UCSD Cognitive Science

Adobe

Trimming and Improving Skip-thought Vectors

--Representing sentences as vectors

Shuai Tang (唐帅)

Why?

Sentence $\longrightarrow$ Vector

Why?

Sentence → Vector

We communicate in sentences,
and they convey our thoughts.

Why?

Sentence $\longrightarrow$ Vector

Vector is an efficient type of representation for machines to operate on.

Why?

Sentence → 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

Machine Learning

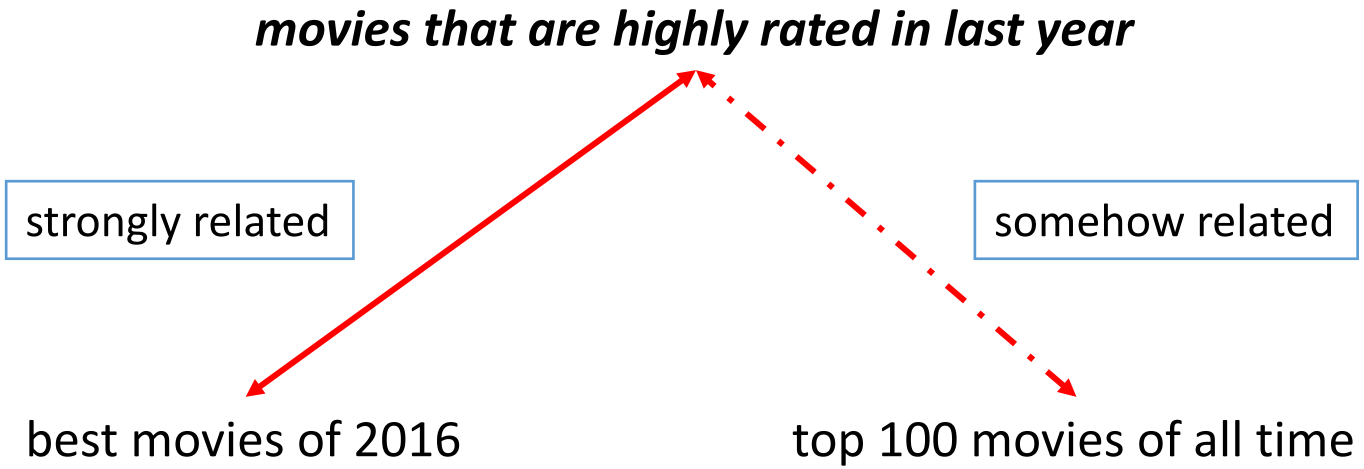
Sentence $\longrightarrow$ Vector

Learn from data!

Supervised Learning

Unsupervised Learning

Supervised Learning



Supervised Learning

labels

strongly related

somehow related

Goodfellow, Bengio & Courville. Deep Learning, 2017
Photo credit: Google Image

Unsupervised Learning

Sentence $\longrightarrow$ Vector

Learn from data!

Without labels!

Goodfellow, Bengio & Courville. Deep Learning, 2017

Existing Models

- Bag-of-words (BOW)
- Continuous Bag-of-words (CBOW)/ Skip-gram
- Sequence to Sequence (Seq2Seq)
- Skip-thought

Existing Models

- **Bag-of-words (BOW)**
  - **Harris, Word1954**
- Continuous Bag-of-words (CBOW)/ Skip-gram
- Sequence to Sequence (Seq2Seq)
- Skip-thought

Bag-of-words model (BOW)

- Corpus
  - i love you.
  - however, you don’t love me.
  - it is a sad story.
- Dictionary
  - {i, love, you, however, don’t, me, it, is, a, sad, story}
- Representations
  - [1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0]
  - [0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0]
  - [0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1]

Bag-of-words model (BOW)

- Corpus
  - i love you.
  - however, you don’t love me.
  - it is a sad story.
- Dictionary
  - {i, love, you, however, don’t, me, it, is, a, sad, story}
- Representations
  - [1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0]
  - [0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0]
  - [0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1]

Bag-of-words model (BOW)

- Corpus
  - i love you.
  - however, you don’t love me.
  - it is a sad story.
- Dictionary
  - {i, love, you, however, don’t, me, it, is, a, sad, story}
- Representations
  - [1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0]
  - [0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0]
  - [0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1]

Bag-of-words model (BOW)

- Corpus
  - i love you.
  - however, you don't love me.
  - it is a sad story.
- Dictionary
  - {i, love, you, however, don't, me, it, is, a, sad, story}
- Representations
  - [1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0]
  - [0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0]
  - [0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1]

Fast!

Bag-of-words model (BOW)

- Corpus
  - i love you.
  - however, you don’t love me.
  - it is a sad story.
- Dictionary
  - {i, love, you, however, don’t, me, it, is, a, sad, story}
- Representations
  - [1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0]
  - [0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0]
  - [0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1]

Existing Models

- Bag-of-words (BOW)
- **Continuous Bag-of-words (CBOW)/ Skip-gram**
  - **Mikolov et al., NIPS2013**
- Sequence to Sequence (Seq2Seq)
- Skip-thought

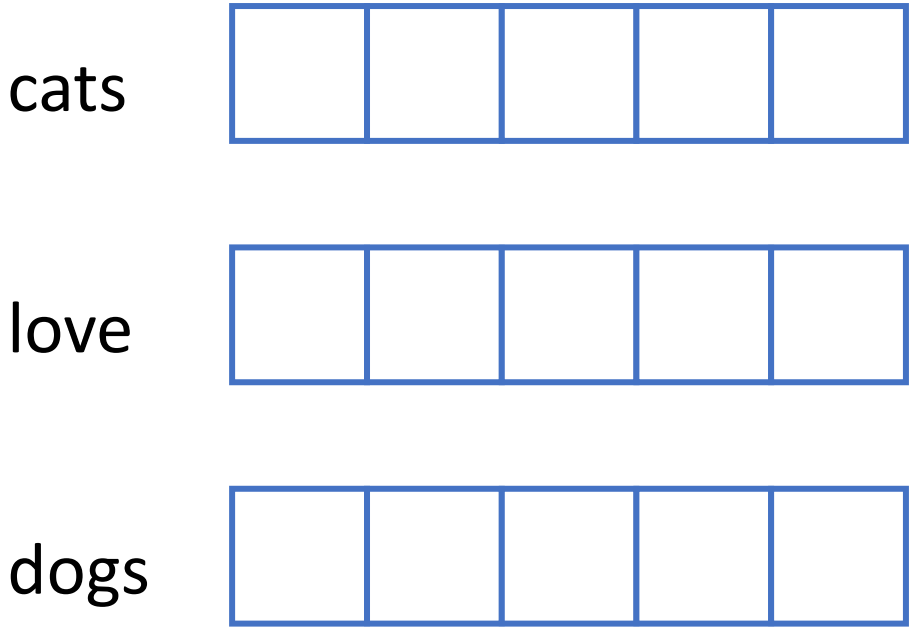
Continuous BOW/ Skip-gram (Mikolov et al. NIPS2013)

cats

love

dogs

Continuous BOW/ Skip-gram (Mikolov et al. NIPS2013)



Continuous BOW/ Skip-gram (Mikolov et al. NIPS2013)

cats	0.4	-0.5	0	-0.1	0.2
love	0	0.1	0.9	-0.3	-0.4
dogs	0.5	-0.4	0.1	-0.2	0.3

Continuous BOW/ Skip-gram (Mikolov et al. NIPS2013)

cats	0.4	-0.5	0	-0.1	0.2
love	0	0.1	0.9	-0.3	-0.4
dogs	0.5	-0.4	0.1	-0.2	0.3
cats love dogs	0.30	-0.27	0.33	-0.20	0.03

Continuous BOW/ Skip-gram (Mikolov et al. NIPS2013)

cats	0.4	-0.5	0	-0.1	0.2
					+
love	0	0.1	0.9	-0.3	-0.4
					+
dogs	0.5	-0.4	0.1	-0.2	0.3
cats love dogs	0.30	-0.27	0.33	-0.20	0.03

Fixed Manually

Continuous BOW/ Skip-gram (Mikolov et al. NIPS2013)

dogs love cats

0.30	-0.27	0.33	-0.20	0.03

Same!

cats love dogs

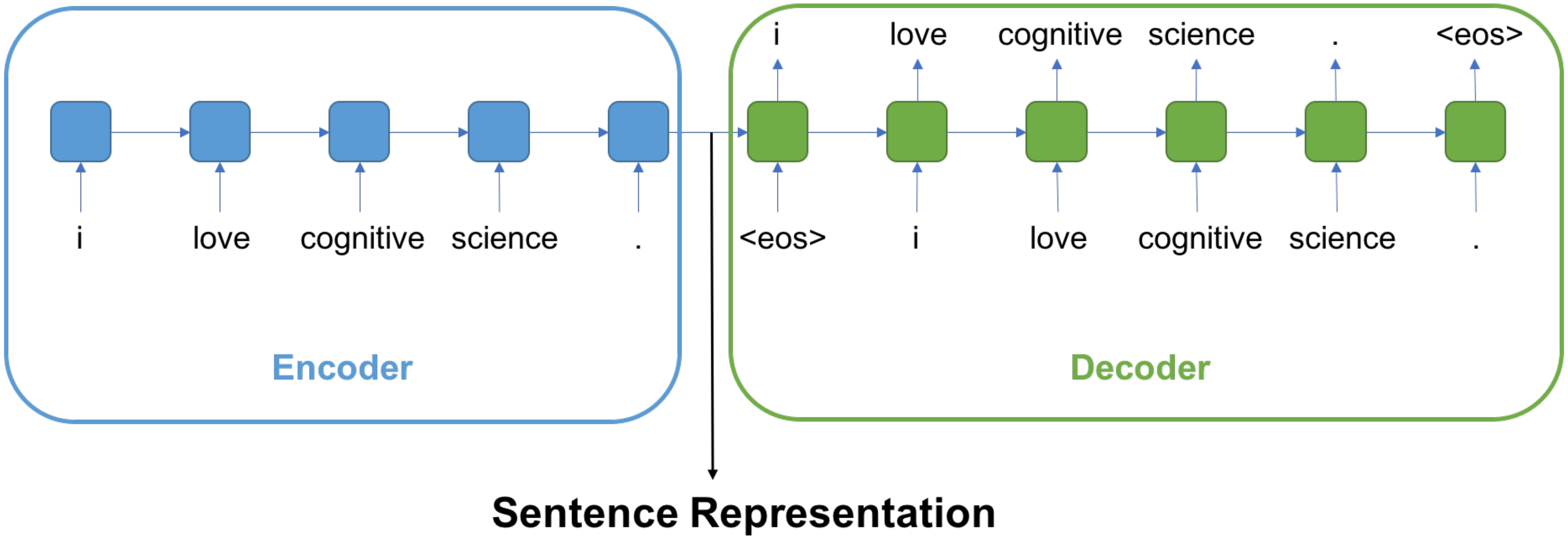
0.30	-0.27	0.33	-0.20	0.03

Existing Models

- Bag-of-words (BOW)
- Continuous Bag-of-words (CBOW)/ Skip-gram
- **Sequence to Sequence (Seq2Seq)**
  - **Sutskever, Vinyals & Le (NIPS2014)**
  - **Dai & Le (NIPS2015)**
- Skip-thought

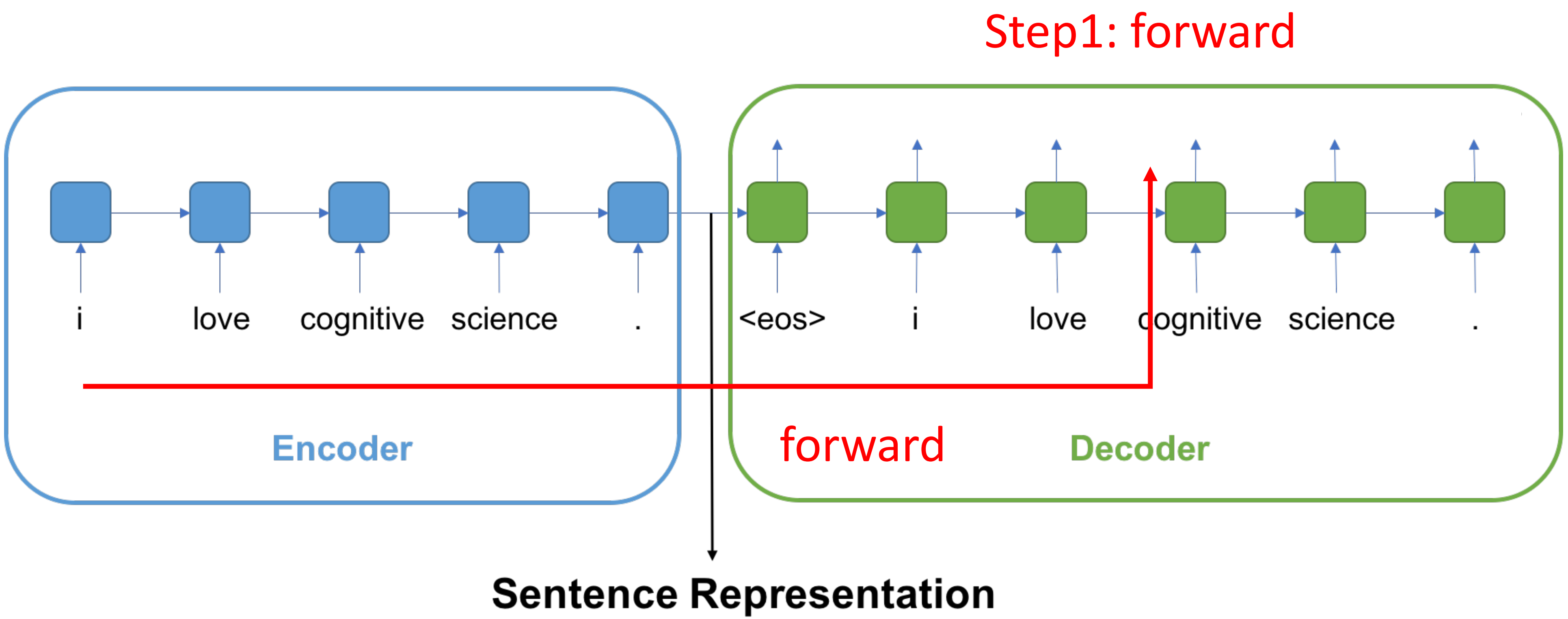
Sequence to Sequence (Dai & Le, NIPS2015)

