Finding Structure in Data
(Bridge) Correlational Neural Networks

B. RAVINDRAN
RECONFIGURABLE AND INTELLIGENT SYSTEMS ENGINEERING (RISE) GROUP
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
AND INTERDISCIPLINARY LABORATORY FOR DATA SCIENCES (ILDS)
INDIAN INSTITUTE OF TECHNOLOGY MADRAS

JOINT WORK WITH
SARATH CHANDAR A.P. UNIVERSITY DE MONTREAL
JANARTHANAN R., IIT MADRAS
MITESH KHAPRA IBM IRL /IIT MADRAS
HUGO LAROCHELLE TWITTER CORTEX, USA.,
UNIVERSITE DE SHERBROOKE
We live in an increasingly multilingual multimodal world.
Learning common representations for multiple views
(i) reconstruct a missing view

(ii) Transfer learning between views

(iii) Cross view retrieval/matching

Why learn such common representations?
Outline

• Related Work
• Proposed Model: CorrNet
• Analysis of the model
• Application to cross language learning
• Beyond 2 views – Bridge CorrNet
• Results
Outline

• Related Work
• Proposed Model: CorrNet
• Analysis of the model
• Application to cross language learning
• Beyond 2 views – Bridge CorrNet
• Results
Any common representation learning method would need this...
Parallel data between two views

Canonical Correlation Analysis (CCA)
This looks good... then why go beyond CCA?

• Scalability issues
• Non-trivial to extend beyond 2 views
• Lacks reconstruction capabilities
  – $A$ and $B$ need not be invertible
  – Only a low rank approximate reconstruction would be possible
Outline

• Related Work
• **Proposed Model: CorrNet**
• Analysis of the model
• Application to cross language learning
• Beyond 2 views – Bridge CorrNet
• Results
Wishlist

(i) good self reconstruction

(ii) good cross reconstruction

(iii) High correlation

(iv) scalable

(v) easily extendable to multiple views
Background: a neural network based single view autoencoder

\[ h(X) = f(X) = \sigma(WX + b) \]  

encoder

\[ X' = g(h(X)) = \sigma(W'h(X) + b') \]  

decoder

\[ \min \sum_{i=1}^{N} (X_i - g(h(X_i)))^2 \]  

use backpropagation
**encoder**

\[ h_x(X) = f_x(X) = \sigma(W_x X + b) \]

\[ h_y(Y) = f_y(Y) = \sigma(W_y Y + b) \]

**decoder**

\[ X' = g_x(h_x(X)) = \sigma(W'_x h_x(X) + b') \]

\[ X' = g_y(h_y(Y)) = \sigma(W'_y h_y(Y) + b') \]
**encoder**

\[ h_x(X) = f_x(X) = \sigma(W_x X + b) \]
\[ h_y(Y) = f_y(Y) = \sigma(W_y Y + b) \]

**decoder**

\[ X' = g_x(h_x(X)) = \sigma(W'_x h_x(X) + b') \]
\[ X' = g_y(h_y(Y)) = \sigma(W'_y h_y(Y) + b') \]

**A multiview autoencoder**

\[ \min \sum_{i=1}^{N} (X_i - g_x(f_x(X_i)))^2 \]
**Correlational Neural Network**

**encoder**

\[ h_x(X) = f_x(X) = \sigma(W_x X + b) \]
\[ h_y(Y) = f_y(Y) = \sigma(W_y Y + b) \]

**decoder**

\[ X' = g_x(h_x(X)) = \sigma(W'_x h_x(X) + b') \]
\[ X' = g_y(h_y(Y)) = \sigma(W'_y h_y(Y) + b') \]

\[
\min \sum_{i=1}^{N} \left( X_i - g_x(f_x(X_i)) \right)^2 + \sum_{i=1}^{N} \left( Y_i - g_y(f_y(Y_i)) \right)^2
\]
A multiview autoencoder

**encoder**

\[ h_x(X) = f_x(X) = \sigma(W_x X + b) \]
\[ h_y(Y) = f_y(Y) = \sigma(W_y Y + b) \]

**decoder**

\[ X' = g_x(h_x(X)) = \sigma(W'_x h_x(X) + b') \]
\[ X' = g_y(h_y(Y)) = \sigma(W'_y h_y(Y) + b') \]

\[
\min \sum_{i=1}^{N} (X_i - g_x(f_x(X_i)))^2 \\
+ \sum_{i=1}^{N} (Y_i - g_y(f_y(Y_i)))^2 \\
+ \sum_{i=1}^{N} (X_i - g_x(f_y(Y_i)))^2
\]
**encoder**

\[ h_x(X) = f_x(X) = \sigma(W_x X + b) \]
\[ h_y(Y) = f_y(Y) = \sigma(W_y Y + b) \]

**decoder**

\[ X' = g_x(h_x(X)) = \sigma(W'h_x(X) + b') \]
\[ X' = g_y(h_y(Y)) = \sigma(W'h_y(Y) + b') \]

\[
\begin{align*}
\text{A multiview autoencoder} \\
\text{min} & \sum_{i=1}^{N} (X_i - g_x(f_x(X_i)))^2 \\
& + \sum_{i=1}^{N} (Y_i - g_y(f_y(Y_i)))^2 \\
& + \sum_{i=1}^{N} (X_i - g_x(f_y(Y_i)))^2 \\
& + \sum_{i=1}^{N} (Y_i - g_y(f_x(X_i)))^2
\end{align*}
\]
**encoder**

\[
h_x(X) = f_x(X) = \sigma(W_x X + b)
\]

\[
h_y(Y) = f_y(Y) = \sigma(W_y Y + b)
\]

