Skip-Thought Vectors & Quick Thoughts Vectors for Sentence Rrepsentation

Xing Wang

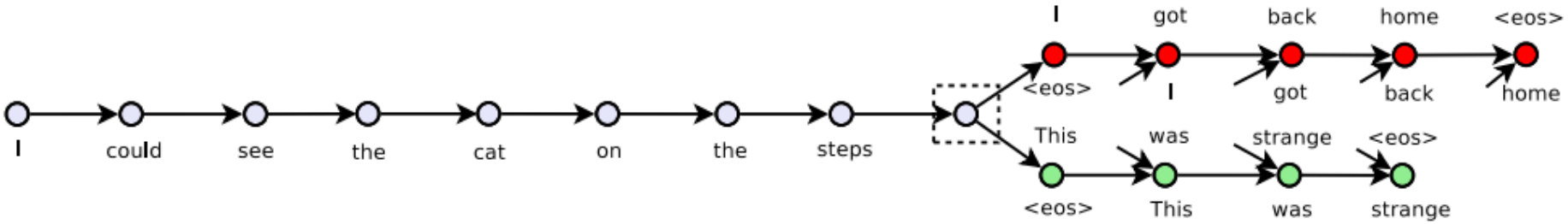
2018/07/02

- Skip-thought vectors, NIPS2015
- An efficient framework for learning sentence representations, ICLR2018

- Word representation:
  - the meaning of a word is characterized by the word-contexts in which it appears.

- Sentence representation:
  - the meaning of a sentence is ?

- Skip-Thought Vectors (ST)



- Loss:

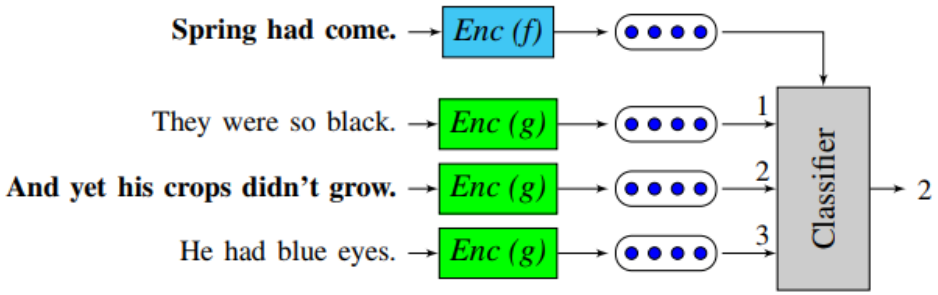
$$\sum_t \log P(w_{i+1}^t | w_{i+1}^{<t}, \mathbf{h}_i) + \sum_t \log P(w_{i-1}^t | w_{i-1}^{<t}, \mathbf{h}_i)$$

- Vocabulary expansion
  - linear transformation from word embedding space to RNN word embedding space

- Issues:
  - reconstruct the surface form of a sentence
  - computational cost

- Given an input sentence, it is encoded as before using some function. But instead of generating the target sentence, the model chooses the correct target sentence from a set of candidate sentences.

- Quick Thoughts Vectors (QT)



- Loss:

$$\sum_{s \in D} \sum_{s_{\text{ctxt}} \in S_{\text{ctxt}}} \log p(s_{\text{ctxt}} | s, S_{\text{cand}})$$

