Multi-view representation learning for speech (and language)

Karen Livescu

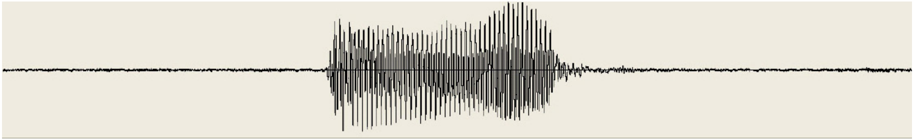
Joint work with

Qingming Tang, Weiran Wang, Raman Arora, Galen Andrew, Jeff Bilmes, ...

TOYOTA TECHNOLOGICAL INSTITUTE AT CHICAGO

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

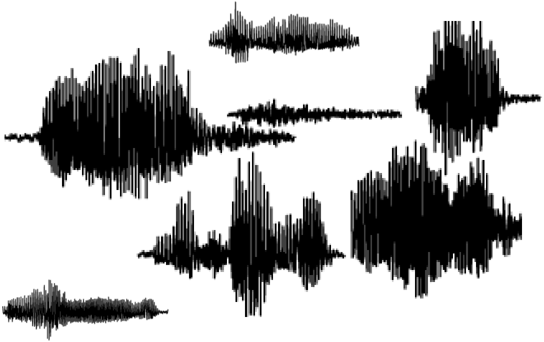
Text vs. speech

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

- haha
- hahahahahahahaha
- haaaahaaaa
- lol
- rotflmao
- lol!!!!!!!!!!!!!
- wow that is big
- that is biiiiiig
- that. is. big.
- waaaaaaay big

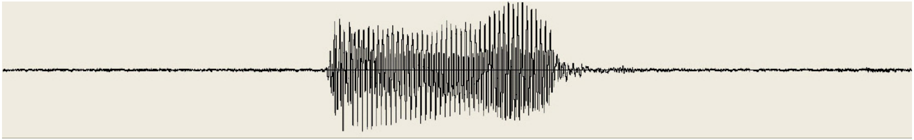
Text vs. speech

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?

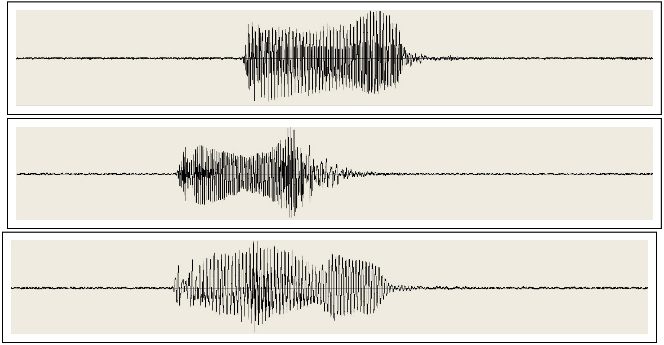


What are we looking at?

- 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. $a(t) = [3, 16, -1, 0, 427, 29, \ldots]$

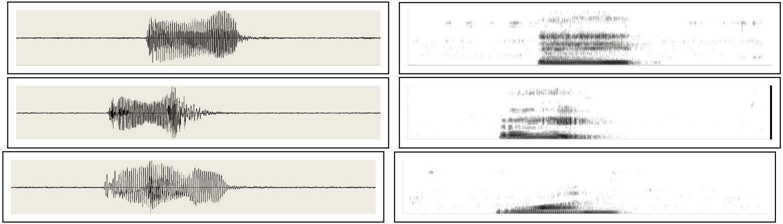
This is hard!

Which two are the same digit?