- With Recurrent Neural Networks



Sequence to Sequence (Dai & Le, NIPS2015)

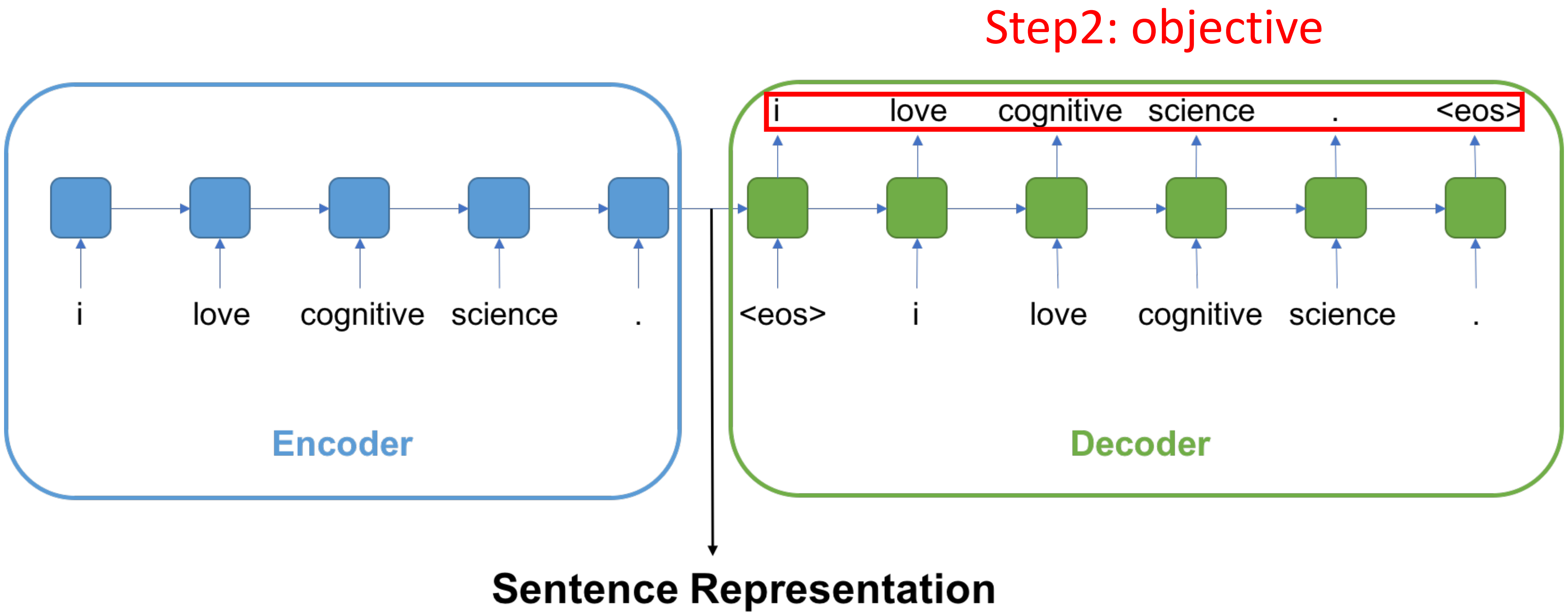
- Each training iteration...



Goodfellow, Bengio & Courville. Deep Learning, 2017

Sequence to Sequence (Dai & Le, NIPS2015)

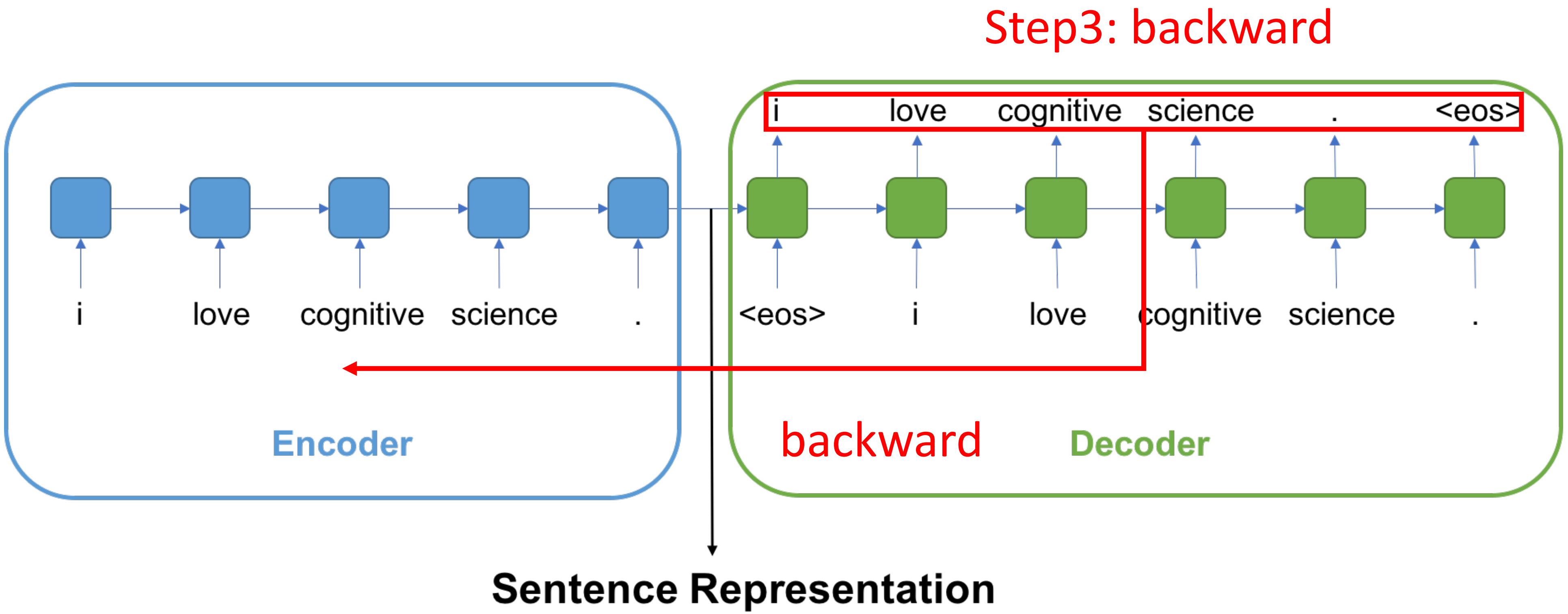
- Each training iteration…



Goodfellow, Bengio & Courville. Deep Learning, 2017

Sequence to Sequence (Dai & Le, NIPS2015)

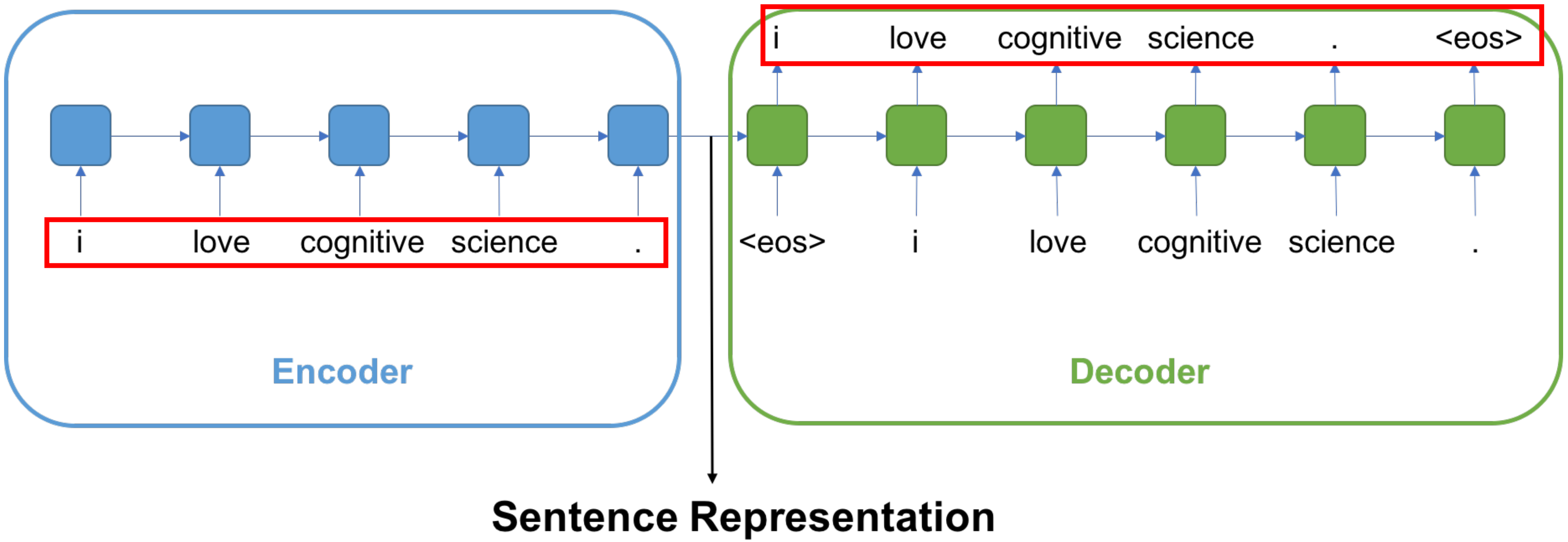
- Each training iteration…



Goodfellow, Bengio & Courville. Deep Learning, 2017

Sequence to Sequence (Dai & Le, NIPS2015)

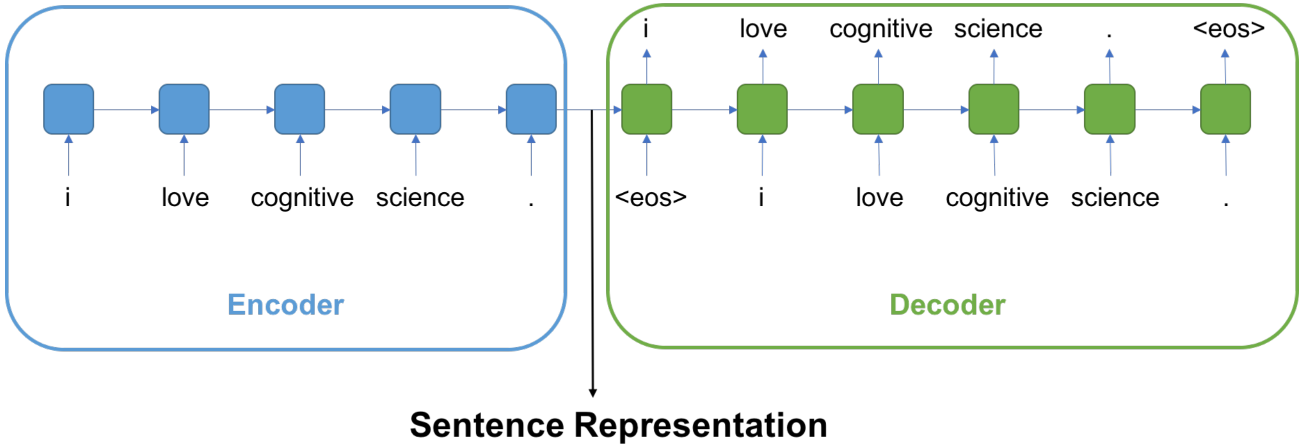
- With Recurrent Neural Networks



Sequence to Sequence (Dai & Le, NIPS2015)

- Pros:
  - Word-order information is utilized in training.

- Cons:
  - Training is slow.



Existing Models

- Bag-of-words (BOW)
- Continuous Bag-of-words (CBOW)/ Skip-gram
- Sequence to Sequence (Seq2Seq)
- **Skip-thought**
  - **Kiros et al., NIPS2015**

Trimming and Improving Skip-thought Vectors

- **Skip-thought**
  - **Kiros et al., NIPS2015**
- Our hypotheses to improve skip-thought
- Comparison between our trimmed skip-thought model and the skip-thought model
- Conclusion

Skip-thought (Kiros et al., NIPS2015)

- Skip-thought model is for learning a generic sentence **encoder** .

Encoder - Decoder

$$\textbf{Sentence} \xrightarrow{\boldsymbol{encode}} \textbf{Representation} \xrightarrow{\boldsymbol{decode}} \textbf{Sentences}$$

Skip-thought (Kiros et al., NIPS2015)

- The skip-thought model learns to encode a sentence, and decode its surrounding two sentences, instead of itself.

