Skip-Thought Vectors

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, Sanja Fidler

Student Presentation By Harish Shanker, Manu Gandham

Problem

Is there a task and a corresponding loss that will allow us to learn highly generic sentence representations?

The team proposes a model for learning high quality sentence vectors without a particular supervised task in mind

- Propose an objective function that abstracts the skip-gram model to sentence level
- Instead of using a word to predicts its surrounding text, encode a sentence to predict the sentences around it

Previous Work

- There have been several approaches developed for learning composition operators that map word vectors to sentence vectors
    - Recursive networks
    - Recurrent networks
    - Convolutional networks
    - Recursive convo network
- All produce sentence representations that are passed to a supervised task
    - Learn high quality sentence representations but are tuned ONLY for their respective task
- Paragraph vector - learn unsupervised sentence representations by introducing a distributed sentence indicator as part of a neural language model
    - Downside: inference needs to be performed to compute a new vector

Skip-Thought vectors

- Represented by an encoder-decoder model
- Encoder: Maps English sentence into a vector
- Decoder: Conditions on this vector to generate surrounding sentences
- Architecture: RNN encoder with GRU activations, RNN decoder with condition GRU
- Benefit: Skip-thoughts yield generic representation that perform robustly across all tasks considered

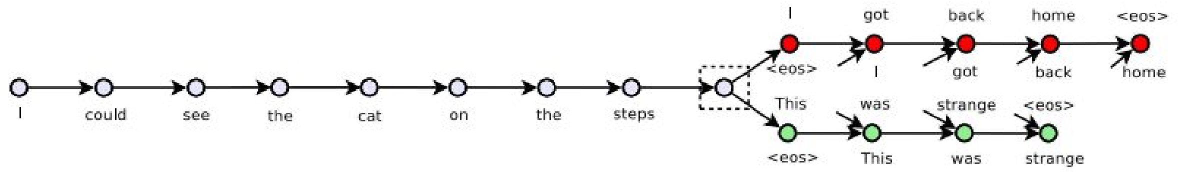


Figure 1: The skip-thoughts model. Given a tuple (s_{i-1}, s_i, s_{i+1}) of contiguous sentences, with s_i the i-th sentence of a book, the sentence s_i is encoded and tries to reconstruct the previous sentence s_{i-1} and next sentence s_{i+1}. In this example, the input is the sentence triplet *I got back home. I could see the cat on the steps. This was strange.* Unattached arrows are connected to the encoder output. Colors indicate which components share parameters. $\langle$eos$\rangle$ is the end of sentence token.

Encoder

- Let $w_i^1, \ldots, w_i^N$ be the words in sentence s_i where N is the number of words in the sentence
- At each time step, the encoder produces a hidden state h_i^t which can be interpreted as the representation of the sequence $w_i^1, \ldots, w_i^t$
- The hidden state h_i^N thus represents the full sentence

$$
\begin{aligned}
\mathbf{r}^t &= \sigma(\mathbf{W}_r \mathbf{x}^t + \mathbf{U}_r \mathbf{h}^{t-1}) \\
\mathbf{z}^t &= \sigma(\mathbf{W}_z \mathbf{x}^t + \mathbf{U}_z \mathbf{h}^{t-1}) \\
\bar{\mathbf{h}}^t &= \tanh(\mathbf{W}\mathbf{x}^t + \mathbf{U}(\mathbf{r}^t \odot \mathbf{h}^{t-1})) \\
\mathbf{h}^t &= (1 - \mathbf{z}^t) \odot \mathbf{h}^{t-1} + \mathbf{z}^t \odot \bar{\mathbf{h}}^t
\end{aligned}
$$

Decoder

- Introduce matrices C_z , C_r, and C that are used to bias the update gate, reset gate and hidden state computation by the sentence vector
- Separate decoders for previous and next sentences (S_{i-1} and S_{i+1})
- Separate params for each decoder

$$\mathbf{r}^t = \sigma(\mathbf{W}_r^d \mathbf{x}^{t-1} + \mathbf{U}_r^d \mathbf{h}^{t-1} + \mathbf{C}_r \mathbf{h}_i) \quad (5)$$

$$\mathbf{z}^t = \sigma(\mathbf{W}_z^d \mathbf{x}^{t-1} + \mathbf{U}_z^d \mathbf{h}^{t-1} + \mathbf{C}_z \mathbf{h}_i) \quad (6)$$

$$\bar{\mathbf{h}}^t = \tanh(\mathbf{W}^d \mathbf{x}^{t-1} + \mathbf{U}^d (\mathbf{r}^t \odot \mathbf{h}^{t-1}) + \mathbf{C}\mathbf{h}_i) \quad (7)$$

$$\mathbf{h}_{i+1}^t = (1 - \mathbf{z}^t) \odot \mathbf{h}^{t-1} + \mathbf{z}^t \odot \bar{\mathbf{h}}^t \quad (8)$$