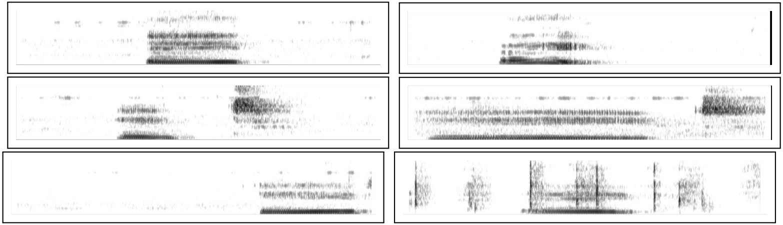
Idea: Use a frequency-domain representation

Spectrogram: Fourier transform over short windows (e.g. 20ms) $\longrightarrow$ plot of energy at each frequency over time $f_1(\omega), f_2(\omega), \ldots$

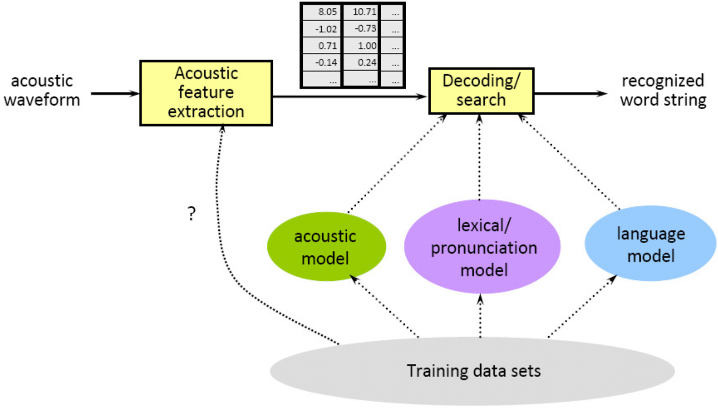


This is still hard!

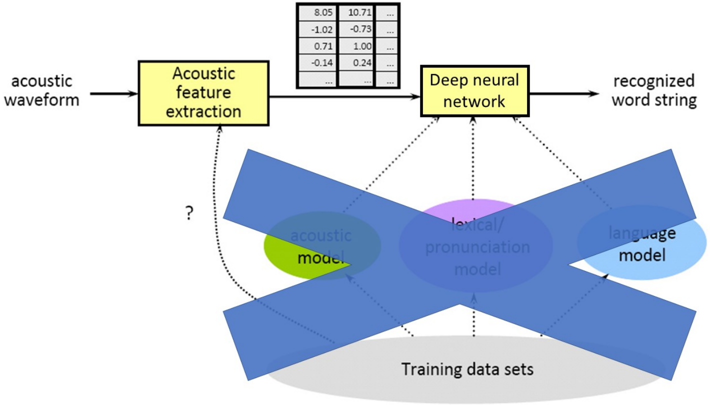
Several examples of the digit "eight"



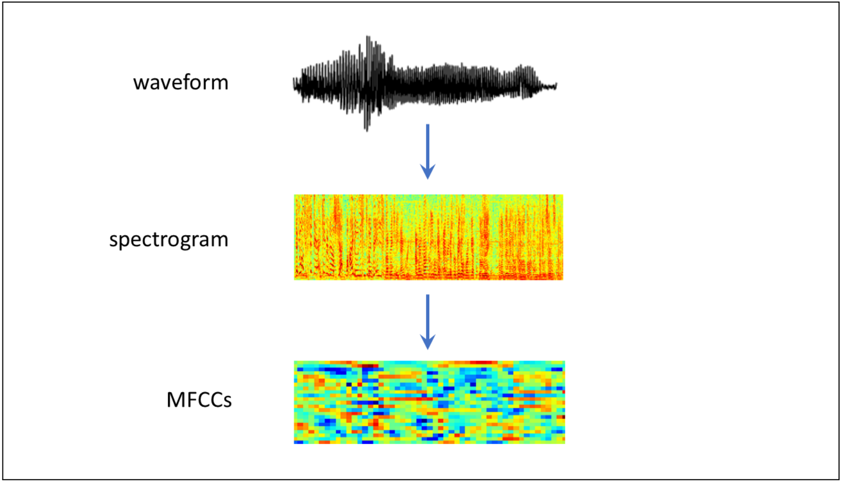
Architecture of a “traditional” speech recognizer



Architecture of an "end-to-end" speech recognizer



Acoustic features **(representations)**



Representations for text

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
- ...

