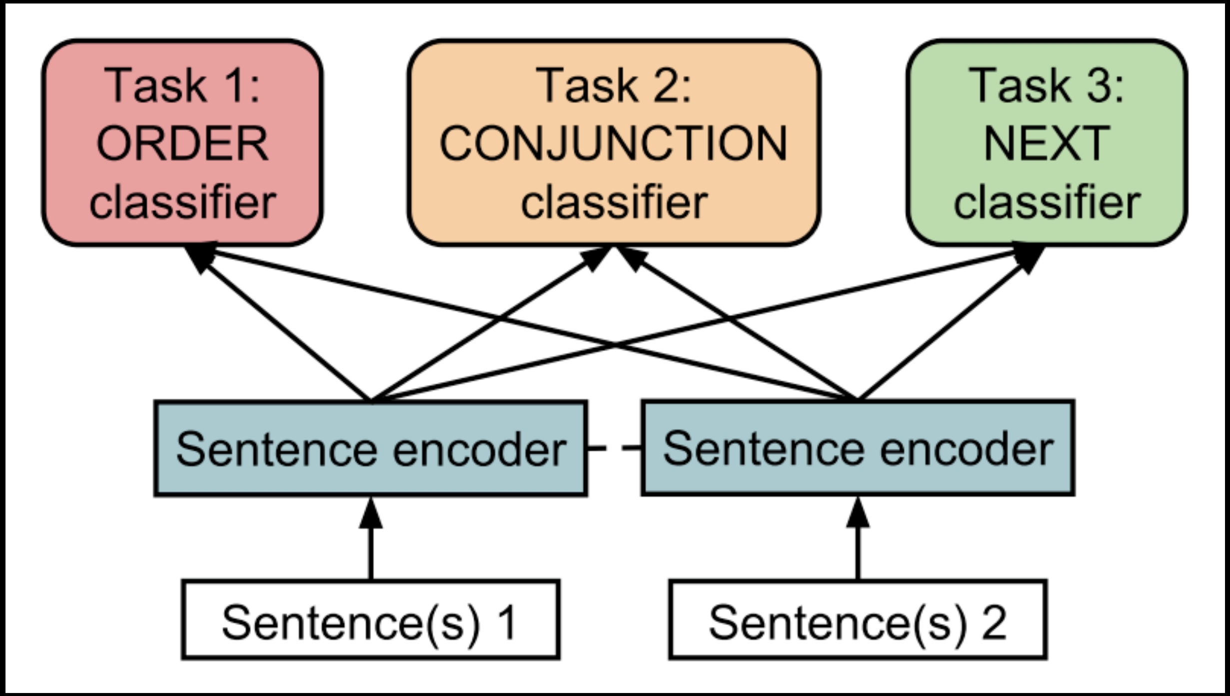
- Generative Objective
  - CBOW & Skip-gram → Paragraph Vectors
  - Skip-thought Vectors → FastSent → CNN-LSTM → Our RNN-CNN model
  - Seq2Seq → Sequential (Denoising) Auto-encoder
  - BTYE m-LSTM
- Discriminative Objective
  - Siamese CBOW → DiscSent → Quick-Thought Vectors

Siamese CBOW



Kenter et al., *Siamese CBOW: Optimizing Word Embeddings for Sentence Representations*, ACL2016