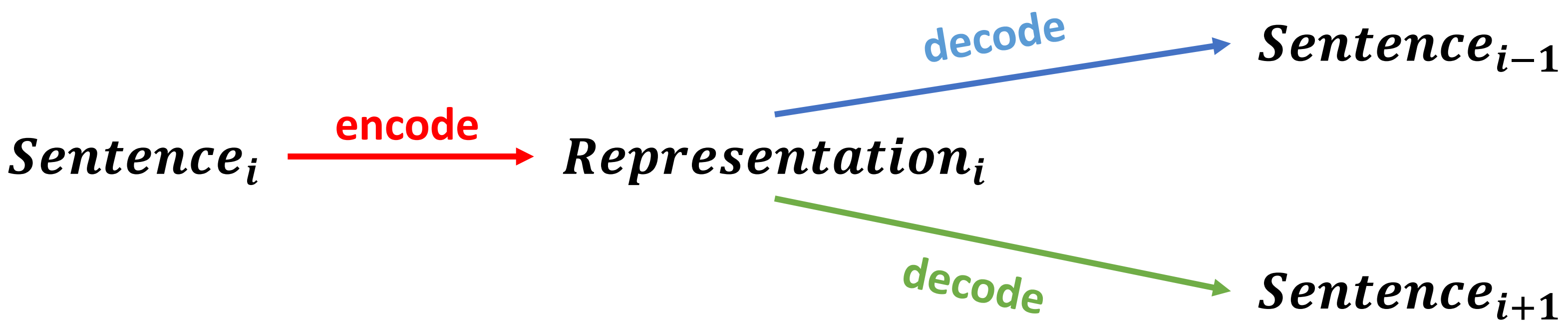
$$
\begin{tikzcd}
 & & \boldsymbol{Sentence_{i-1}} \\
\boldsymbol{Sentence_i} \arrow[r, "\text{encode}"] & \boldsymbol{Representation_i} \arrow[ur, "\text{decode}"] \arrow[dr, "\text{decode}"'] & \\
 & & \boldsymbol{Sentence_{i+1}}
\end{tikzcd}
$$

The context in which words and sentences are understood plays an important role in human comprehension. (Altmann & Steedman, 1988; Binder & Desai, 2011)

Skip-thought (Kiros et al., NIPS2015)

- The model contains
  - an encoder
  - a previous decoder
  - a next decoder

3 **parametric** functions needs to be learned



Skip-thought (Kiros et al., NIPS2015)

- Given a sentence tuple
    - i love you.
    - however, you don't love me.
    - it is a sad story.
- Detailed encoding schemes
    - Uni-skip/ Bi-skip/ Combine-skip

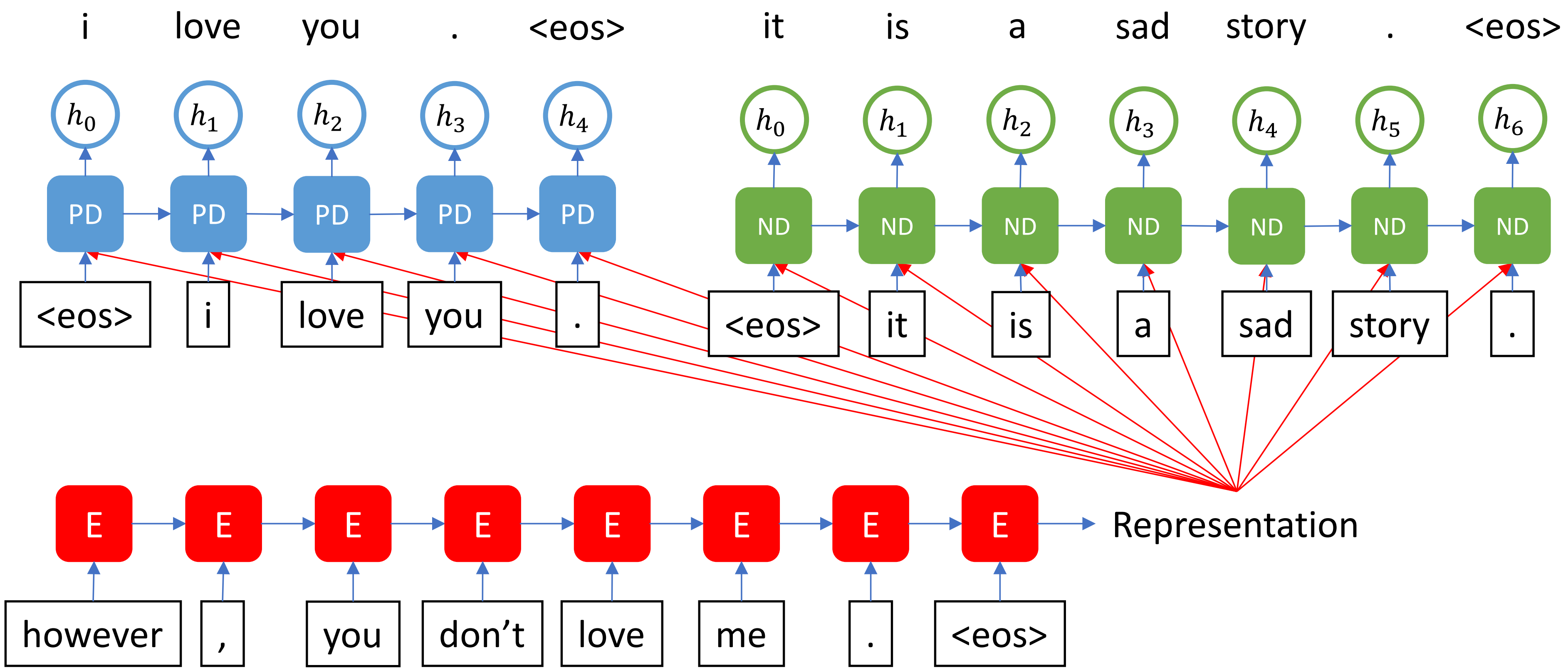
Skip-thought (Kiros et al., NIPS2015)

- Given a sentence tuple

  - i love you.
  - however, you don't love me.
  - it is a sad story.

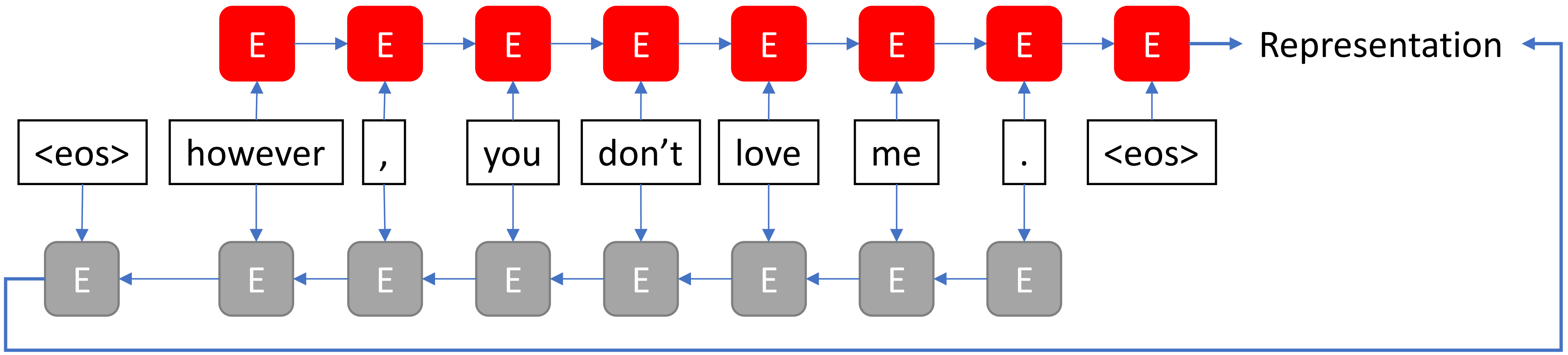
- Detailed encoding schemes
  - Uni-skip/ Bi-skip/ Combine-skip

Uni-Skip



i
love
you
.
<eos>
it
is
a
sad
story
.
<eos>
h_0
h_1
h_2
h_3
h_4
h_0
h_1
h_2
h_3
h_4
h_5
h_6
PD
ND
<eos>
i
love
you
.
<eos>
it
is
a
sad
story
.
E
Representation
however
,
you
don’t
love
me
.
<eos>

Bi-Skip



Combine-Skip

uni-skip

concatenate → representation

bi-skip

Trimming and Improving Skip-thought Vectors

- Skip-thought (Kiros et al., NIPS2015)
- Our hypotheses to improve skip-thought
  - **Neighborhood hypothesis**
  - Average+Max Connection
  - Word Vectors Initialization
- Comparison between our trimmed skip-thought model and the skip-thought model
- Conclusion

Neighborhood Hypothesis

- Does this model really need a previous decoder and a next decoder?

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

Neighborhood Hypothesis

- Skip-thought model

$$\begin{tikzcd}[column sep=large] & & \boldsymbol{Sentence_{i-1}} \\ \boldsymbol{Sentence_i} \arrow[r, "\text{encode}"] & \boldsymbol{Representation_i} \arrow[ur, "\text{decode}"] \arrow[dr, "\text{decode}"'] & \\ & & \boldsymbol{Sentence_{i+1}} \end{tikzcd}$$

- Neighborhood Hypothesis

$$\begin{tikzcd}[column sep=large] & & \boldsymbol{Sentence_{i-1}} \\ \boldsymbol{Sentence_i} \arrow[r, "\text{encode}"] & \boldsymbol{Representation_i} \arrow[ur, "\text{decode}"] \arrow[dr, "\text{decode}"'] & \\ & & \boldsymbol{Sentence_{i+1}} \end{tikzcd}$$

Neighborhood Hypothesis

- Can we further simplify the skip-thought model?

Yes!

Neighborhood Hypothesis

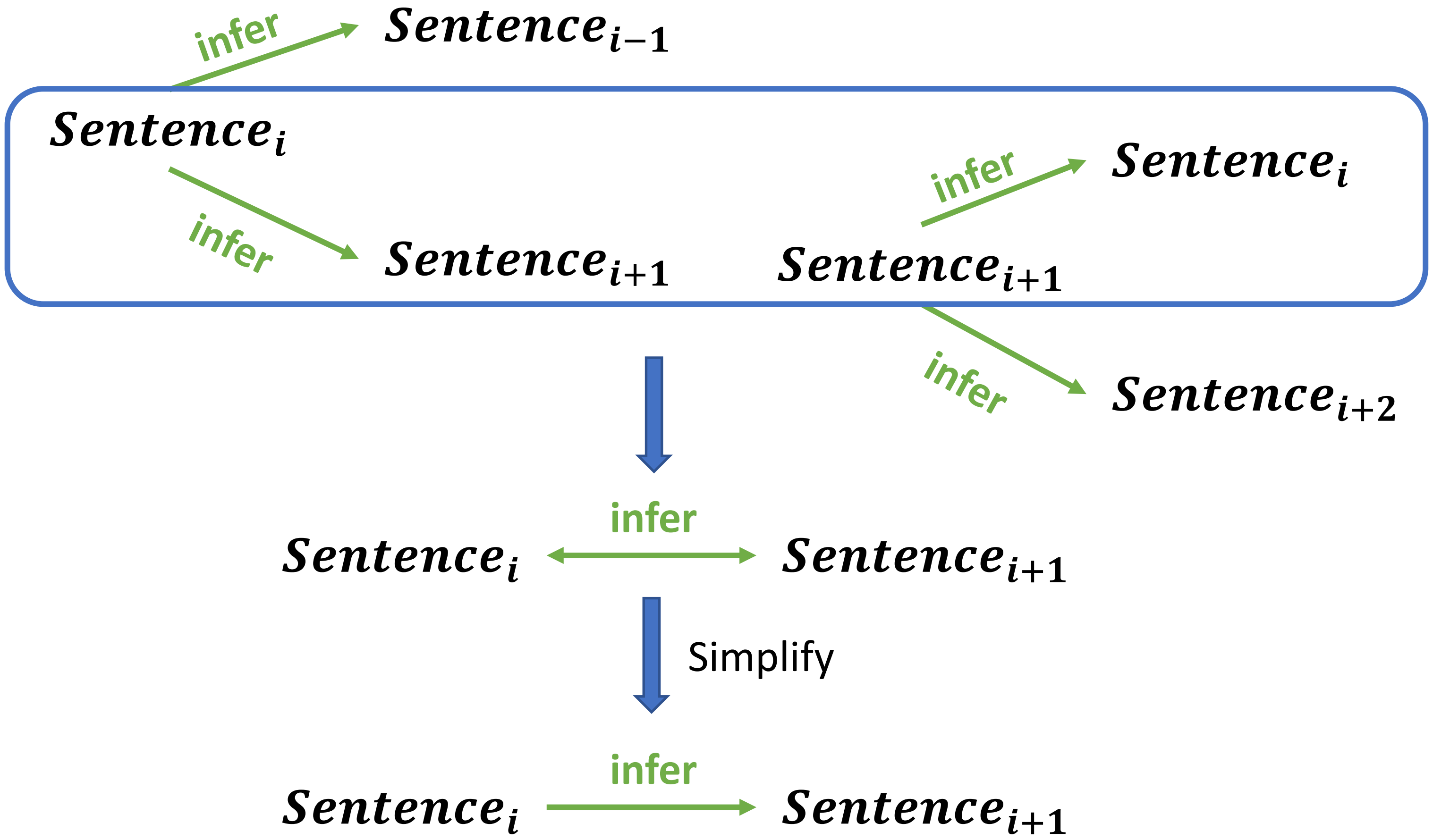
$$\boldsymbol{Sentence_i} \xrightarrow{\text{infer}} \boldsymbol{Sentence_{i-1}}$$

$$\boldsymbol{Sentence_i} \xrightarrow{\text{infer}} \boldsymbol{Sentence_{i+1}}$$

$$\boldsymbol{Sentence_{i+1}} \xrightarrow{\text{infer}} \boldsymbol{Sentence_i}$$

$$\boldsymbol{Sentence_{i+1}} \xrightarrow{\text{infer}} \boldsymbol{Sentence_{i+2}}$$