TTIC

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, $\mathbf{x} = [x_1\ x_2\ \ldots]$
- Learn a function $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x})\ f_2(\mathbf{x})\ \ldots]$
- $\mathbf{f}(\mathbf{x})$ should map $\mathbf{x}$ to a more useful (typically, smaller) representation
- $\mathbf{f}(\mathbf{x})$ should discard the noise (nuisance variables)

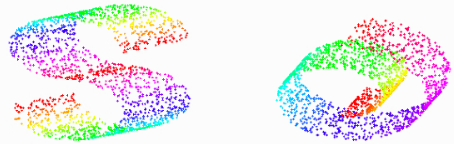
Some representation learning algorithms:

- 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

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



- **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

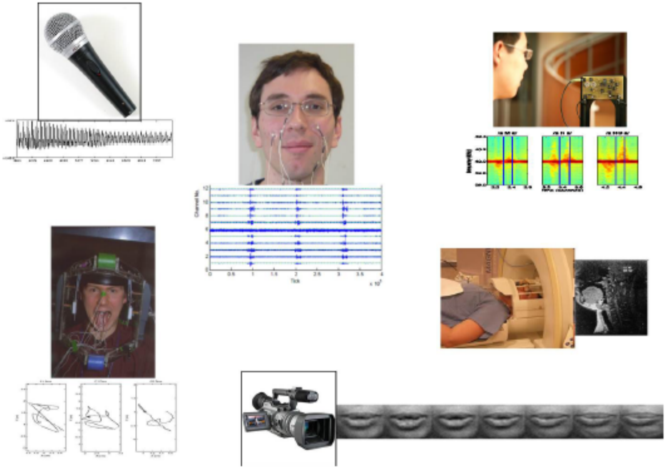
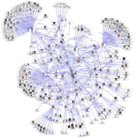


Figure credits: [Schultz & Wand *Sp. Comm.* 2009, Zhu+ Interspeech 2007, Lingala+ Mag. Res. Med. 2016, Saenko+ *PAMI* 2009, Paula West]

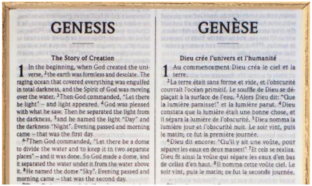
Multiple views of text



Family with 2 kids and a dog

Several kinds of winter squash

The Earth as seen from space



GENESIS

The Story of Creation

1 In the beginning, when God created the uni-
verse, 2 the earth was formless and desolate. The
raging ocean that covered everything was engulfed
in total darkness, and the Spirit of God was moving
over the water. 3 Then God commanded, "Let there
be light" – and light appeared. 4 God was pleased
with what he saw. Then he separated the light from
the darkness, 5 and he named the light "Day" and
the darkness "Night". Evening passed and morning
came – that was the first day.
6-7 Then God commanded, "Let there be a dome
to divide the water and to keep it in two separate
places" – and it was done. So God made a dome, and
it separated the water under it from the water above
it. 8 He named the dome "Sky". Evening passed and
morning came – that was the second day.

GENÈSE

Dieu crée l'univers et l'humanité

1 Au commencement Dieu créa le ciel et la
terre.
2 La terre était sans forme et vide, et l'obscurité
couvrait l'océan primitif. Le souffle de Dieu se dé-
plaçait à la surface de l'eau. 3 Alors Dieu dit: "Que
la lumière paraisse!" et la lumière parut. 4 Dieu
constata que la lumière était une bonne chose, et
il sépara la lumière de l'obscurité. 5 Dieu nomma la
lumière jour et l'obscurité nuit. Le soir vint, puis
le matin; ce fut la première journée.
6 Dieu dit encore: "Qu'il y ait une voûte, pour
séparer les eaux en deux masses!" 7 Et cela se réalisa.
Dieu fit ainsi la voûte qui sépare les eaux d'en bas
de celles d'en haut. 8 Il nomma cette voûte ciel. Le
soir vint, puis le matin; ce fut la seconde journée.