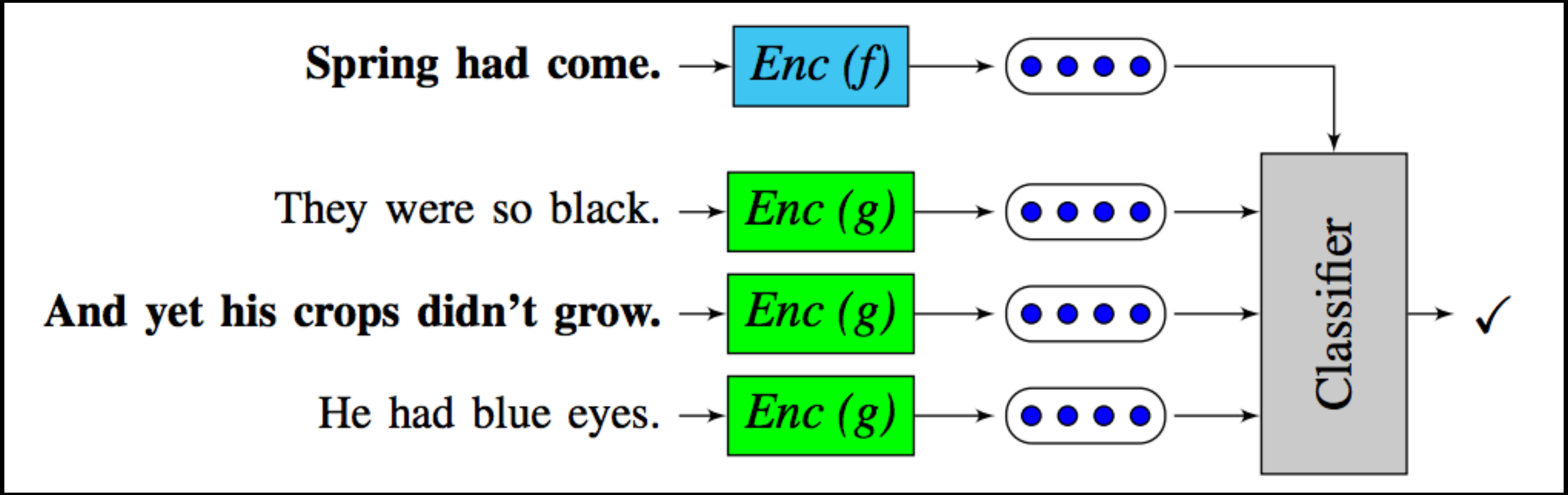
Siamese CBOW → DiscSent



Sentence Pair	**Label**
He had a point. *For good measure, I pouted.*	RETURN *(Still)*
It doesn't hurt at all. *It's exhilarating.*	STRENGTHEN *(In fact)*
The waterwheel hammered on. *There was silence.*	CONTRAST *(Otherwise)*

Jernite et al., *Discourse-Based Objectives for Fast Unsupervised Sentence Representation Learning*, ArXiv2017

Siamese CBOW → DiscSent → Quick-thought



Anonymous., *An Efficient Framework for Learning Sentence Representations*, Submitted to ICLR2018

… from unlabeled data (Context-based)

- Generative Objective
  - CBOW & Skip-gram → Paragraph Vectors
  - Skip-thought Vectors → FastSent → CNN-LSTM → Our RNN-CNN model
  - Seq2Seq → Sequential (Denoising) Auto-encoder
  - BTYE m-LSTM
- Discriminative Objective
  - Siamese CBOW → DiscSent → Quick-Thought Vectors

Data

- BookCorpus (Zhu et al., ICCV2015)
  - Romance, Fantasy, Science fiction, Teen, etc.
- Wikipedia
  - Scientific description
- Amazon Review Data (McAuley et al., SIGIR2015)
  - Reviews with relatively strong personal preference

…from labeled data

… from labeled data

- Natural Language Inference (NLI) datasets
  - Stanford NLI *(Bowman et al., EMNLP2015)*
  - Multi-genre NLI *(Williams et al., ArXiv2017)*

A man inspects the uniform of a figure in some East Asian country.	**contradiction** C C C C C	The man is sleeping
An older and younger man smiling.	**neutral** N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	**contradiction** C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	**entailment** E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	**neutral** N N E C N	A happy woman in a fairy costume holds an umbrella.