Neighborhood Hypothesis



Neighborhood Hypothesis

- Skip-thought Model

$$\begin{array}{l} \nearrow^{\text{decode}} \; \boldsymbol{Sentence_{i-1}} \\ \boldsymbol{Sentence_i} \xrightarrow{\text{encode}} \boldsymbol{Representation_i} \\ \searrow_{\text{decode}} \; \boldsymbol{Sentence_{i+1}} \end{array}$$

- Our **Trimmed** Skip-thought Model

$$\boldsymbol{Sentence_i} \xrightarrow{\text{encode}} \boldsymbol{Representation_i} \xrightarrow{\text{decode}} \boldsymbol{Sentence_{i+1}}$$

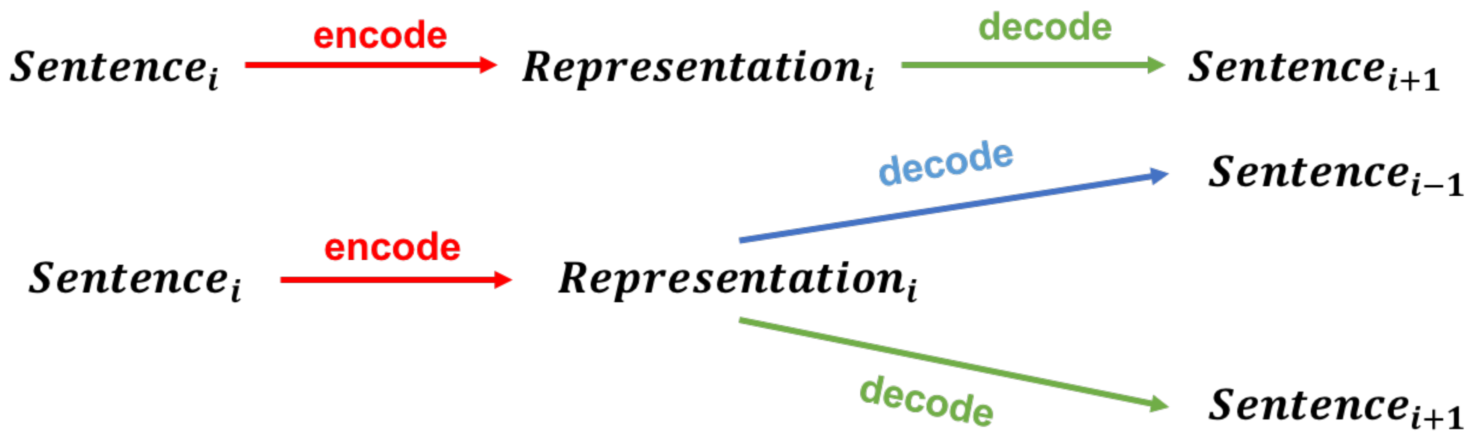
Neighborhood Hypothesis

- **BookCorpus** dataset (Zhu et al., ICCV2015)
  - 74 million contiguous sentences from 7,000 books

Encoder - Decoder

- Then, the sentence **encoder** was evaluated on 7 natural language processing (NLP) tasks.

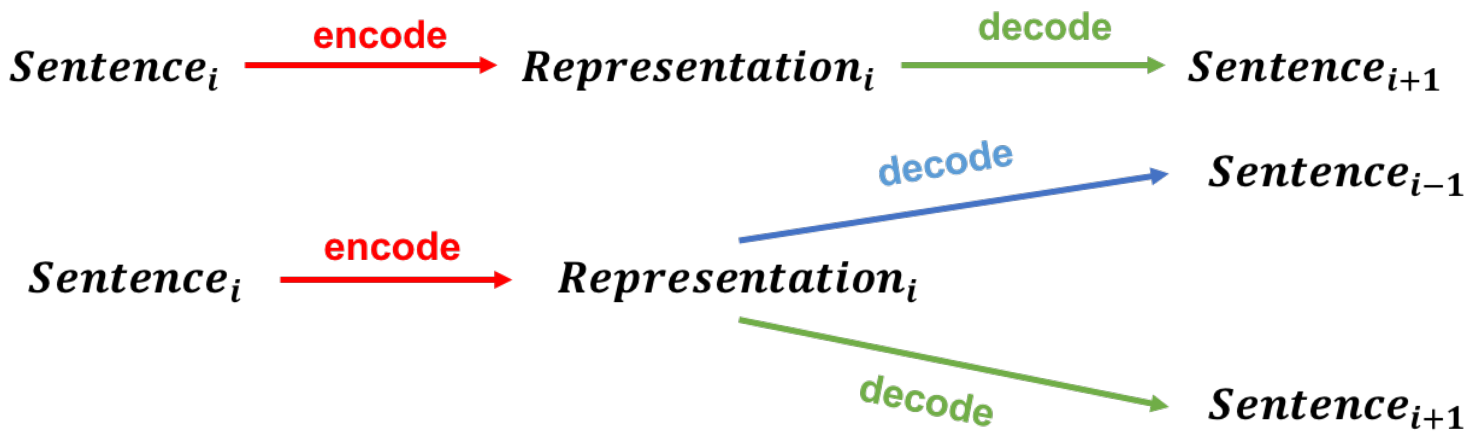
Neighborhood Hypothesis



Relationship of sentence pair — Classification on single sentence

Model	WE	SICK r	SICK ρ	SICK MSE	MSRP (Acc/F1)	MR	CR	SUBJ	MPQA	TREC
Plain Connection										
bi-T-skip	word2vec	0.8408	0.7649	0.2994	75.3 / **83.0**	76.1	80.3	92.3	87.5	86.6
uni-T-skip		0.8349	0.7629	0.3084	73.7 / 81.9	75.7	82.1	91.3	87.4	86.4
C-T-skip		**0.8518**	**0.7808**	**0.2802**	**75.7 / 83.0**	76.8	**83.2**	**92.8**	**88.4**	87.5
bi-skip	word2vec	0.8385	0.7618	0.3028	73.9 / 82.0	75.7	81.4	92.1	87.2	88.4
uni-skip		0.8344	0.7586	0.3098	73.6 / 81.6	76.2	81.8	92.2	87.6	87.0
C-skip		0.8492	0.7738	0.2844	74.6 / 82.3	**77.0**	83.0	92.7	87.9	**89.2**

Neighborhood Hypothesis

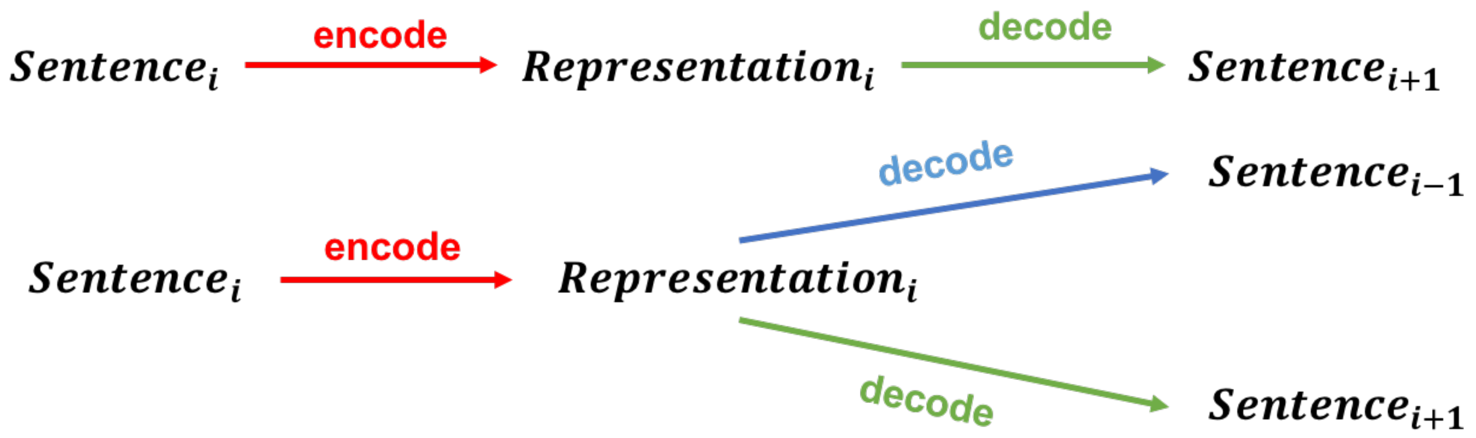


Relationship of sentence pair

Classification on single sentence

Model	WE	SICK			MSRP (Acc/F1)	MR	CR	SUBJ	MPQA	TREC
		r	ρ	MSE						
Plain Connection										
bi-T-skip	word2vec	0.8408	0.7649	0.2994	75.3 / **83.0**	76.1	80.3	92.3	87.5	86.6
uni-T-skip		0.8349	0.7629	0.3084	73.7 / 81.9	75.7	82.1	91.3	87.4	86.4
C-T-skip		**0.8518**	**0.7808**	**0.2802**	**75.7 / 83.0**	76.8	**83.2**	**92.8**	**88.4**	87.5
bi-skip	word2vec	0.8385	0.7618	0.3028	73.9 / 82.0	75.7	81.4	92.1	87.2	88.4
uni-skip		0.8344	0.7586	0.3098	73.6 / 81.6	76.2	81.8	92.2	87.6	87.0
C-skip		0.8492	0.7738	0.2844	74.6 / 82.3	**77.0**	83.0	92.7	87.9	**89.2**

Neighborhood Hypothesis



Relationship of sentence pair

Classification on single sentence

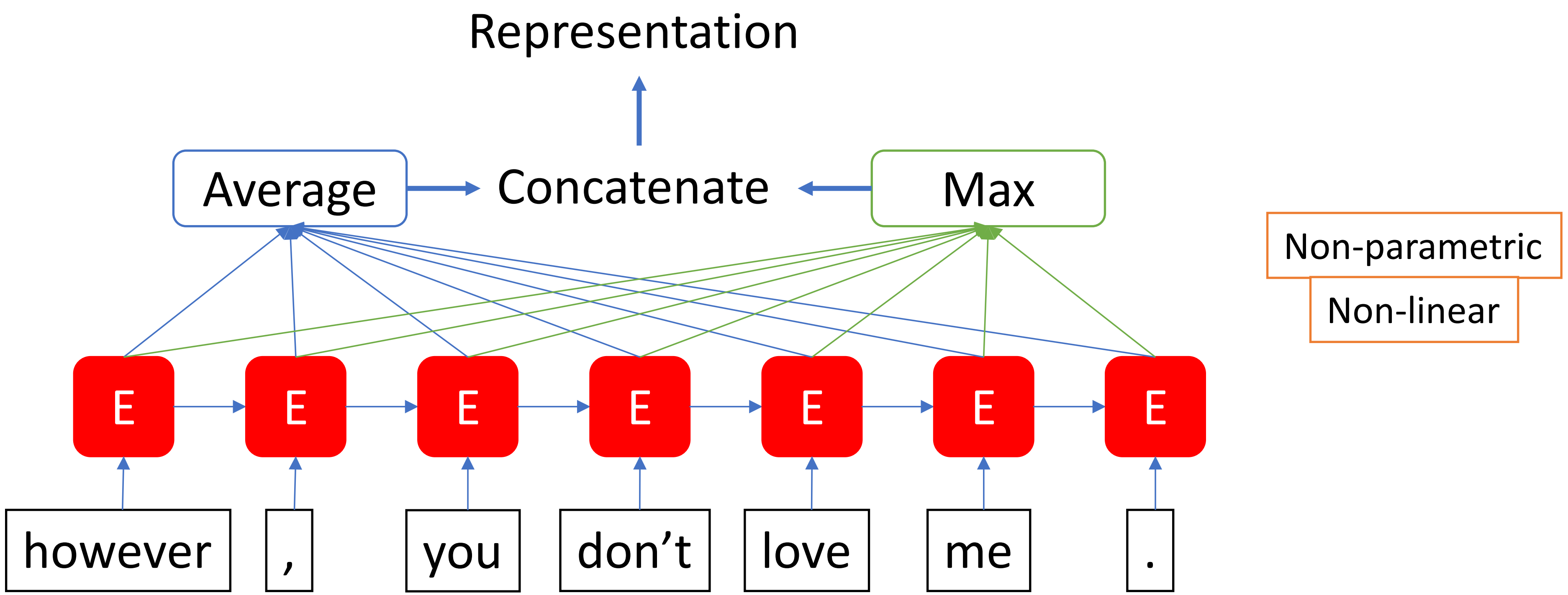
Model	WE	SICK			MSRP (Acc/F1)	MR	CR	SUBJ	MPQA	TREC
		r	ρ	MSE						
Plain Connection										
bi-T-skip	word2vec	0.8408	0.7649	0.2994	75.3 / **83.0**	76.1	80.3	92.3	87.5	86.6
uni-T-skip	word2vec	0.8349	0.7629	0.3084	73.7 / 81.9	75.7	82.1	91.3	87.4	86.4
C-T-skip	word2vec	**0.8518**	**0.7808**	**0.2802**	**75.7 / 83.0**	76.8	**83.2**	**92.8**	**88.4**	87.5
bi-skip	word2vec	0.8385	0.7618	0.3028	73.9 / 82.0	75.7	81.4	92.1	87.2	88.4
uni-skip	word2vec	0.8344	0.7586	0.3098	73.6 / 81.6	76.2	81.8	92.2	87.6	87.0
C-skip	word2vec	0.8492	0.7738	0.2844	74.6 / 82.3	**77.0**	83.0	92.7	87.9	**89.2**

