Multi-view representation learning for speech (and language)

Karen Livescu

Joint work with
Qingming Tang, Weiran Wang, Raman Arora, Galen Andrew, Jeff Bilmes, ...
Text vs. speech

This is a speech signal:

Some differences between text and speech:

- Speech is continuous-valued, text is discrete
- Speech is also continuous in time: Words and sounds are not separated

Speech and NLP research have a lot in common...

- Similar problems
- Many of the same algorithms
- Many researchers work on both
Text vs. speech

Text can be more or less formal. Informal text has many variants.

- haha
- hahahahahahahaha
- haaaaahaaaa
- lol
- rotflmao
- lol!!!!!!!!!!!!!
- wow that is big
- that is biiiiiiig
- that. is. big.
- waaaaaaay big
Spoken words have even more variants: pronunciation, speaker, acoustic environment, mood, state of inebriation...
A “simple” speech task: Single-digit classification

This is a 1-second speech waveform. Which digit (0-9) was spoken?

What are we looking at?

- Recording from a microphone: instantaneous air pressure vs. time
- Discretized in time (in this case, to 16,000 samples, i.e. sampling rate of 16kHz)
- Discretized in magnitude (in this case, to 16 bits per sample)
- Result: 16,000-dimensional vector, e.g. $a(t) = [3, 16, -1, 0, 427, 29, \ldots]$
This is hard!

Which two are the same digit?
Idea: Use a frequency-domain representation

**Spectrogram:** Fourier transform over short windows (e.g. 20ms) → plot of energy at each frequency over time $f_1(\omega), f_2(\omega), \ldots$
This is still hard!

Several examples of the digit “eight”
Architecture of a “traditional” speech recognizer
Architecture of an “end-to-end” speech recognizer
Acoustic features (representations)

- **Waveform**
- **Spectrogram**
- **MFCCs**
Are these reviews positive or negative?

- This is the best mattress I have ever had. It is a perfect combination of firmness and support. I have never slept better. ...

- I hate this mattress. I can’t believe I bought it. It seemed good in the store but when it was delivered I noticed it had a strange smell and was already lumpy. This is not what a new mattress should feel like. I want to tear it up and dump it on the ...
Representations for text

A possible feature vector (representation) for the review sentiment classification task:

- \# words from the set \{ good, great, best, lovely, perfect, ... \}
- \# words from the set \{ bad, horrible, worst, irritating, ... \}
- total \# words (?)

Some features that we would probably not use:

- \# words that start with “t”
- \# capital letters

...
Representation learning

Maybe we can design an algorithm to automatically learn what are good features?

- Start with a very long vector of possibly useful features, $x = [x_1 \ x_2 \ldots ]$
- Learn a function $f(x) = [f_1(x) \ f_2(x) \ldots ]$
- $f(x)$ should map $x$ to a more useful (typically, smaller) representation
- $f(x)$ should discard the noise (nuisance variables)

Some representation learning algorithms:

- Principal components analysis (PCA)
- Linear discriminant analysis (LDA)
- Deep autoencoders
Multi-view representation learning

Training data: samples of a $d$-dimensional random vector that has some natural split into two sub-vectors

$$\begin{bmatrix} x \\ y \end{bmatrix}, \ x \in \mathbb{R}^{d_x}, \ y \in \mathbb{R}^{d_y}, \ d_x + d_y = d$$

- Multi-view representation learning: Find representations of each view that are predictive of the other, or that are common to both
- Intuition: If the noise/nuisance parameters in the two views are independent, then the shared information must be signal!
- At test time, all views or only a subset may be available
Multiple views of speech

**Figure credits:** [Schultz & Wand *Sp. Comm.* 2009, Zhu+ *Interspeech* 2007, Lingala+ *Mag. Res. Med.* 2016, Saenko+ *PAMI* 2009, Paula West]
Family with 2 kids and a dog
Several kinds of winter squash
The Earth as seen from space

Figure credits: [http://www.fmsasg.com, http://www.bibleexpo.com]
Method 1: Canonical correlation analysis (CCA)

[Hotelling 1936]

One of the oldest and most popular multi-view techniques

- **Given**: data set of \( n \) paired vectors \( \{(x_1, y_1), \ldots, (x_n, y_n)\} \), which are samples of random vectors \( X \in \mathbb{R}^{d_x}, Y \in \mathbb{R}^{d_y} \)

- **Find**: direction vectors \( v_j, w_j, j \in \{1, \ldots, k\} \) that maximize the correlation between the projections \( v_j^T X \) and \( w_j^T Y \) while being minimally redundant

\[
\begin{align*}
    v_j, w_j &= \arg \max_{v, w} \text{corr}(v^T X, w^T Y) \\
    \text{such that} \quad \text{corr}(v_j^T X, v_k^T X) &= 0, \ k < j \\
    \text{corr}(w_j^T Y, w_k^T Y) &= 0, \ k < j
\end{align*}
\]
CCA: Toy examples

- Theoretical results (e.g., [Chaudhuri+ 2009]) show discriminative properties of CCA projections, assuming the views are uncorrelated given a class label.
Method 2: Deep CCA [Andrew+ 2013]