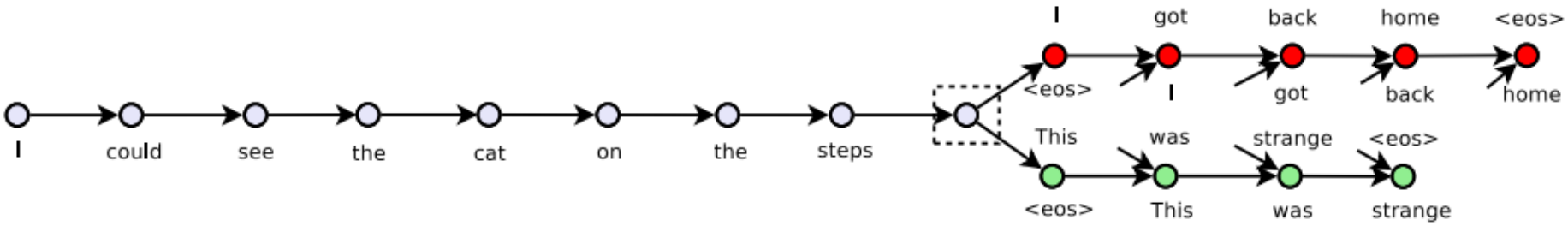
$$p(s_{\text{cand}} | s, S_{\text{cand}}) = \frac{\exp[c(f(s), g(s_{\text{cand}}))]}{\sum_{s' \in S_{\text{cand}}} \exp[c(f(s), g(s'))]}$$

- trained to reconstruct the surface form of a sentence

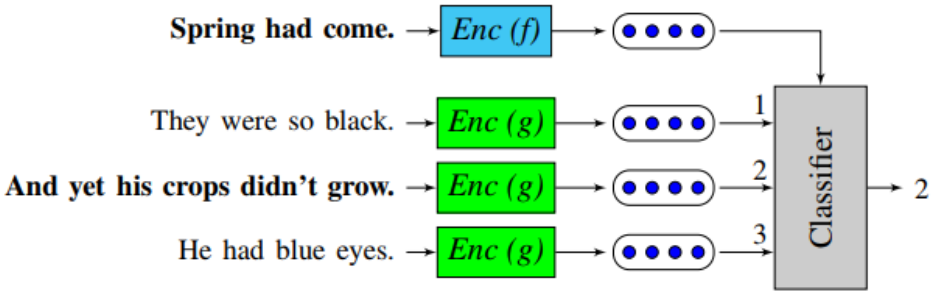
- There are numerous ways of expressing an idea in the form of a sentence. The ideal semantic representation is insensitive to the form in which meaning is expressed.

- Instead of training a model to reconstruct the surface form of the input sentence or its neighbors, our formulation attempts to focus on the semantic aspects of sentences. ~~The meaning of a sentence is the property that creates bonds between a sequence of sentences and makes it logically flow.~~
- Reviewer's comments: It's hard to pin down exactly what this means, but it sounds like you're making an empirical claim here: semantic information is more important than non-semantic sources of variation (syntactic/lexical/morphological factors) in predicting the flow of a text. Provide some evidence for this, or cut it.

- Skip-Thought Vectors (ST)



- Quick Thoughts Vectors (QT)



- Evaluation:
  1. Use the learned encoder as a feature extractor.
  2. If the task involves computing scores between pairs of sentences, compute component-wise features between pairs.
  3. Train a linear classifier on top of the extracted features, with no additional fine-tuning or backpropagation through the skip-thoughts model.

- Sentence representation in ST
  - use the encoder as a generic feature extractor.

- Sentence representation in QT
  - for a given sentence s, we consider its representation to be the concatenation of the outputs of the two encoders [f(s) g(s)].

- Sentence pair representation in ST
  - Given two skip-thought vectors u and v, we compute their component-wise product $u \cdot v$ and their absolute difference $|u - v|$ and concatenate them together.

- Experiments:
  - Semantic relatedness: SemEval 2014 Task 1: semantic relatedness SICK dataset
  - Paraphrase detection: Microsoft Research Paraphrase Corpus
  - Image-sentence ranking: Microsoft COCO dataset
  - Classification benchmarks: movie review sentiment (MR), customer product reviews (CR), subjectivity/objectivity classification (SUBJ), opinion polarity (MPQA) and question-type classification (TREC)
  - Nearest Neighbors/Visualization