Trimming and Improving Skip-thought Vectors

- Skip-thought
- Our hypotheses to improve skip-thought
  - Neighborhood hypothesis
  - **Average+Max Connection**
  - Word Vector Initialization
- Comparison between our trimmed skip-thought model and the skip-thought model
- Conclusion

Average+Max Connection

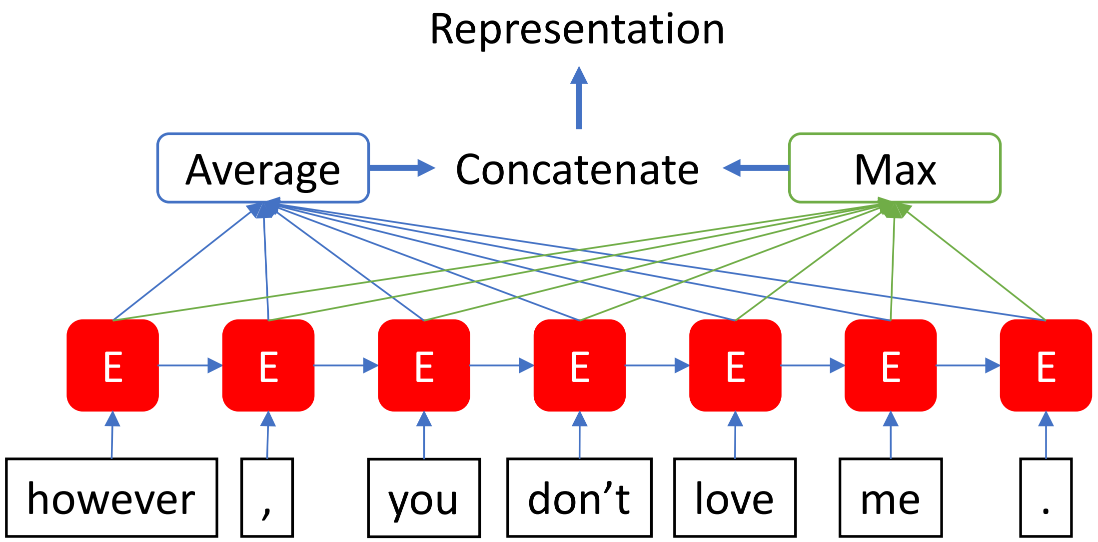
- Average+Max Connection (Chen et al., arXiv2017)



Average+Max Connection

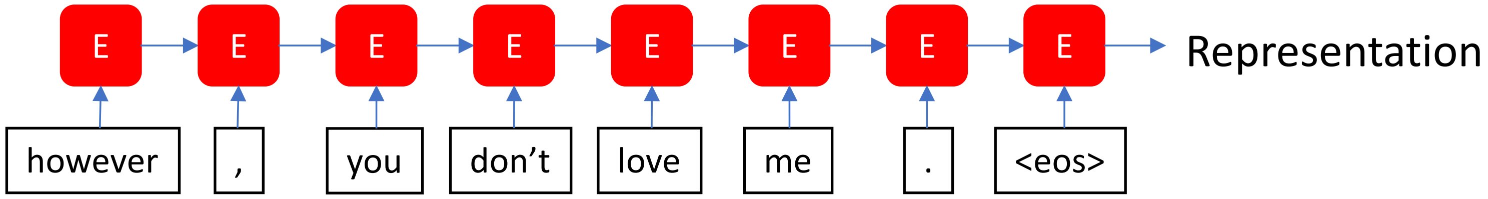
- Plain Connection (Kiros et al., NIPS2015)

- Average+Max Connection (Chen et al., arXiv2017)

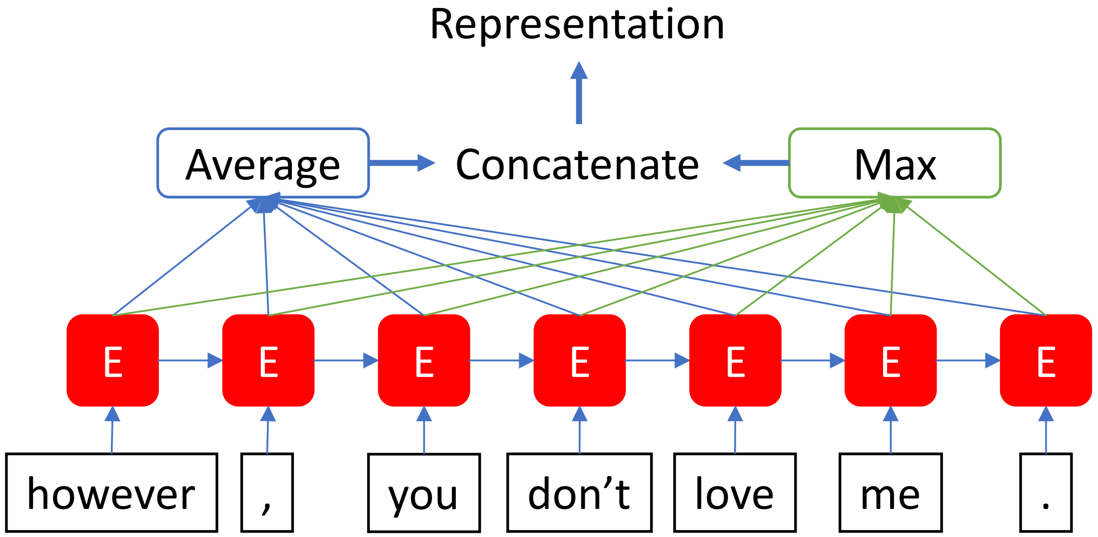


Average+Max Connection

- Plain Connection (Kiros et al., NIPS2015)



- Average+Max Connection (Chen et al., arXiv2017)



Average+Max Connection

- The average+max connection was proposed for **supervised** tasks, and specifically for Stanford Natural Language Inference corpus.
  - (Chen et al., arXiv2017; Bowman et al., EMNLP2015)

supervised → **unsupervised**

- We hypothesize that, our trimmed skip-thought model could also **benefit** from the average+max connection .

Average+Max Connection

- The average+max connection was proposed for **supervised** tasks, and specifically for Stanford Natural Language Inference corpus.
  - (Chen et al., arXiv2017; Bowman et al., EMNLP2015)

supervised $\longrightarrow$ **unsupervised**

- We hypothesize that, our trimmed skip-thought model could also **benefit** from the average+max connection .

Average+Max Connection

		Relationship of sentence pair				Classification on single sentence				
Model	WE	SICK			MSRP (Acc/F1)	MR	CR	SUBJ	MPQA	TREC
		r	ρ	MSE						
Plain Connection										
bi-T-skip	word2vec	0.8408	0.7649	0.2994	75.3 / **83.0**	76.1	80.3	92.3	87.5	86.6
uni-T-skip		0.8349	0.7629	0.3084	73.7 / 81.9	75.7	82.1	91.3	87.4	86.4
C-T-skip		0.8518	0.7808	0.2802	**75.7 / 83.0**	**76.8**	**83.2**	**92.8**	**88.4**	87.5
Average+Max Connection										
bi-T-skip	word2vec	0.8463	0.7744	0.2894	73.3 / 81.6	74.4	78.6	91.3	86.2	88.8
uni-T-skip		0.8466	0.7705	0.2884	74.0 / 81.7	73.0	78.6	91.3	85.2	88.4
C-T-skip		**0.8598**	**0.7892**	**0.2654**	75.0 / 82.2	75.1	80.0	92.2	87.2	**90.0**

Average+Max Connection

Relationship of sentence pair | Classification on single sentence

Model	WE	SICK r	SICK ρ	SICK MSE	MSRP (Acc/F1)	MR	CR	SUBJ	MPQA	TREC
Plain Connection										
bi-T-skip	word2vec	0.8408	0.7649	0.2994	75.3 / **83.0**	76.1	80.3	92.3	87.5	86.6
uni-T-skip	word2vec	0.8349	0.7629	0.3084	73.7 / 81.9	75.7	82.1	91.3	87.4	86.4
C-T-skip	word2vec	0.8518	0.7808	0.2802	**75.7 / 83.0**	**76.8**	**83.2**	**92.8**	**88.4**	87.5
Average+Max Connection										
bi-T-skip	word2vec	0.8463	0.7744	0.2894	73.3 / 81.6	74.4	78.6	91.3	86.2	88.8
uni-T-skip	word2vec	0.8466	0.7705	0.2884	74.0 / 81.7	73.0	78.6	91.3	85.2	88.4
C-T-skip	word2vec	**0.8598**	**0.7892**	**0.2654**	75.0 / 82.2	75.1	80.0	92.2	87.2	**90.0**

Average+Max Connection

Relationship of sentence pair

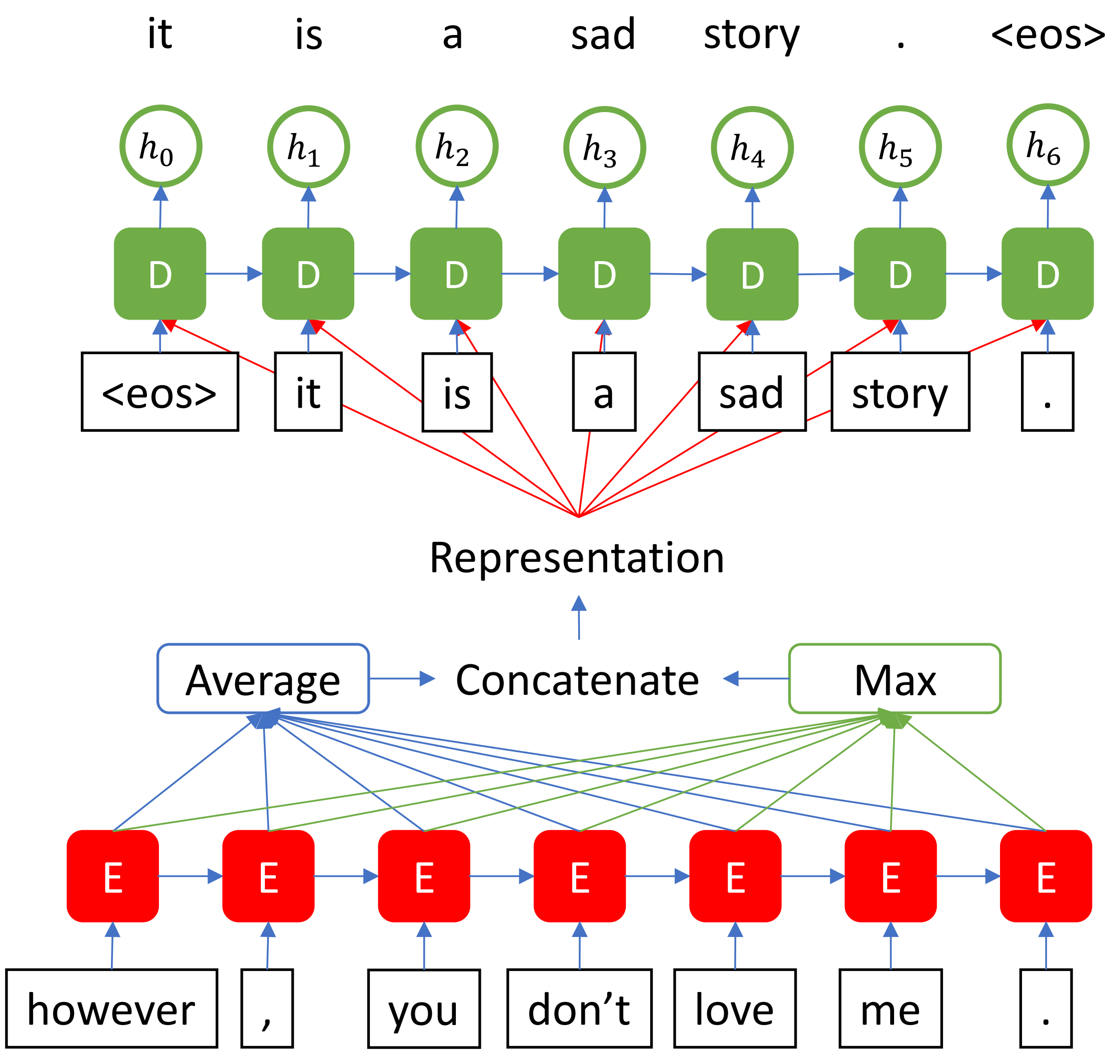
Classification on single sentence

Model	WE	SICK			MSRP (Acc/F1)	MR	CR	SUBJ	MPQA	TREC
		r	ρ	MSE						
Plain Connection										
bi-T-skip	word2vec	0.8408	0.7649	0.2994	75.3 / **83.0**	76.1	80.3	92.3	87.5	86.6
uni-T-skip		0.8349	0.7629	0.3084	73.7 / 81.9	75.7	82.1	91.3	87.4	86.4
C-T-skip		0.8518	0.7808	0.2802	**75.7 / 83.0**	**76.8**	**83.2**	**92.8**	**88.4**	87.5
Average+Max Connection										
bi-T-skip	word2vec	0.8463	0.7744	0.2894	73.3 / 81.6	74.4	78.6	91.3	86.2	88.8
uni-T-skip		0.8466	0.7705	0.2884	74.0 / 81.7	73.0	78.6	91.3	85.2	88.4
C-T-skip		**0.8598**	**0.7892**	**0.2654**	75.0 / 82.2	75.1	80.0	92.2	87.2	**90.0**