Given $\mathbf{h}_{i+1}^t$, the probability of word w_{i+1}^t given the previous $t-1$ words and the encoder vector is

$$P(w_{i+1}^t | w_{i+1}^{<t}, \mathbf{h}_i) \propto \exp(\mathbf{v}_{w_{i+1}^t} \mathbf{h}_{i+1}^t) \quad (9)$$

Objective Function

- Given a tuple (s_{i-1}, s_i , s_{i+1}), the objective optimized is the sum of the log-probabilities for the forward and backward sentences conditioned on the encoder representation:

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

- The total objective is the above summed over all such training tuples

Experiment Setup

- Using the learned encoder as a feature extractor, extract skip-thought vectors for all sentences
- If the task involves computing scores between pairs of sentences, compute component-wise features between pairs
- Train a linear classifier on top of the extracted features, with no additional fine-tuning or backpropagation

Data

Book Corpus dataset

- Large collection of novels (free books written by unpublished authors)
- 16 different genres (Romance, Fantasy, Science Fiction, etc.)

# of books	# of sentences	# of words	# of unique words	mean # of words per sentence
11,038	74,004,228	984,846,357	1,316,420	13

Training Details

- 3 types of embeddings were created: uni-skip, bi-skip, and combine-skip
- Minibatch size: 128, Gradients are clipped if the norm of the vector exceeds 10, Adam algorithm for optimization
- Trained on 20,000 word vocabulary from the Book Corpus database
- Expanded to 930,911 word vocabulary using vocab expansion and CBOW word vectors
- Since the goal is to evaluate skip-thoughts as a general feature extractor, pre-processing is kept to a minimum
- When encoding new sentences, no additional preprocessing is done other than basic tokenization - this is done to test the robustness of the skip-thought vectors

Vocabulary expansion

- Map the embedding space of the desired vocabulary to the input shape of the RNN encoder
- To do this: solve for the matrix W which can be used to transform between vocabulary spaces
- L2 linear regression problem

choreograph	**modulation**	**vindicate**	**neuronal**	**screwy**	**Mykonos**	**Tupac**
choreography	transimpedance	vindicates	synaptic	wacky	Glyfada	2Pac
choreographs	harmonics	exonerate	neural	nutty	Santorini	Cormega
choreographing	Modulation	exculpate	axonal	iffy	Dubrovnik	Biggie
rehearse	##QAM	absolve	glial	loopy	Seminyak	Gridlock'd
choreographed	amplitude	undermine	neuron	zany	Skiathos	Nas
Choreography	upmixing	invalidate	apoptotic	kooky	Hersonissos	Cent
choreographer	modulations	refute	endogenous	dodgy	Kefalonia	Shakur

Table 3: Nearest neighbours of words after vocabulary expansion. Each query is a word that does not appear in our 20,000 word training vocabulary.

Experiment: Semantic Relatedness

- SICK dataset: Humans scored sentences on how similar they are
- Authors trained a logistic regression classifier to predict semantic relatedness for two encoded skip-thought vectors
- Given two skip-thought vectors u and v, compute their component-wise product $u \cdot v$ and their absolute difference $|u - v|$ and concatenate them together

Sentence 1	Sentence 2	GT	pred
A little girl is looking at a woman in costume	A young girl is looking at a woman in costume	4.7	4.5
A little girl is looking at a woman in costume	The little girl is looking at a man in costume	3.8	4.0
A little girl is looking at a woman in costume	A little girl in costume looks like a woman	2.9	3.5

Experiment: Paraphrase Detection

- Task: 2 sentences are given and one must predict whether or not they are paraphrases (using MSR Paraphrase Corpus)
- Skip-thought encoding + linear classifier works just as well as RNNs for some tasks, unless the features are hand selected
- Observations:
  - Skip-thoughts alone outperform recursive nets with dynamic pooling when no hand-crafted features are used
  - when other features are used, recursive nets with dynamic pooling works better
  - when skip-thoughts are combined with basic pairwise statistics, it becomes competitive with the state-of-the-art which incorporate much more complicated features and hand-engineering

Experiment: Image Sentence Ranking

- Best results for image sentence ranking achieved with RNNs
- Fisher vectors + linear CCA has been shown
- Images represented by features from OxfordNet