Figure credits: [http://www.fmsasg.com, http://www.bibleexpo.com]

Method 1: Canonical correlation analysis (CCA)

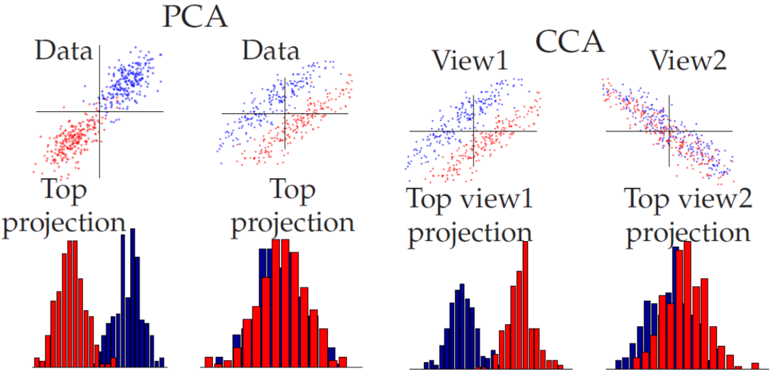
[Hotelling 1936]

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

$$
\begin{aligned}
v_j, w_j = & \quad \arg\max_{v,w} \ \mathrm{corr}(v^T X, w^T Y) \\
\text{such that} & \quad \mathrm{corr}(v_j^T X, v_k^T X) = 0, \ k < j \\
& \quad \mathrm{corr}(w_j^T Y, w_k^T Y) = 0, \ k < j
\end{aligned}
$$

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

$$\max \operatorname{tr}(U^T \Sigma_{12} V)$$

$$\text{s.t.} \quad U^T \Sigma_{11} V = I \,, \; V^T \Sigma_{22} V = I$$

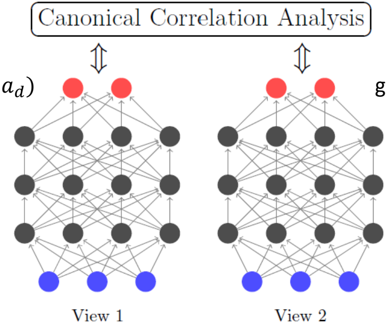
$f(x) = s(V_d h_{d-1} + a_d)$

$\vdots$

$h_2 = s(V_2 h_1 + a_2)$

$h_1 = s(V_1 x + a_1)$

x



$g(y) = s(W_d l_{d-1} + b_d)$

$\vdots$

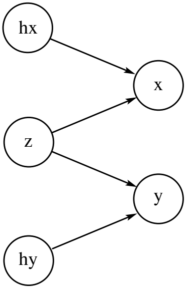
$l_2 = s(W_2 l_1 + b_2)$

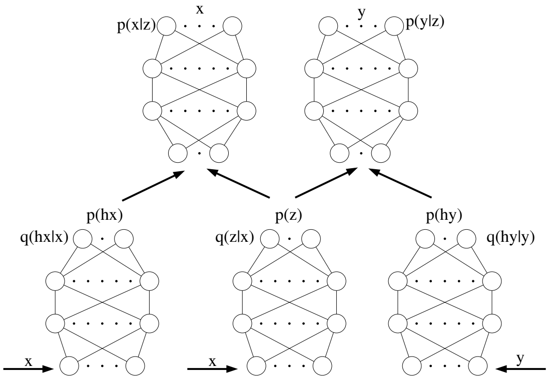
$l_1 = s(W_1 y + b_1)$

y

Method 3: Deep variational CCA [Wang+ 2016, Tang+ 2017]

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

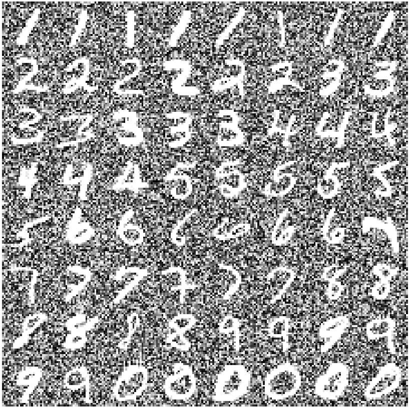
$$\min_{f,g} \frac{1}{N} \sum_{i=1}^{N} \max\left(0, m + \mathrm{dist}(f(x_i^+), g(y_i^+)) - \mathrm{dist}(f(x_i^+), g(y_i^-))\right)$$

Other methods

- Kernel CCA [Lai & Fyfe 2000, Akaho+ 2001, Melzer+ 2001]
- Multi-view autoencoders [Ngiam+ 2011]
- Multimodal deep Boltzmann machines [Srivastava & Salakhutdinov 2014, Sohn+ 2014]
- ...