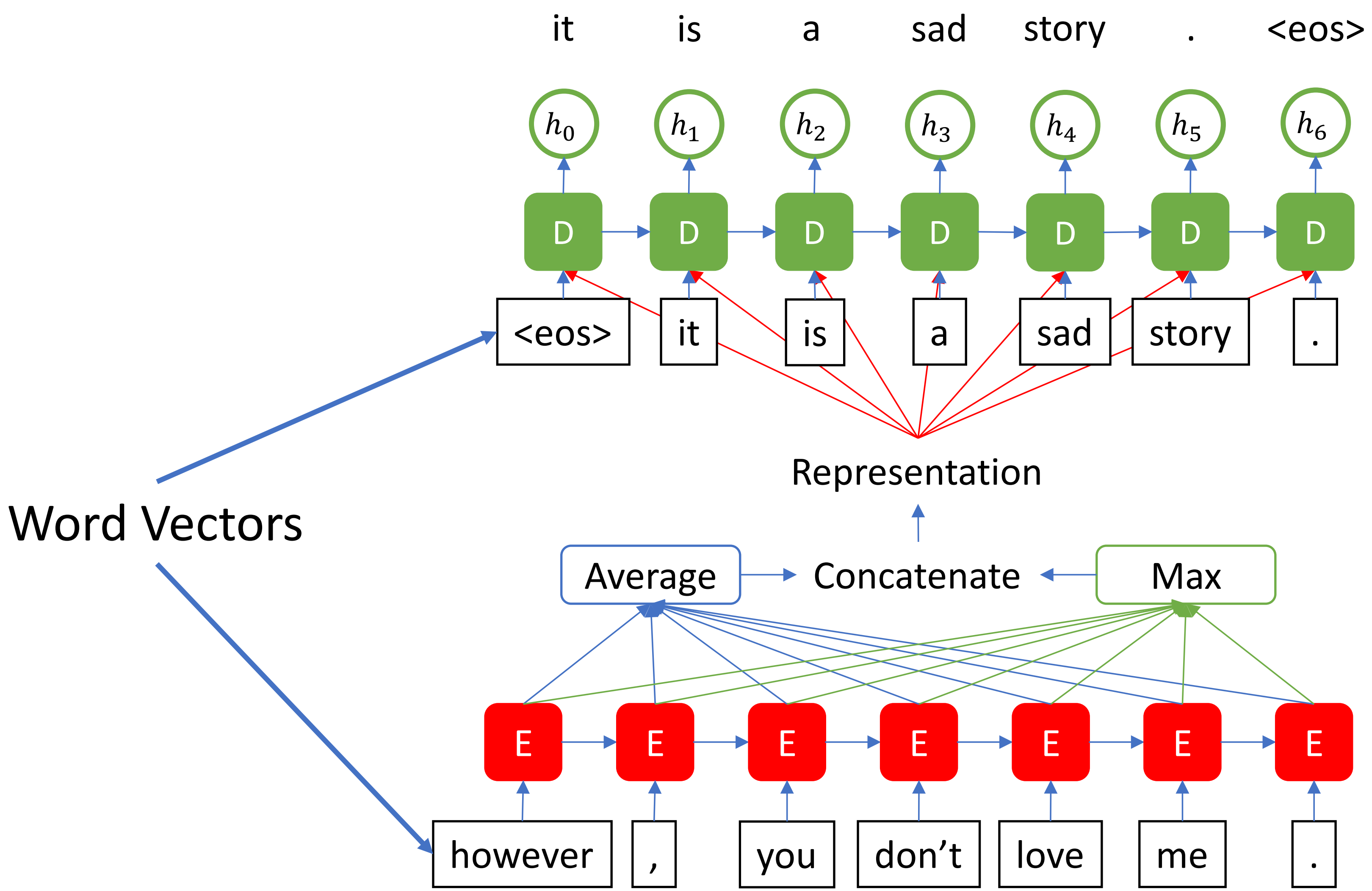
Trimming and Improving Skip-thought Vectors

- Skip-thought
- Our hypotheses to improve skip-thought
  - Neighborhood hypothesis
  - Average+Max Connection
  - **Word Vectors Initialization**
- Comparison between our trimmed skip-thought model and the skip-thought model
- Conclusion

Word Vectors Initialization



Word Vectors Initialization



Word Vectors Initialization

- Pretrained word vectors usually help the deep learning models to perform better on **supervised** tasks. (Collobert et al., JMLR2011)
    - word2vec (Mikolov et al., NIPS2013)
    - GloVe (Pennington et al., EMNLP2014)

- We hypothesize that, initializing our model with **pretrained word vectors** could also help the model to learn better sentence representation.

Word Vectors Initialization

- Pretrained word vectors usually help the deep learning models to perform better on **supervised** tasks. (Collobert et al., JMLR2011)
  - word2vec (Mikolov et al., NIPS2013)
  - GloVe (Pennington et al., EMNLP2014)

- We hypothesize that, initializing our model with **pretrained word vectors** could also help the model to learn better sentence representation.

Word Vectors Initialization

- Based on our trimmed skip-thought model with Average+Max connection, we implement models with 3 initializations.

Model	WE	SICK			MSRP (Acc/F1)	MR	CR	SUBJ	MPQA	TREC
		r	ρ	MSE						
		Average+Max Connection								
bi-T-skip	random	0.8336	0.7612	0.3112	73.2 / 81.3	69.7	76.0	89.6	83.5	86.6
uni-T-skip	random	0.8293	0.7555	0.3180	72.5 / 81.0	67.3	74.9	89.0	81.1	83.6
C-T-skip	random	0.8458	0.7755	0.2902	74.7 / 82.1	70.4	76.7	90.4	83.8	84.8
bi-T-skip	GloVe	0.8444	0.7739	0.2922	75.1 / 82.4	74.4	79.5	90.9	85.3	87.6
uni-T-skip	GloVe	0.8485	0.7711	0.2854	73.7 / 81.8	74.6	78.8	91.1	86.2	87.0
C-T-skip	GloVe	0.8596	**0.7903**	0.2665	**75.4 / 82.6**	**75.6**	**80.4**	91.9	87.0	89.0
bi-T-skip	word2vec	0.8463	0.7744	0.2894	73.3 / 81.6	74.4	78.6	91.3	86.2	88.8
uni-T-skip	word2vec	0.8466	0.7705	0.2884	74.0 / 81.7	73.0	78.6	91.3	85.2	88.4
C-T-skip	word2vec	**0.8598**	0.7892	**0.2654**	75.0 / 82.2	75.1	80.0	**92.2**	**87.2**	**90.0**

Word Vectors Initialization

		Relationship of sentence pair				Classification on single sentence				
Model	WE	SICK			MSRP (Acc/F1)	MR	CR	SUBJ	MPQA	TREC
		r	ρ	MSE						
Average+Max Connection										
bi-T-skip	random	0.8336	0.7612	0.3112	73.2 / 81.3	69.7	76.0	89.6	83.5	86.6
uni-T-skip		0.8293	0.7555	0.3180	72.5 / 81.0	67.3	74.9	89.0	81.1	83.6
C-T-skip		0.8458	0.7755	0.2902	74.7 / 82.1	70.4	76.7	90.4	83.8	84.8
bi-T-skip	GloVe	0.8444	0.7739	0.2922	75.1 / 82.4	74.4	79.5	90.9	85.3	87.6
uni-T-skip		0.8485	0.7711	0.2854	73.7 / 81.8	74.6	78.8	91.1	86.2	87.0
C-T-skip		0.8596	**0.7903**	0.2665	**75.4 / 82.6**	**75.6**	**80.4**	91.9	87.0	89.0
bi-T-skip	word2vec	0.8463	0.7744	0.2894	73.3 / 81.6	74.4	78.6	91.3	86.2	88.8
uni-T-skip		0.8466	0.7705	0.2884	74.0 / 81.7	73.0	78.6	91.3	85.2	88.4
C-T-skip		**0.8598**	0.7892	**0.2654**	75.0 / 82.2	75.1	80.0	**92.2**	**87.2**	**90.0**

Furthermore…

- We wonder if adding more parameters could improve our model.

Furthermore…

- Double-sized encoder gave us further improvement.

Relationship of sentence pair (SICK, MSRP) — Classification on single sentence (MR, CR, SUBJ, MPQA, TREC)

	Model	WE	SICK			MSRP (Acc/F1)	MR	CR	SUBJ	MPQA	TREC
			r	ρ	MSE						
	Doubled Encoder's Dimension vs. Results reported by [6]										
Ours	bi-T-skip	word2vec	0.8503	0.7796	0.2823	74.4 / 82.2	74.8	80.3	91.8	87.0	88.2
	uni-T-skip		0.8486	0.7784	0.2857	74.3 / 82.4	72.9	78.0	90.7	85.7	86.4
	C-T-skip		**0.8611**	**0.7946**	**0.2634**	**74.5 / 82.2**	75.4	**80.3**	92.2	**87.4**	88.4
Kiros et al.	bi-skip [6]	random	0.8405	0.7696	0.2995	71.2 / 81.2	73.9	77.9	92.5	83.3	89.4
	uni-skip [6]		0.8477	0.7780	0.2872	73.0 / 81.9	75.5	79.3	92.1	86.9	91.4
	C-skip [6]		0.8584	0.7916	0.2687	73.0 / 82.0	**76.5**	80.1	**93.6**	87.1	**92.2**

Furthermore…

- Double-sized encoder gave us further improvement.

Relationship of sentence pair: SICK, MSRP (Acc/F1)

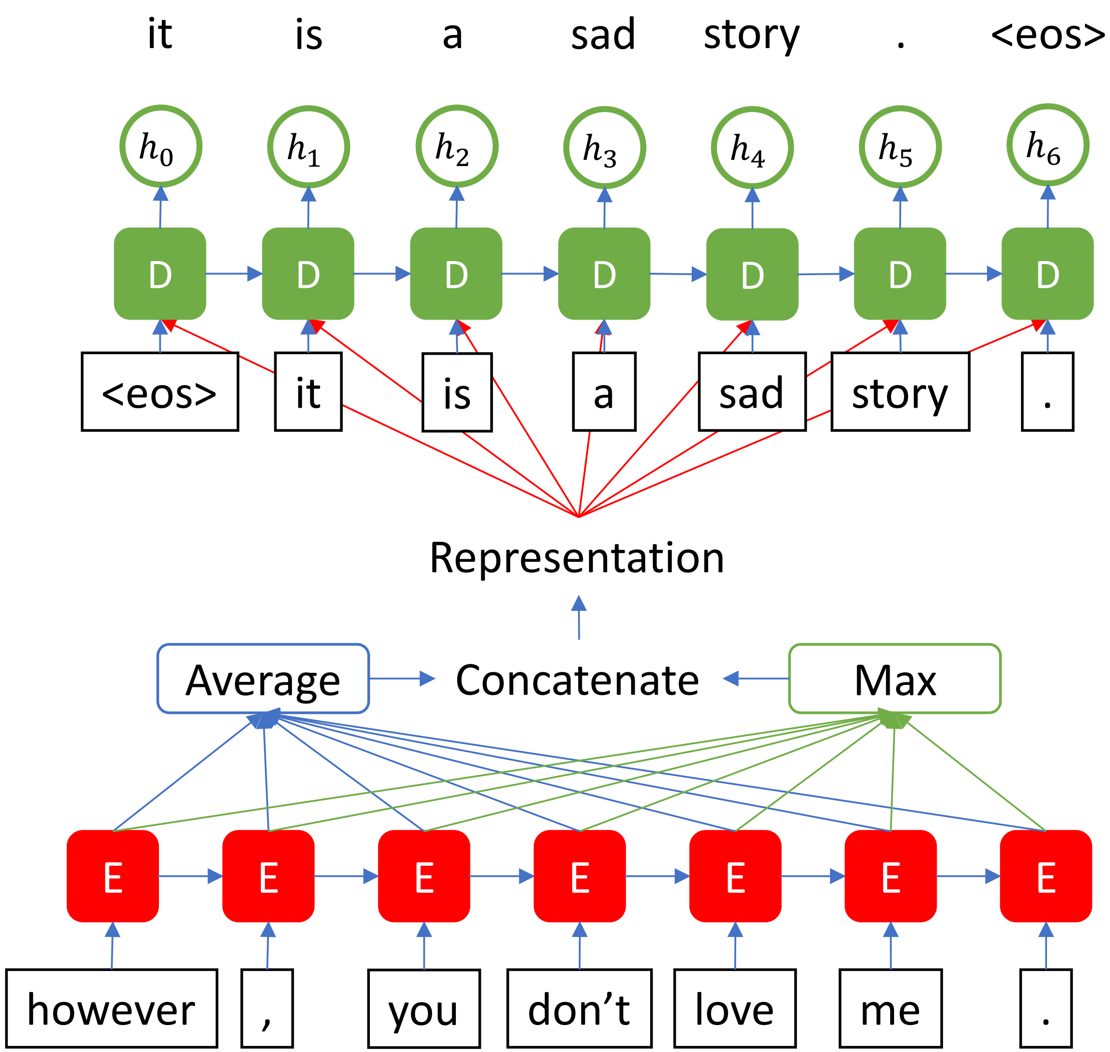
Classification on single sentence: MR, CR, SUBJ, MPQA, TREC