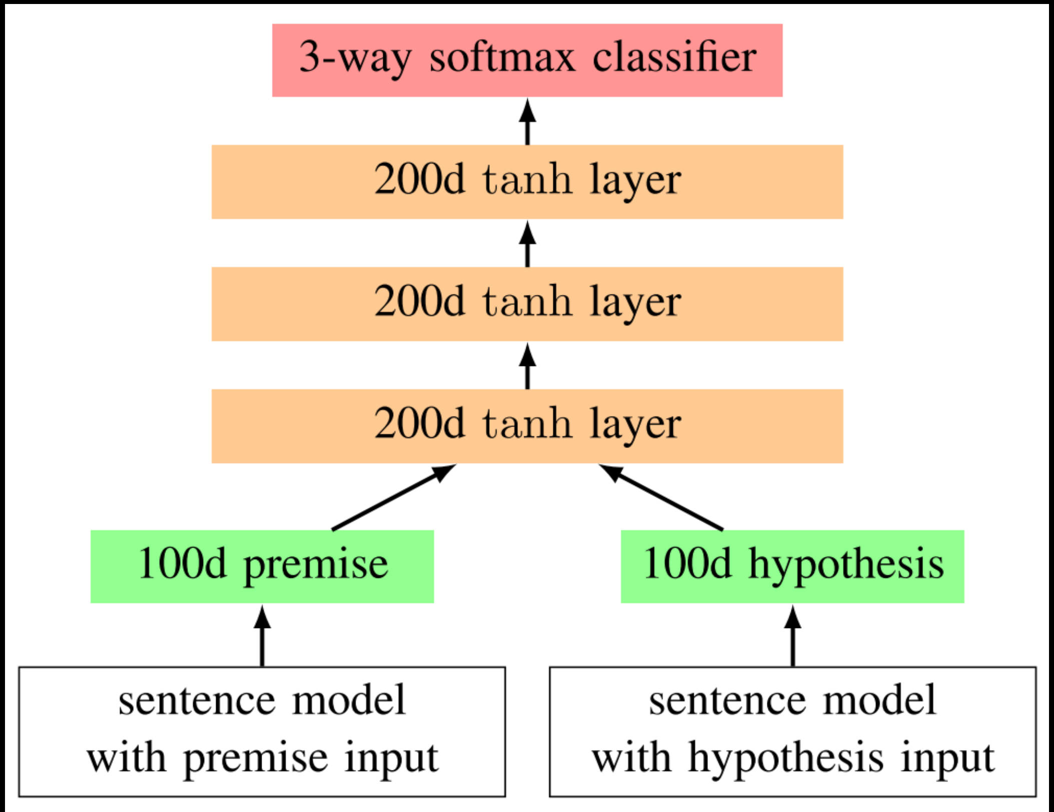
Premise **Hypothesis**

… from labeled data

- Natural Language Inference (NLI) datasets
  - Stanford NLI *(Bowman et al., EMNLP2015)*
  - Multi-genre NLI *(Williams et al., ArXiv2017)*

	#Examples		
Genre	**Train**	**Dev.**	**Test**
SNLI	*550,152*	*10,000*	*10,000*
FICTION	77,348	2,000	2,000
GOVERNMENT	77,350	2,000	2,000
SLATE	77,306	2,000	2,000
TELEPHONE	83,348	2,000	2,000
TRAVEL	77,350	2,000	2,000
9/11	0	2,000	2,000
FACE-TO-FACE	0	2,000	2,000
LETTERS	0	2,000	2,000
OUP	0	2,000	2,000
VERBATIM	0	2,000	2,000
MultiNLI Overall	**392,702**	**20,000**	**20,000**

Bowman et al., EMNLP2015



Sentence model	Train	Test
100d Sum of words	79.3	75.3
100d RNN	73.1	72.2
100d LSTM RNN	84.8	**77.6**

Bowman et al., A Large Annotated Corpus for Learning Natural Language Inference, EMNLP2015