- Nonlinear extension of CCA
- Each view's representation is the output of a neural network
- All parameters learned jointly via backpropagation

\[
\begin{align*}
\text{max } & \text{tr}(U^T \Sigma_{12} V) \\
\text{s.t. } & U^T \Sigma_{11} V = I, \ V^T \Sigma_{22} V = I
\end{align*}
\]

\[
\begin{align*}
f(x) &= s(V_d h_{d-1} + a_d) \\
& \vdots \\
h_2 &= s(V_2 h_1 + a_2) \\
h_1 &= s(V_1 x + a_1)
\end{align*}
\]

\[
\begin{align*}
g(y) &= s(W_d l_{d-1} + b_d) \\
& \vdots \\
l_2 &= s(W_2 l_1 + b_2) \\
l_1 &= s(W_1 y + b_1)
\end{align*}
\]

Inspired by generative interpretation of CCA [Bach & Jordan 2005]
Method 4: Multi-view contrastive loss

[Hermann & Blunsom 2014]

Competitive alternative to CCA

- Try to bring paired examples closer together
- While keeping random unpaired examples farther apart by some margin

\[
\min_{f,g} \frac{1}{N} \sum_{i=1}^{N} \max \left( 0, m + \text{dist}(f(x_i^+), g(y_i^+)) - \text{dist}(f(x_i^+), g(y_i^-)) \right)
\]
Other methods

- Multi-view autoencoders [Ngiam+ 2011]
- Multimodal deep Boltzmann machines [Srivastava & Salakhutdinov 2014, Sohn+ 2014]
- ...


Toy example: Noisy MNIST digits

A synthetic dataset that perfectly satisfies the uncorrelated noise multi-view assumption
Noisy MNIST visualization [Wang+ 2015, Wang+ 2016]

Visualization via t-SNE [van der Maaten & Hinton 2008]
VCCA: Shared vs. private dimensions
Speech recognition experiments

U. Wisconsin X-ray Microbeam Database (XRMB) [Westbury+ 1994]
Phonetic recognition results

Cross-domain phonetic recognition [Tang+ 2018]

- Would like to use the learned features on typical acoustic-only data sets
- **Approach**: Multi-task learning combining the multi-view loss with recognizer loss on target domain
Multi-lingual word embedding learning

[Lu+ 2015, Wang+ 2015]

English

word vector 1

foul  magnificent  cute
beastly  awful  horrid
ugly  grotesque  gorgeous
marvelous  charming  splendid
elegant

German

word vector 2

haassliche  bezaubernder
foul  abscheulichen
ziemlich  aufzuklaaren
grotesk  schrecklichen
grobbartige  elegante
gebot  clever
hervorragende  blonden
wunderbaren
• Different languages provide different views of a single “concept”
• Consider pairs of translationally equivalent words (e.g., \( Mr. \leftrightarrow Herr \))
• Can we improve monolingual word embeddings using both views?
• Monolingual embeddings often conflate antonyms; translational context should help!
Procedure [Faruqui & Dyer 2014]

- Start with off-the-shelf word embeddings, learned independently for each language (via LSA [Deerwester et al. 1990], word2vec, etc.)
- Do unsupervised word alignment on parallel sentences
- Extract aligned word pairs:
  - i  ich
  - and  und
  - the  die
  - mr  herr
  - correlation  Zusammenhang
  - ...
- CCA on set of paired word vectors to map them to a shared space
Word embedding results

[Lu+ 2015, Wang+ 2015]

**Table:** Spearman’s correlation ($\rho$) for bigram similarities.

<table>
<thead>
<tr>
<th>Method</th>
<th>AN</th>
<th>VN</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>45.0</td>
<td>39.1</td>
<td>42.1</td>
</tr>
<tr>
<td>CCA</td>
<td>46.6</td>
<td>37.7</td>
<td>42.2</td>
</tr>
<tr>
<td>1→2AE</td>
<td>47.0</td>
<td>45.0</td>
<td>46.0</td>
</tr>
<tr>
<td>CorrAE</td>
<td>43.0</td>
<td>42.0</td>
<td>42.5</td>
</tr>
<tr>
<td>DistAE</td>
<td>43.6</td>
<td>39.4</td>
<td>41.5</td>
</tr>
<tr>
<td>FKCCA</td>
<td>46.4</td>
<td>42.9</td>
<td>44.7</td>
</tr>
<tr>
<td>NKCCA</td>
<td>44.3</td>
<td>39.5</td>
<td>41.9</td>
</tr>
<tr>
<td>DCCA</td>
<td>48.5</td>
<td>42.5</td>
<td>45.5</td>
</tr>
<tr>
<td>DCCAE</td>
<td>49.1</td>
<td>43.2</td>
<td><strong>46.2</strong></td>
</tr>
</tbody>
</table>
Other applications [Wang+ 2016]

Image tagging

- Mean average precision

Digit classification

- Accuracy (%)
Even more applications

CCA and related methods have been used for...

- Learning word embeddings, where the views are past + present word context [Dhillon+ 2011] or word + context [Stratos+ 2015]
- Learning probabilistic context-free grammars, using inside + outside trees [Cohen+ 2012]
- Learning hidden Markov models [Hsu+ 2012]
- Localizing a sound source in video [Kidron+ 2005]
- Decoding brain signals, using stimulus + response pairs [de Cheveigné+ 2018]
Summary

It is often possible to learn better representations using multi-view learning

- CCA is often a good baseline method
- Nonlinear (deep neural) extensions can be a lot better
- Contrastive learning often a good (sometimes better) alternative
- A key step is defining the views
- Applications in speech, NLP, computer vision, neuroscience, ...

Try it at home!

- http://ttic.edu/livescu/software/dcca.tgz
- https://bitbucket.org/qingming_tang/interspeech2017_vccap
References

References