	Model	WE	SICK			MSRP (Acc/F1)	MR	CR	SUBJ	MPQA	TREC
			r	ρ	MSE						
	Doubled Encoder's Dimension vs. Results reported by [6]										
Ours	bi-T-skip	word2vec	0.8503	0.7796	0.2823	74.4 / 82.2	74.8	80.3	91.8	87.0	88.2
	uni-T-skip		0.8486	0.7784	0.2857	74.3 / 82.4	72.9	78.0	90.7	85.7	86.4
	C-T-skip		**0.8611**	**0.7946**	**0.2634**	**74.5 / 82.2**	75.4	**80.3**	92.2	**87.4**	88.4
Kiros et al.	bi-skip [6]	random	0.8405	0.7696	0.2995	71.2 / 81.2	73.9	77.9	92.5	83.3	89.4
	uni-skip [6]		0.8477	0.7780	0.2872	73.0 / 81.9	75.5	79.3	92.1	86.9	91.4
	C-skip [6]		0.8584	0.7916	0.2687	73.0 / 82.0	**76.5**	80.1	**93.6**	87.1	**92.2**

Trimming and Improving Skip-thought Vectors

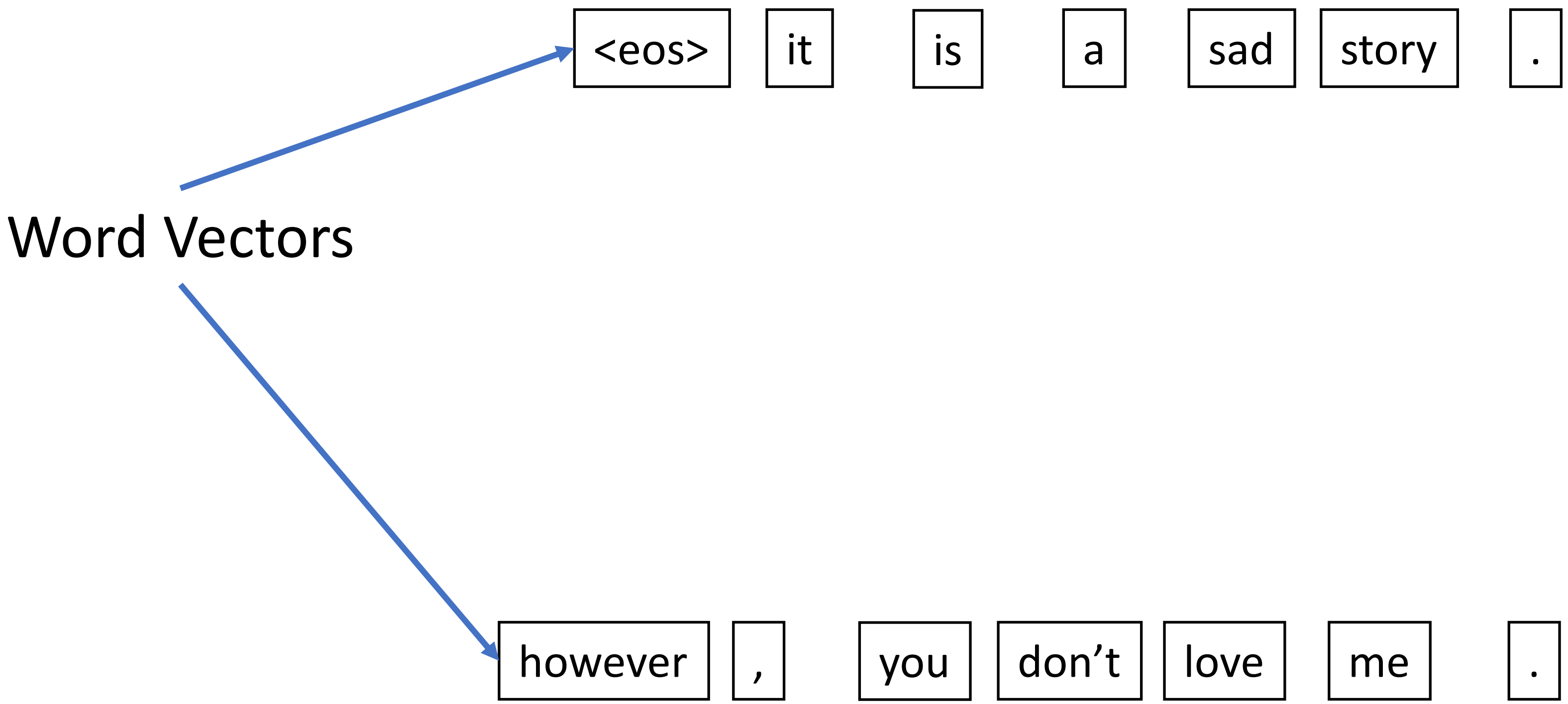
- Skip-thought
- Our hypotheses to improve skip-thought
- **Comparison between our trimmed skip-thought model and the skip-thought model**
  - Number of Parameters
  - Training Time
- Conclusion

Our Trimmed Skip-thought

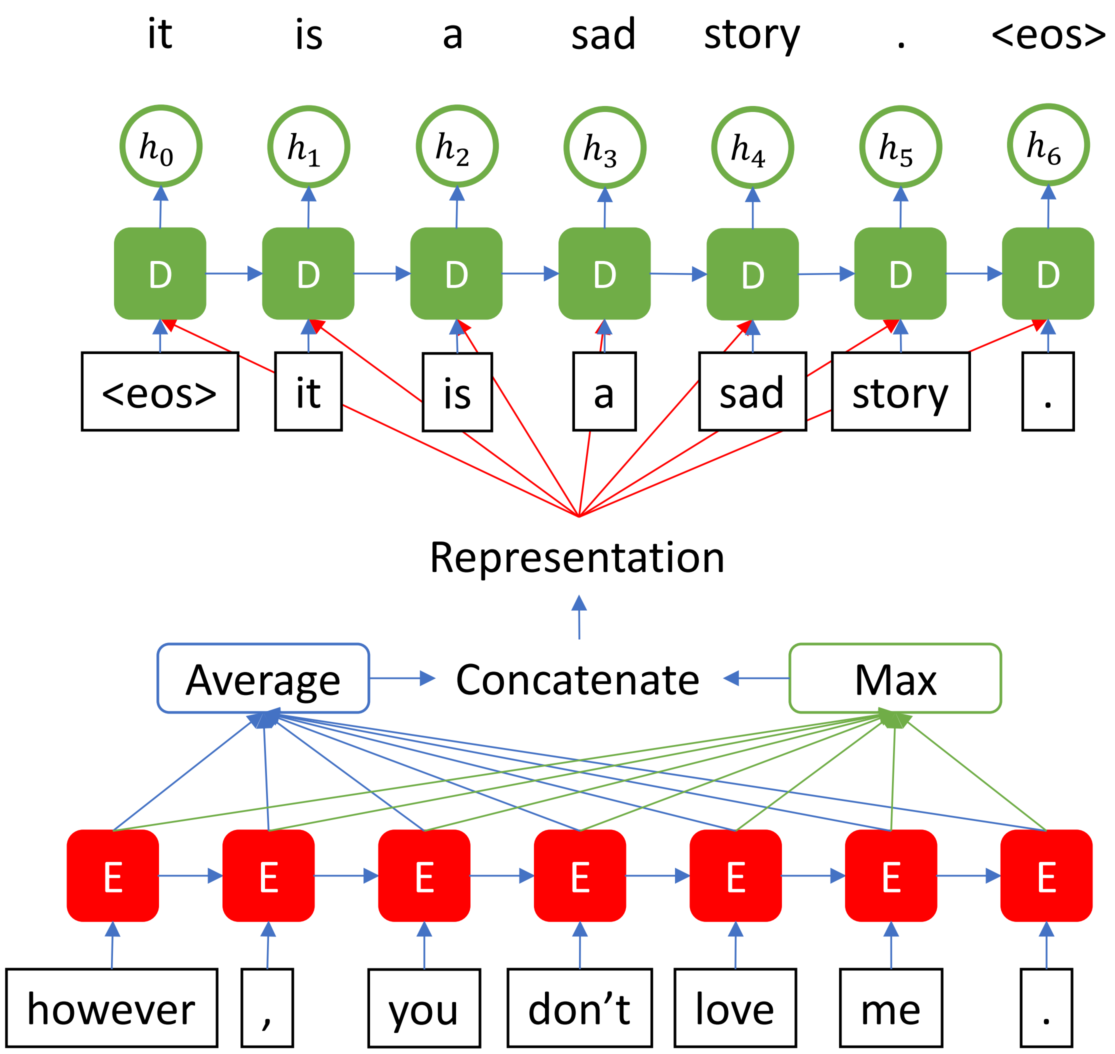


Our Trimmed Skip-thought

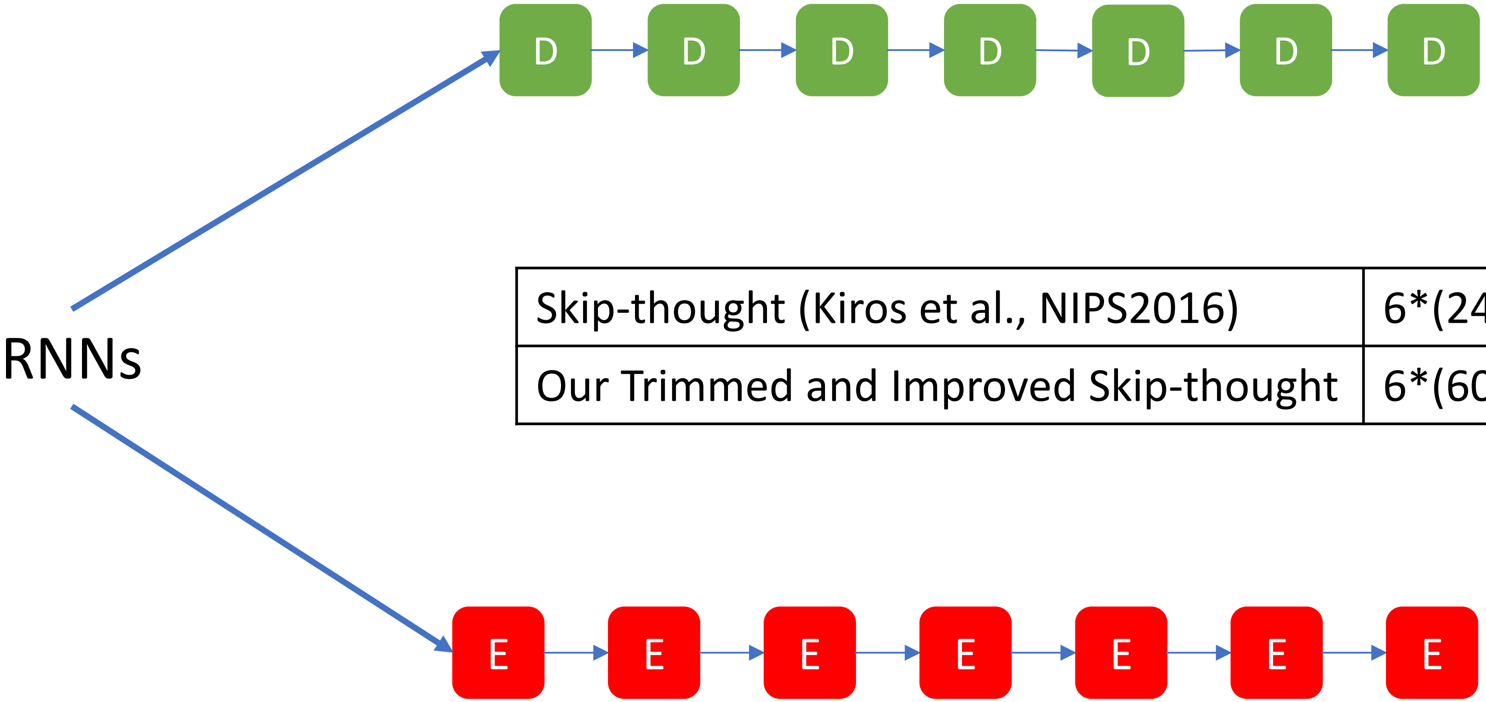
Skip-thought (Kiros et al., NIPS2016)	620*20000
Our Trimmed and Improved Skip-thought	300*20000



