Unsupervised learning of word embeddings from speech

Yu-An Chung
April 11, 2018

Spoken Language Systems Group
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Motivation

• NLP techniques such as Word2Vec and GloVe transform words in a given text corpus into vector representations of fixed dimensionality (embeddings).

• Obtained via unsupervised learning from co-occurrence information in text

• Contain semantic information of the words
Speech and text are languages in different forms

- Can machines learn meaningful vector representations from speech and only from speech as well?

- If yes, what kind of information do these vector representations contain?
Audio signal processing is currently undergoing a paradigm change, where data-driven machine learning is replacing hand-crafted feature design. This has led some to ask whether audio signal processing is still useful in the era of machine learning.

**Text (written language)**

- Input
- **Learning machine** such as word2vec
- Output
- Word embeddings
  - audio
  - signal
  - processing
  - :
  - learning

**Speech (spoken language)**

- Input
- **Learning machine** our goal
- Output
- Speech segment embeddings
  - :
Word2Vec (Skip-gram) Recap

Audio signal processing is currently undergoing a paradigm change ...

Window size = 2

All represented as one-hot vectors

Softmax probability estimator

Single layer fully-connected neural network (linear)

Word embedding of $x_t$

$x_t$ represented as one-hot vector
Our proposed model: Speech2Vec

Speech

Represented as a sequence of acoustic feature vectors

Variable-length sequence?

RNN (acts as an encoder)

Another RNN as decoder

Embedding of $x_t$

Represented as a sequence of acoustic feature vectors such as MFCCs
Evaluation of the Speech2Vec word embeddings

Corpus
- LibriSpeech - a large corpus of read English speech (500 hours)
- Acoustic features consisted of 13-dim MFCCs produced every 10ms
- Corpus was segmented via forced alignment such that each speech segment corresponds to a spoken word

Model Architecture
- Encoder: A single-layered bidirectional LSTM
- Decoder: A single-layered unidirectional LSTM
- Window size is set to 3
- A fixed learning rate of 1e-3

Comparing Model
- Word2Vec (skip-gram and CBOW) trained on the LibriSpeech transcriptions
13 Word Similarity Benchmarks

• Contain different numbers of pairs of English words that have been assigned similarity ratings by humans

• Commonly used for evaluating how well the word embeddings capture the semantics of the words they represent

• During testing:
  – Given a pair of words, their similarity was calculated by computing the cosine similarity between their corresponding word embeddings.
  – Spearman’s rank correlation coefficient $\rho$ between the rankings produced by the machine against the human rankings were reported.
  – The higher $\rho$ the better
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Speech2Vec</th>
<th>Word2Vec</th>
<th>Speech2Vec</th>
<th>Word2Vec</th>
<th>Speech2Vec</th>
<th>Word2Vec</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cbow</td>
<td>skipgrams</td>
<td>cbow</td>
<td>skipgrams</td>
<td>cbow</td>
<td>skipgrams</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>50</td>
<td>100</td>
<td>200</td>
<td>10</td>
<td>50</td>
</tr>
<tr>
<td>Verb-143</td>
<td>0.182</td>
<td><strong>0.223</strong></td>
<td>0.203</td>
<td>0.205</td>
<td>0.263</td>
<td><strong>0.315</strong></td>
</tr>
<tr>
<td>SimLex-999</td>
<td>0.183</td>
<td>0.235</td>
<td><strong>0.238</strong></td>
<td>0.237</td>
<td>0.200</td>
<td>0.292</td>
</tr>
<tr>
<td>MC-30</td>
<td>0.680</td>
<td><strong>0.716</strong></td>
<td>0.688</td>
<td>0.684</td>
<td>0.701</td>
<td><strong>0.846</strong></td>
</tr>
<tr>
<td>WS-353</td>
<td>0.305</td>
<td><strong>0.343</strong></td>
<td>0.336</td>
<td>0.335</td>
<td>0.370</td>
<td><strong>0.508</strong></td>
</tr>
<tr>
<td>WS-353-SIM</td>
<td>0.461</td>
<td><strong>0.484</strong></td>
<td>0.474</td>
<td>0.471</td>
<td>0.533</td>
<td><strong>0.663</strong></td>
</tr>
<tr>
<td>WS-353-REL</td>
<td>0.122</td>
<td><strong>0.192</strong></td>
<td>0.189</td>
<td>0.186</td>
<td>0.207</td>
<td><strong>0.346</strong></td>
</tr>
<tr>
<td>RG-65</td>
<td>0.676</td>
<td><strong>0.705</strong></td>
<td>0.699</td>
<td>0.697</td>
<td>0.702</td>
<td><strong>0.790</strong></td>
</tr>
<tr>
<td>MEN</td>
<td>0.476</td>
<td><strong>0.509</strong></td>
<td>0.501</td>
<td>0.498</td>
<td>0.543</td>
<td><strong>0.619</strong></td>
</tr>
<tr>
<td>MTurk-287</td>
<td>0.346</td>
<td><strong>0.349</strong></td>
<td>0.336</td>
<td>0.331</td>
<td>0.426</td>
<td><strong>0.468</strong></td>
</tr>
<tr>
<td>MTurk-771</td>
<td>0.356</td>
<td><strong>0.391</strong></td>
<td>0.380</td>
<td>0.377</td>
<td>0.445</td>
<td><strong>0.521</strong></td>
</tr>
<tr>
<td>SimVerb-3500</td>
<td>0.098</td>
<td>0.122</td>
<td><strong>0.126</strong></td>
<td>0.125</td>
<td>0.100</td>
<td>0.157</td>
</tr>
<tr>
<td>Rare-Word</td>
<td>0.240</td>
<td>0.273</td>
<td><strong>0.275</strong></td>
<td>0.269</td>
<td>0.249</td>
<td><strong>0.323</strong></td>
</tr>
<tr>
<td>YP-130</td>
<td>0.198</td>
<td><strong>0.216</strong></td>
<td>0.211</td>
<td>0.214</td>
<td>0.322</td>
<td><strong>0.334</strong></td>
</tr>
</tbody>
</table>

## Discussions

1. Skip-grams outperforms CBOW
2. Word embeddings of 50-dim perform the best
3. Speech2Vec outperforms Word2Vec (why?)
Conclusions

• We propose Speech2Vec, a speech version of Word2Vec, for unsupervised learning of word embeddings from speech.

• In word similarity task, Speech2Vec trained on the LibriSpeech corpus outperforms Word2Vec trained on the LibriSpeech transcriptions.

• Future Works
  – Try Speech2Vec on non pre-segmented speech corpus (truly unsupervised)
  – Explore the possibility of learning the link (alignment) between speech and text embedding spaces.

• Publications
  – Learning word embeddings from speech (Chung and Glass, 2017)
  – Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech (Chung and Glass, 2018)
Thank you!