Conneau et al., EMNLP2017

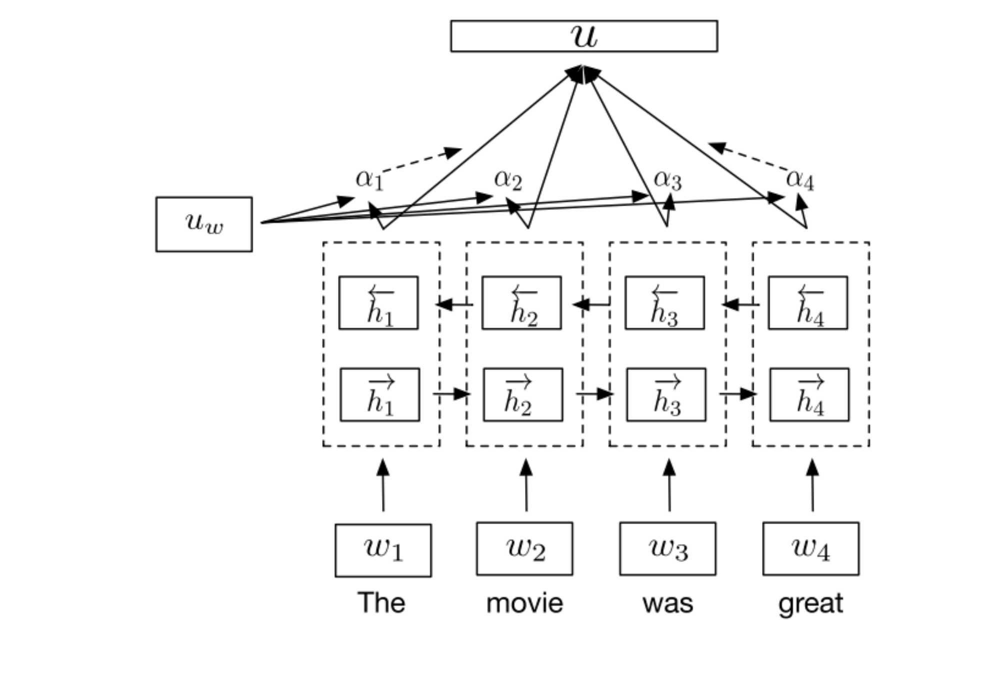


Figure 3: **Inner Attention network architecture.**