COCO Retrieval

	Image Annotation				Image Search			
Model	**R@1**	**R@5**	**R@10**	**Med *r***	**R@1**	**R@5**	**R@10**	**Med *r***
Random Ranking	0.1	0.6	1.1	631	0.1	0.5	1.0	500
DVSA [31]	38.4	69.6	80.5	**1**	27.4	**60.2**	74.8	**3**
GMM+HGLMM [32]	39.4	67.9	80.9	2	25.1	59.8	76.6	4
m-RNN [33]	**41.0**	**73.0**	**83.5**	2	**29.0**	42.2	**77.0**	**3**
uni-skip	30.6	64.5	79.8	3	22.7	56.4	71.7	4
bi-skip	32.7	67.3	79.6	3	24.2	57.1	73.2	4
combine-skip	33.8	67.7	82.1	3	25.9	60.0	74.6	4

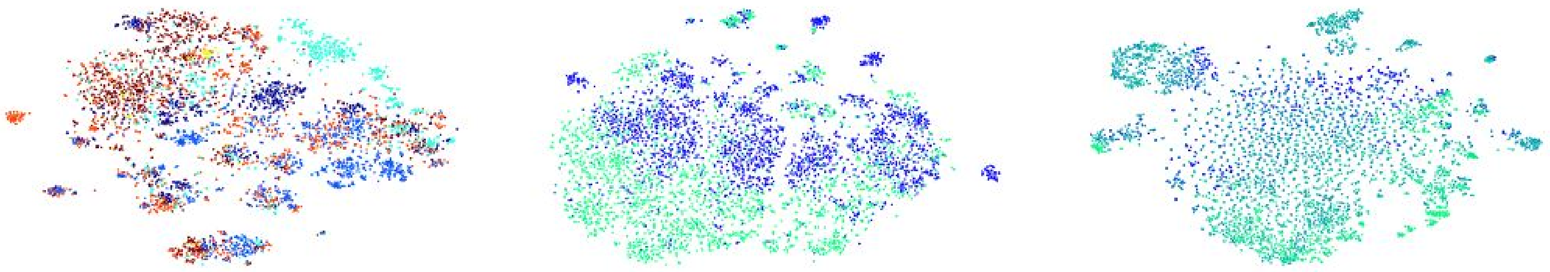
Classification Benchmarks

- Use 5 datasets: movie review sentiment (MR), customer product reviews (CR), subjectivity/objectivity classification (SUBJ), opinion polarity (MPQA), and question-type classification (TREC)
- Extracted skip-thought vectors and trained a logistic regression classifier on top
- Skip-thoughts performs about as well as the bag-of-words baselines
- However, fails to improve over methods whose sentence representations are learned directly for the task at hand

Method	MR	CR	SUBJ	MPQA	TREC
NB-SVM [41]	79.4	81.8	93.2	86.3	
MNB [41]	79.0	80.0	93.6	86.3	
cBoW [6]	77.2	79.9	91.3	86.4	87.3
GrConv [6]	76.3	81.3	89.5	84.5	88.4
RNN [6]	77.2	82.3	93.7	90.1	90.2
BRNN [6]	82.3	82.6	94.2	90.3	91.0
CNN [4]	81.5	85.0	93.4	89.6	**93.6**
AdaSent [6]	**83.1**	**86.3**	**95.5**	**93.3**	92.4
Paragraph-vector [7]	74.8	78.1	90.5	74.2	91.8
uni-skip	75.5	79.3	92.1	86.9	91.4
bi-skip	73.9	77.9	92.5	83.3	89.4
combine-skip	76.5	80.1	93.6	87.1	92.2
combine-skip + NB	80.4	81.3	93.6	87.5	

Visualizing Skip-Thoughts

- Sentence pairs that are similar to each other are embedded next to other similar pairs
- Even without the use of relatedness labels, skip-thought vectors learn to accurately capture this property



(a) TREC (b) SUBJ (c) SICK

Novel Generation

- Perform generation by conditioning on a sentence, generating a new sentence, concatenating the generated example to the previous text and continuing
- Model was trained on books, the generated samples is a nonsensical novel

she grabbed my hand . " come on . " she fluttered her bag in the air . " i think we 're at your place . i ca n't come get you . " he locked himself back up . " no . she will . " kyrian shook his head . " we met ... that congratulations ... said no . " the sweat on their fingertips 's deeper from what had done it all of his flesh hard did n't fade . cassie tensed between her arms suddenly grasping him as her sudden her senses returned to its big form . her chin trembled softly as she felt something unreadable in her light . it was dark . my body shook as i lost what i knew and be betrayed and i realize just how it ended . it was n't as if i did n't open a vein . this was all my fault , damaged me . i should have told toby before i was screaming . i should 've told someone that was an accident . never helped it . how can i do this , to steal my baby 's prints ? "

Conclusion

- Skip-thought vectors perform well on MANY tasks, demonstrating the robustness of this representation
- Experiments only scratch the surface, lot of variations for improvement:
    - Deep encoders and decoders
    - Larger context windows
    - Encoding and decoding paragraphs
    - Other encoders, such as convnets