Our Trimmed Skip-thought

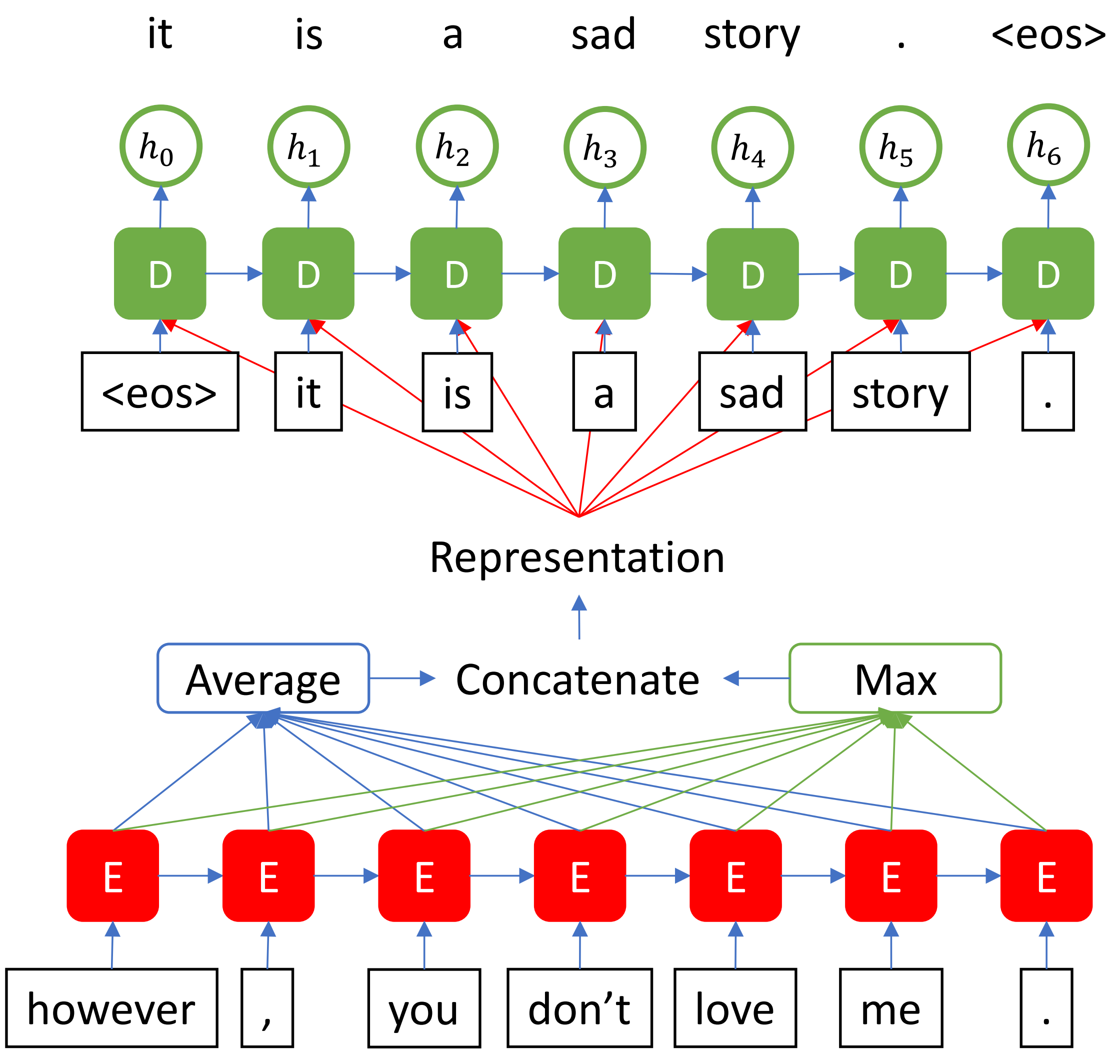


Our Trimmed Skip-thought

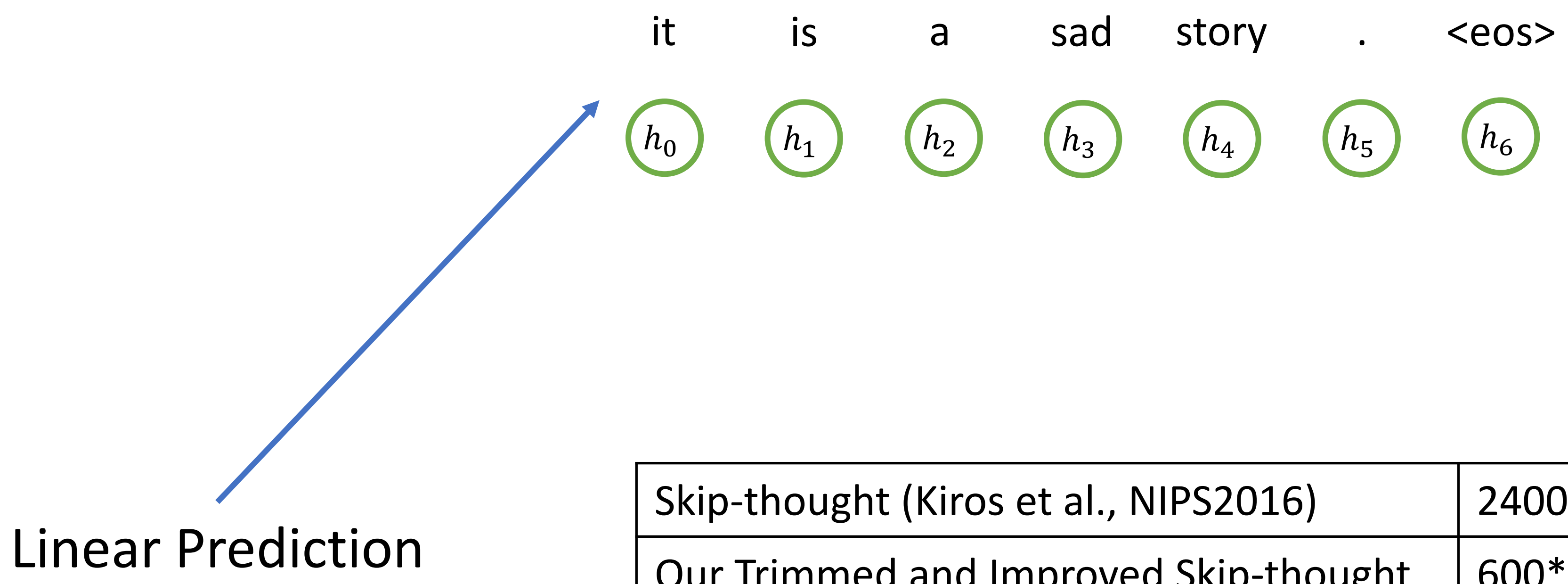


Skip-thought (Kiros et al., NIPS2016)	6*(2400*2400)*2
Our Trimmed and Improved Skip-thought	6*(600*600)*2

Our Trimmed Skip-thought



Our Trimmed Skip-thought



Skip-thought (Kiros et al., NIPS2016)	2400*20000
Our Trimmed and Improved Skip-thought	600*20000

Number of Parameters

Model	RNNs	Word Vectors	Linear Prediction
uni-T-skip (ours)	4.32M	6M	12M
bi-T-skip (ours)	3.24M		
uni-T-skip-double (ours)	10.80M		
bi-T-skip-double (ours)	6.48M		
uni-skip (Kiros et al., NIPS2015)	69.12M	12.4M	48M
bi-skip (Kiros et al., NIPS2015)	51.84M		

``RNNs'' refers to recurrent networks in the encoder and the decoder.
``Word Embedding'' refers to all word vectors in unsupervised training.
``Linear Prediction'' refers to the linear prediction layer in the decoder.

Training Time

Model	Training Time
Skip-thought (Kiros et al., NIPS2015)	2 weeks

Training Time

Model		Training Time
Skip-thought	(Kiros et al., NIPS2015)	2 weeks
Skip-thought	(our implementation)	4 days

Training Time

Model		Training Time
Skip-thought	(Kiros et al., NIPS2015)	2 weeks
Skip-thought	(our implementation)	4 days
Our Trimmed and Improved Skip-thought		1 day

Trimming and Improving Skip-thought Vectors

- Skip-thought
- Our hypotheses to improve skip-thought
- Comparison between our trimmed skip-thought model and the skip-thought model
- **Conclusion**

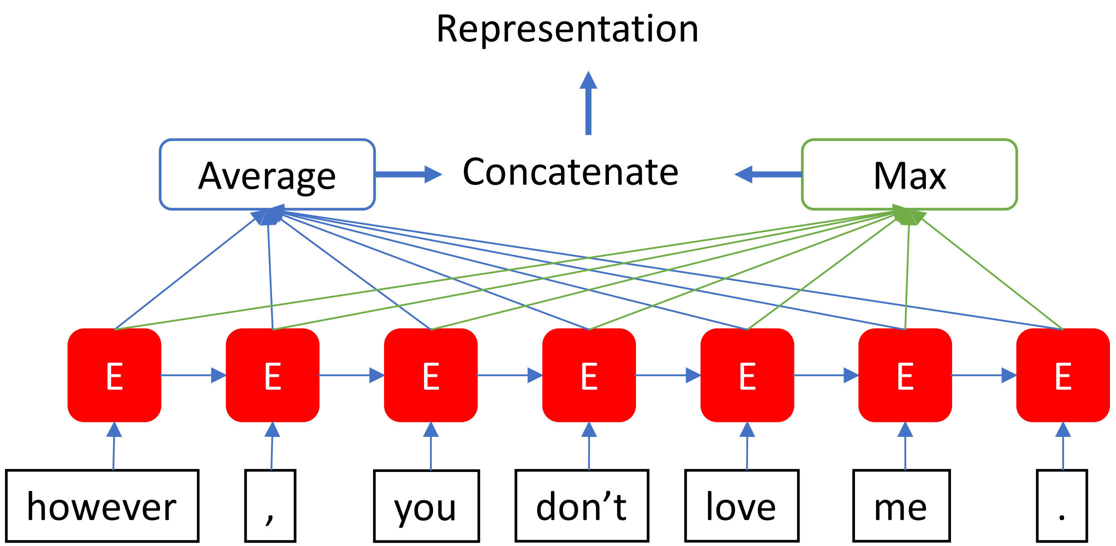
Conclusion: We improved skip-thought model

- by **dropping one decoder**

$$\boldsymbol{Sentence_i} \xrightarrow{\text{encode}} \boldsymbol{Representation_i} \xrightarrow{\text{decode}} \boldsymbol{Sentence_{i+1}}$$

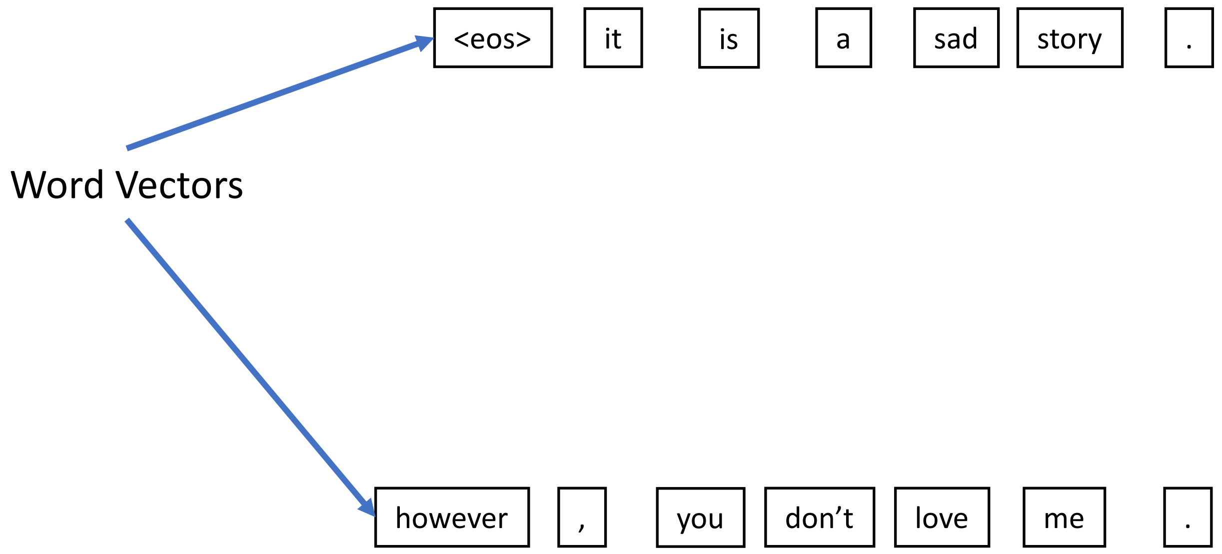
Conclusion: We improved skip-thought model

- by applying the **average+max connection** between the encoder and the decoder



Conclusion: We improved skip-thought model

- by **initializing** the model with **pretrained word vectors** instead of random values



Conclusion: We improved skip-thought model

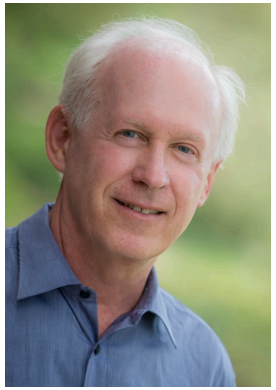
- by **accelerating** the training procedure, because we cut out 80% parameters in the skip-thought model.

Model		Training Time
Skip-thought	(Kiros et al., NIPS2015)	2 weeks
Skip-thought	(our implementation)	4 days
Our Trimmed and Improved Skip-thought		1 day

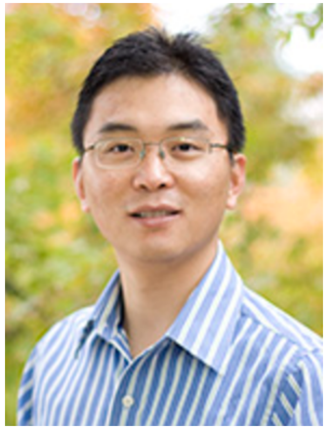
Committee & Collaborators

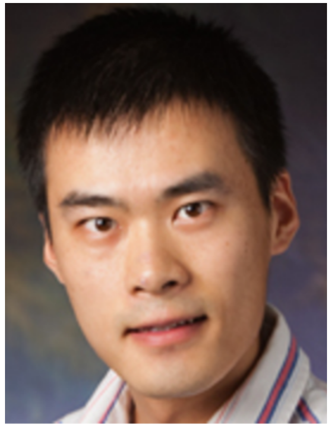
- Committee members
  - Virginia R. de Sa, CogSci
  - Benjamin K. Bergen, CogSci
  - Jeffrey L. Elman, CogSci
  - Julian J. McAuley, CSE






- Collaborators
  - Hailin Jin, Chen Fang, Zhaowen Wang
  - Researchers at Adobe research lab





Acknowledgements

- Many thanks to Adobe Research Lab for GPUs support, also to NVIDIA for DGX-1 trial.

Acknowledgements

- Many thanks to cohort, and Marta.

Q & A