Lin et al., ICLR2017

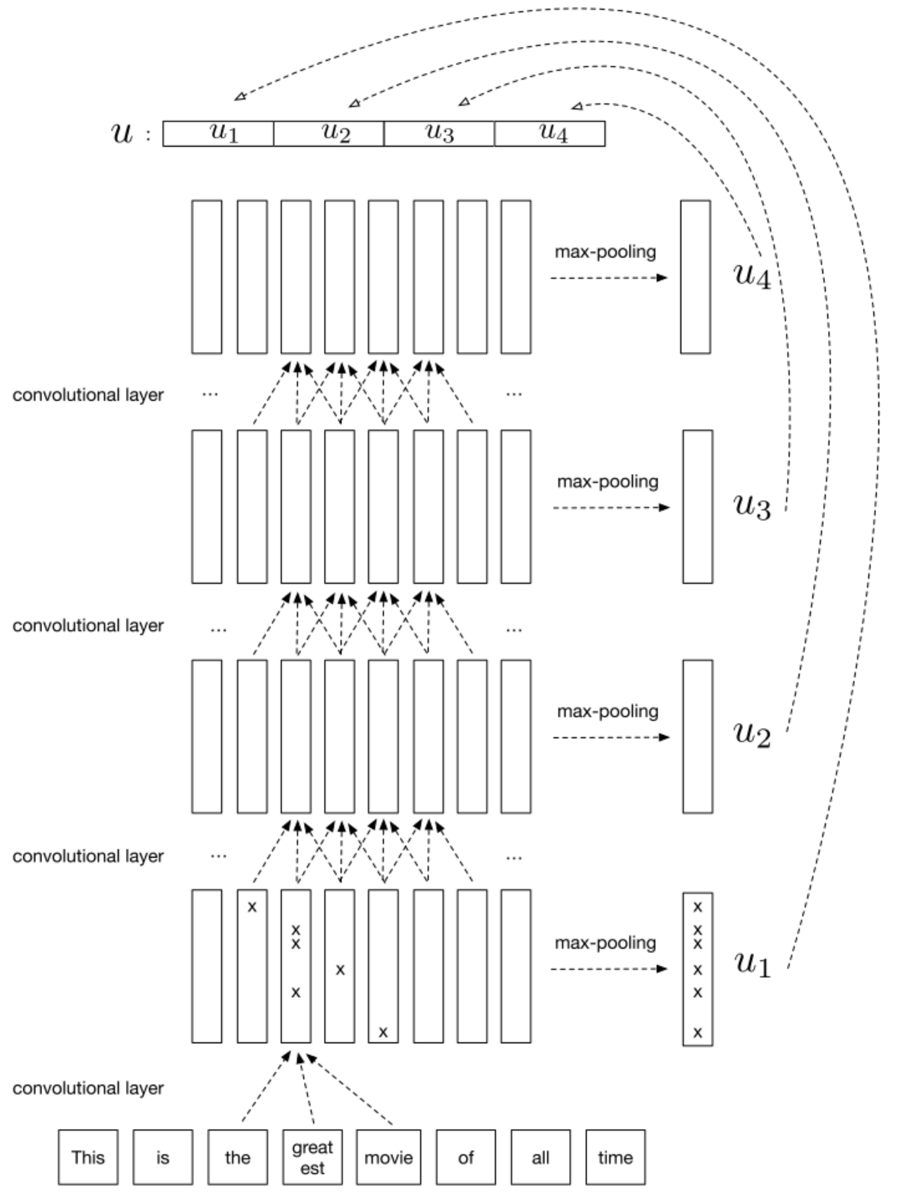
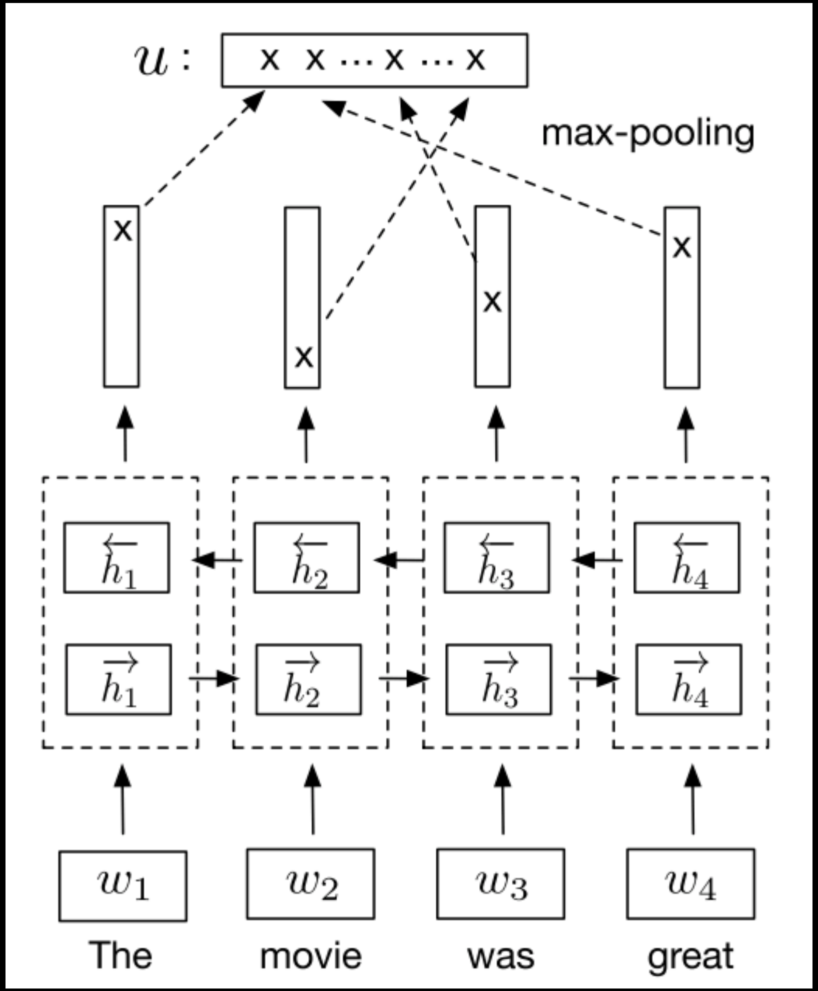