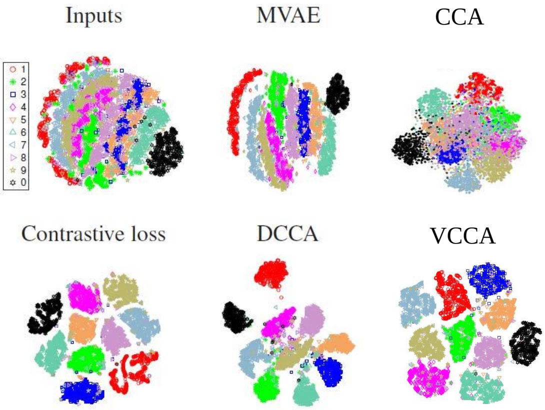
Toy example: Noisy MNIST digits

A synthetic dataset that perfectly satisfies the uncorrelated noise multi-view assumption

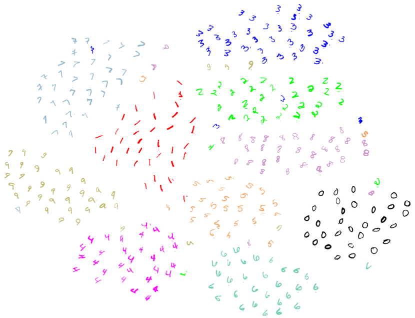


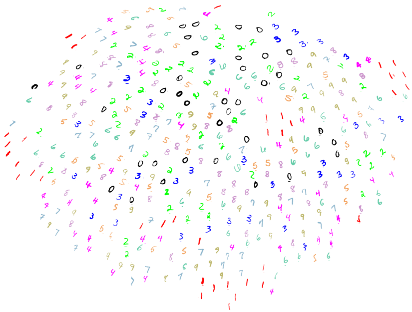
Noisy MNIST visualization [Wang+ 2015, Wang+ 2016]

Visualization via t-SNE [van der Maaten & Hinton 2008]