Model	Dim	Training time (h)	MR	CR	SUBJ	MPQA	TREC	MSRP (Acc)	MSRP (F1)	SICK r	SICK ρ	SICK MSE
GloVe BoW	300	-	78.1	80.4	91.9	87.8	85.2	72.5	81.1	0.764	0.687	0.425
Trained from scratch on BookCorpus data												
SDAE	2400	192	67.6	74.0	89.3	81.3	77.6	**76.4**	83.4	N/A	N/A	N/A
FastSent	<500	2*	71.8	78.4	88.7	81.5	76.8	72.2	80.3	N/A	N/A	N/A
ParagraphVec	<500	4*	61.5	68.6	76.4	78.1	55.8	73.6	81.9	N/A	N/A	N/A
uni-skip	2400	336	75.5	79.3	92.1	86.9	91.4	73.0	81.9	0.848	0.778	0.287
bi-skip	2400	336	73.9	77.9	92.5	83.3	89.4	71.2	81.2	0.841	0.770	0.300
combine-skip	4800	336†	76.5	80.1	**93.6**	87.1	**92.2**	73.0	82.0	0.858	0.792	0.269
combine-cnn	4800	-	77.2	80.9	93.1	**89.1**	91.8	75.5	82.6	0.853	0.789	0.279
uni-QT	2400	11	77.2	82.8	92.4	87.2	90.6	74.7	82.7	0.844	0.778	0.293
bi-QT	2400	9	77.0	83.5	92.3	87.5	89.4	74.8	82.9	0.855	0.787	0.274
combine-QT	4800	11†	**78.2**	**84.4**	93.3	88.0	90.8	76.2	**83.5**	**0.860**	**0.796**	**0.267**
Trained on BookCorpus, pre-trained word vectors are used												
combine-cnn	4800	-	77.8	82.1	93.6	**89.4**	92.6	76.5	83.8	0.862	0.798	0.267
MC-QT	4800	11	**80.4**	**85.2**	**93.9**	**89.4**	**92.8**	**76.9**	**84.0**	**0.868**	**0.801**	**0.256**
Trained on (BookCorpus + UMBC) data, from scratch and using pre-trained word vectors												
combine-QT	4800	28	81.3	84.5	94.6	89.5	92.4	75.9	83.3	0.871	0.807	0.247
MC-QT	4800	28	**82.4**	**86.0**	**94.8**	**90.2**	92.4	**76.9**	**84.0**	**0.874**	**0.811**	**0.243**

- MultiChannel-QT.
  - The MultiChannel-QT model (MC-QT) is defined as the concatenation of two bi-directional RNNs.
  - One of these uses fixed pre-trained word embeddings coming from a large vocabulary (~ 3M) as input. While the other uses tunable word embeddings trained from scratch (from a smaller vocabulary ~ 50k).

- Quick !
  - The BookCorpus dataset: 45M ordered sentences
  - the UMBC corpus: a dataset of 100M web pages crawled from the internet, preprocessed and tokenized into paragraphs. The dataset has 129M sentences, about three times larger than BookCorpus.

- Nearest Neighbors(corpus:Wikipedia dump)

Query	Seizures may occur as the glucose falls further .
ST	It may also occur during an excessively rapid entry into autorotation .
QT	When brain glucose levels are sufficiently low , seizures may result .
Query	This evidence was only made public after both enquiries were completed .
ST	This visa was provided for under Republic Act No .
QT	These evidence were made public by the United States but concealed the names of sources .
Query	He kept both medals in a biscuit tin .
ST	He kept wicket for Middlesex in two first-class cricket matches during the 1891 County Championship .
QT	He won a three medals at four Winter Olympics .
Query	The American alligator is the only known natural predator of the panther .
ST	Their mascot is the panther .
QT	The American alligator is a fairly large species of crocodilian .
Query	Several of them died prematurely : Carmen and Toms very young , while Carlos and Pablo both died .
ST	At the age of 13 , Ahmed Sher died .
QT	Many of them died in prison .
Query	Music for “ Expo 2068 ” originated from the same studio session .
ST	His 1994 work “ Dialogue ” was premiered at the Merkin Concert Hall in New York City .
QT	Music from “ Korra ” and “ Avatar ” was also played in concert at the PlayFest festival in Mlaga , Spain in September 2014 .

- Conclusion:
  - Skip-Thoughts
  - Evaluation: “niubious” representation + “lowbious” classifier