Figure 4: **Hierarchical ConvNet architecture.**

Zhao et al., IJCAI2015

Conneau et al., *Supervised Learning of Universal Sentence Representations from Natural Language Inference Data*, EMNLP2017

Bi-LSTM-Max



Model	dim	NLI dev	NLI test
LSTM	2048	81.9	80.7
GRU	4096	82.4	81.8
BiGRU-last	4096	81.3	80.9
BiLSTM-Mean	4096	79.0	78.2
Inner-attention	4096	82.3	82.5
HConvNet	4096	83.7	83.4
BiLSTM-Max	4096	**85.0**	**84.5**

Conneau et al., *Supervised Learning of Universal Sentence Representations from Natural Language Inference Data*, EMNLP2017

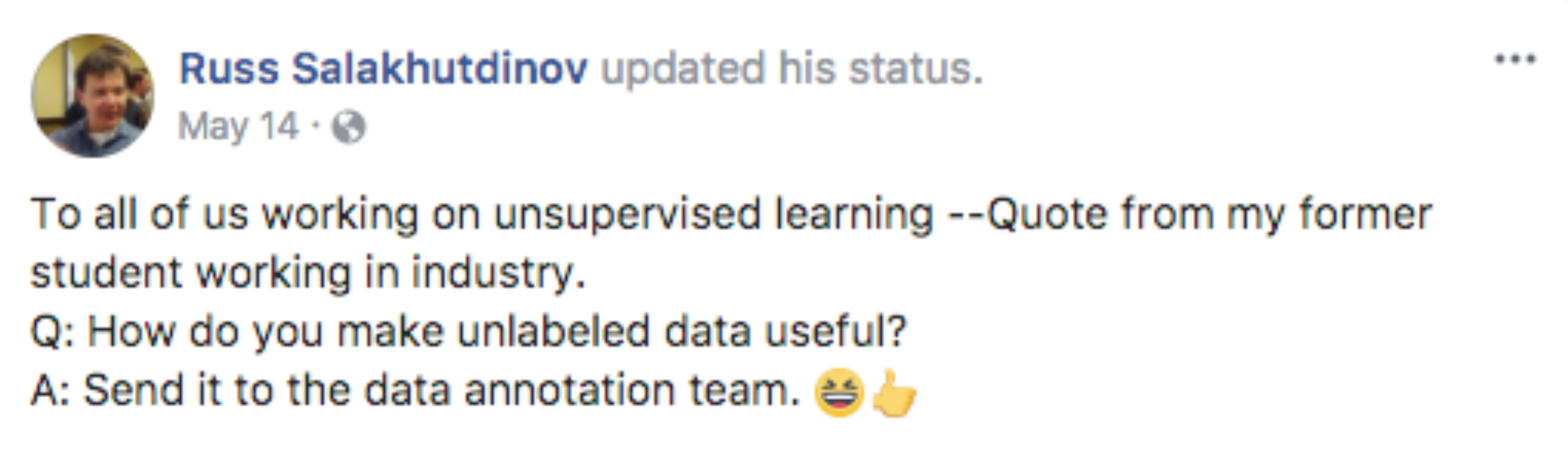
Learning Distributed Representations of Sentences

- … from unlabeled data (Context-based)
  - Generative Objective
  - Discriminative Objective
- … from labeled data
  - Natural Language Inference (NLI) datasets
- Evaluation

To conclude:

- Unlabeled data is enormous, thus building efficient algorithms for representation learning is critical.
- The usage of the context information is not sufficient, so we still need to come up with new ways of exploiting context information.
- There lacks a unified framework/guide for model design.
- Supervised transfer learning is promising, but labeling is costly and time-consuming.

To conclude again…


Russ Salakhutdinov updated his status.
May 14 ·

To all of us working on unsupervised learning --Quote from my former student working in industry.
Q: How do you make unlabeled data useful?
A: Send it to the data annotation team. 😆👆

Thank you!