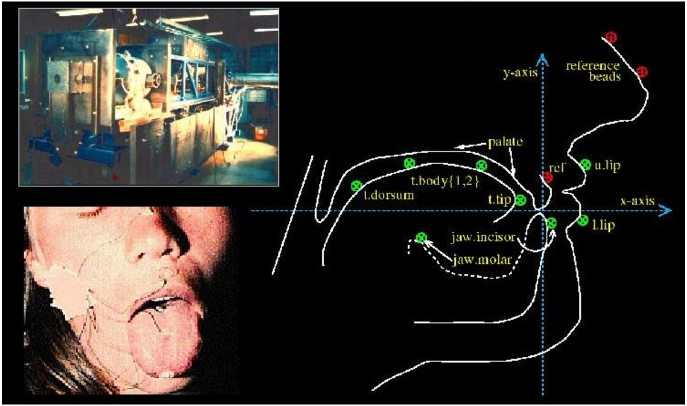
VCCA: Shared vs. private dimensions





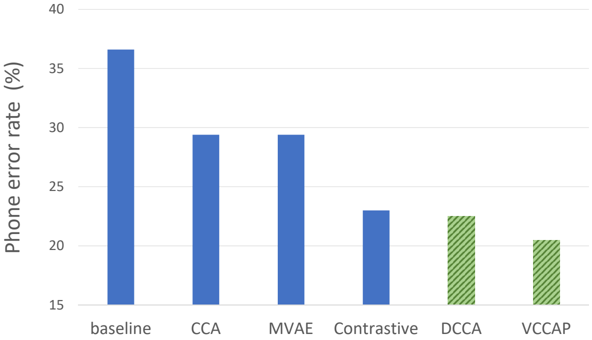
Speech recognition experiments

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



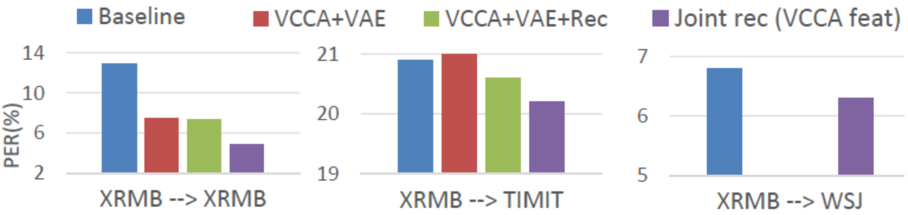
Phonetic recognition results

[Wang+ 2015, Wang+ 2016, Tang+ 2017]



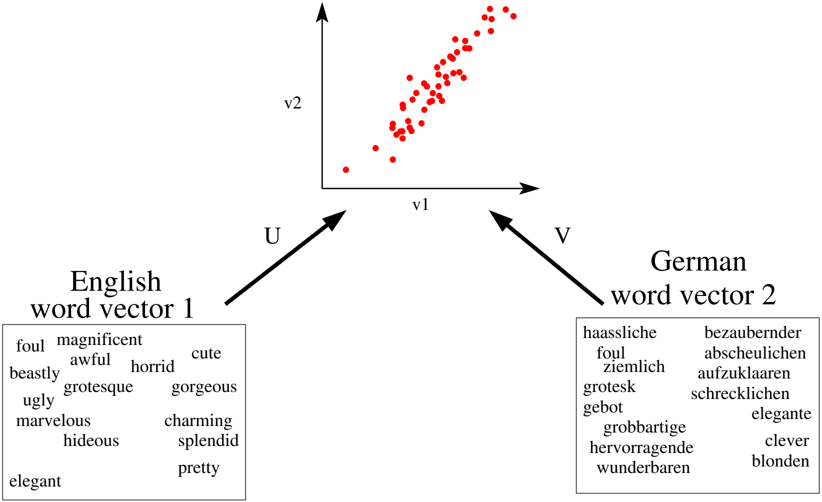
Cross-domain phonetic recognition [Tang+ 2018]

- 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]



CCA on translation pairs [Faruqui & Dyer 2014]

- 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

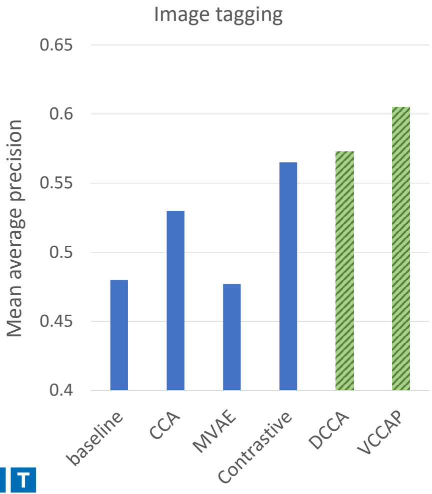
Word embedding results

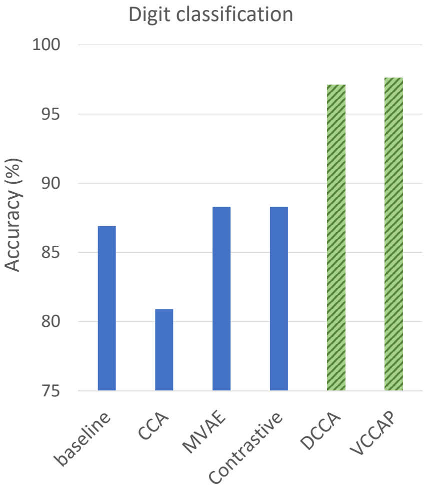
[Lu+ 2015, Wang+ 2015]

Table: Spearman's correlation (ρ) for bigram similarities.

Method	AN	VN	Avg.
Baseline	45.0	39.1	42.1
CCA	46.6	37.7	42.2
1→2AE	47.0	45.0	46.0
CorrAE	43.0	42.0	42.5
DistAE	43.6	39.4	41.5
FKCCA	46.4	42.9	44.7
NKCCA	44.3	39.5	41.9
DCCA	48.5	42.5	45.5
DCCAE	49.1	43.2	**46.2**

Other applications [Wang+ 2016]





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 Cheveign'e+ 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

- S. Akaho. *A kernel method for canonical correlation analysis.* Proc. Int'l Mtg. of the Psychometric Soc. (IMPS2001), 2001.
- G. Andrew, R. Arora, K. Livescu, and J. Bilmes. *Deep canonical correlation analysis.* ICML 2013.
- F. R. Bach and M. I. Jordan. *A probabilistic interpretation of canonical correlation analysis.* Technical Report 688, Dept. of Statistics, U. C. Berkeley, 2005.
- K. Chaudhuri, S. Kakade, K. Livescu, and K. Sridharan. *Multi-view clustering via canonical correlation analysis.* ICML 2009.
- S. B. Cohen, K. Stratos, M. Collins, D. P. Foster, and L. Ungar. *Spectral learning of latent-variable PCFGs.* ACL 2012.
- A. de Cheveigné, D. D. Wong, G. M. Di Liberto, J. Hjortkjær, M. Slaney, and E. Lalor. *Decoding the auditory brain with canonical component analysis. NeuroImage* **172**, 2018.
- P. Dhillon, D. P. Foster, and L. H. Ungar. *Multi-view learning of word embeddings via CCA.* NIPS 2011.
- M. Faruqui and C. Dyer. *Improving vector space word representations using multilingual correlation.* EACL 2014.
- K. M. Hermann and P. Blunsom. *Multilingual distributed representations without word alignment.* ICLR 2014.
- H. Hotelling. *Relations between two sets of variates. Biometrika* **28**(3/4):321–377, 1936.
- D. Hsu, S. M. Kakade, and T. Zhang. *A spectral algorithm for learning hidden Markov models. Jnl Computer & System Sci.* **78**(5), 2012.
- E. Kidron, Y. Y. Schechner, and M. Elad. *Pixels that sound.* CVPR 2005.
- P. L. Lai and C. Fyfe. *Kernel and nonlinear canonical correlation analysis. Int. J. Neural Syst.* **10**(5):365–377, 2000.
- A. Lu, W. Wang, M. Bansal, K. Gimpel, and K. Livescu. *Deep multilingual correlation for improved word embeddings.* NAACL 2015.
- T. Melzer, M. Reiter, and H. Bischof. *Nonlinear feature extraction using generalized canonical correlation analysis.* ICANN 2001.
- J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, *Multimodal deep learning.* ICML 2011.
- K. Sohn, W. Shang, and H. Lee.*Improved multimodal deep learning with variation of information.* NIPS 2014.
- N. Srivastava and R. Salakhutdinov. *Multimodal learning with deep boltzmann machines.* JMLR 2014.
- K. Stratos, M. Collins, and D. Hsu. *Model-based word embeddings from decompositions of count matrices.* ACL 2015.

References

- Q. Tang, W. Wang, and K. Livescu. *Acoustic feature learning via deep variational canonical correlation analysis.* Interspeech 2017.
- Q. Tang, W. Wang, and K. Livescu. *Acoustic feature learning using cross-domain articulatory measurements.* ICASSP 2018.
- L. van der Maaten and G. Hinton. *Visualizing data using t-SNE.* JMLR 2008.
- W. Wang, R. Arora, K. Livescu, and J. Bilmes. *Unsupervised learning of acoustic features via deep canonical correlation analysis.* ICASSP 2015.
- W. Wang, R. Arora, K. Livescu, and J. Bilmes. *On deep multi-view representation learning.* ICML 2015.
- W. Wang, R. Arora, N. Srebro, and K. Livescu. *Stochastic optimization for deep CCA via nonlinear orthogonal iterations.* 53nd Annual Allerton Conference on Communication, Control and Computing, 2015.
- W. Wang, X. Yan, H. Lee, and K. Livescu. *Deep variational canonical correlation analysis.* arXiv:1610.03454, 2016.
- J. Westbury, P. Milenkovic, G. Weismer, and R. Kent. *X-ray microbeam speech production database.* JASA 1990.