**decoder**

\[
X' = g_x(h_x(X)) = \sigma(W'h_x(X) + b')
\]

\[
X' = g_y(h_y(Y)) = \sigma(W'y_y(Y) + b')
\]

\[
\text{min} \sum_{i=1}^{N} (X_i - g_x(f_x(X_i)))^2 + \sum_{i=1}^{N} (Y_i - g_y(f_y(Y_i)))^2
\]

\[
+ \sum_{i=1}^{N} (X_i - g_x(f_y(Y_i)))^2 + \sum_{i=1}^{N} (Y_i - g_y(f_x(X_i)))^2
\]

**Correlational Neural Network**
So far so good…. But will the representations $h(X)$ and $h(Y)$ be correlated?

Turns out that there is no guarantee for this!
So far so good.... But will the representations $h(X)$ and $h(Y)$ be correlated?

*Turns out that there is no guarantee for this!*
So far so good.... But will the representations $h(X)$ and $h(Y)$ be correlated?

*Turns out that there is no guarantee for this!*
encoder

\[ h_x(X) = f_x(X) = \sigma(W_x X + b) \]

\[ h_y(Y) = f_y(Y) = \sigma(W_y Y + b) \]

decoder

\[ X' = g_x(h_x(X)) = \sigma(W' h_x(X) + b') \]

\[ X' = g_y(h_y(Y)) = \sigma(W' h_y(Y) + b') \]

\[ \min \sum_{i=1}^{N} (X_i - g_x(f_x(X_i)))^2 \]

\[ + \sum_{i=1}^{N} (Y_i - g_y(f_y(Y_i)))^2 \]

\[ + \sum_{i=1}^{N} (X_i - g_x(f_y(Y_i)))^2 \]

\[ + \sum_{i=1}^{N} (Y_i - g_y(f_x(X_i)))^2 \]

\[ - \text{corr}(h(X), h(Y)) \]

**Correlational Neural Network**
Step 1: Train a shallow Corrnet
Greedy Unsupervised Layerwise pretraining (Hinton et al., 2006; Bengio et al., 2007)
Greedy Unsupervised Layerwise pretraining
(Hinton et al., 2006; Bengio et al., 2007)
Greedy Unsupervised Layerwise pretraining
(Hinton et al., 2006; Bengio et al., 2007)
Greedy Unsupervised Layerwise pretraining
(Hinton et al., 2006; Bengio et al., 2007)
Greedy Unsupervised Layerwise pretraining
(Hinton et al., 2006; Bengio et al., 2007)
Greedy Unsupervised Layerwise pretraining
(Hinton et al., 2006; Bengio et al., 2007)
Greedy Unsupervised Layerwise pretraining
(Hinton et al., 2006; Bengio et al., 2007)
Greedy Unsupervised Layerwise pretraining
(Hinton et al., 2006; Bengio et al., 2007)
Step 1: Train a shallow Corrnet

Step 2: Decompose the common hidden layer but copy the weights learned earlier

Step 3: Connect a new hidden layer and re-train

Building Deep Correlational Neural Networks
Outline

• Related Work
• Proposed Model: CorrNet
  • Analysis of the model
• Application to cross language learning
• Beyond 2 views – Bridge CorrNet
• Results
Task

MNIST Digit Recognition

• Data
  – 60K train
  – 10K test
  – 28*28 pixels
  – R: 14*28 dim
  – L: 14*28 dim

• Evaluate
  – ability to learn correlated representations
  – effectiveness in transfer learning
  – ability to do self and cross reconstruction
  – whether adding layers helps

Left and right half of the image are two views of the same data
Ability to learn correlated representations

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCA</td>
<td>17.05</td>
</tr>
<tr>
<td>KCCA</td>
<td>30.58</td>
</tr>
<tr>
<td>MAE</td>
<td>24.40</td>
</tr>
<tr>
<td>CorrNet</td>
<td><strong>45.47</strong></td>
</tr>
</tbody>
</table>

- Learn 50 dimensional representations
- Calculate sum of correlation across all dimensions
Effectiveness in transfer learning

<table>
<thead>
<tr>
<th>Model</th>
<th>Left to Right</th>
<th>Right to Left</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCA</td>
<td>66.13</td>
<td>66.71</td>
</tr>
<tr>
<td>KCCA</td>
<td>70.68</td>
<td>70.83</td>
</tr>
<tr>
<td>MAE</td>
<td>68.69</td>
<td>72.54</td>
</tr>
<tr>
<td>CorrNet</td>
<td><strong>76.60</strong></td>
<td><strong>79.51</strong></td>
</tr>
<tr>
<td>Self learning</td>
<td>81.62</td>
<td>80.06</td>
</tr>
</tbody>
</table>

- Train a digit classifier using common representation of only right half of train images
- Test the classifier using common representation of only left half of test images
- and vice versa
Ability to do self and cross reconstruction

<table>
<thead>
<tr>
<th>Model</th>
<th>MSE for self reconstruction</th>
<th>MSE for cross reconstruction</th>
</tr>
</thead>
<tbody>
<tr>
<td>CorrNet</td>
<td>3.6</td>
<td>4.3</td>
</tr>
<tr>
<td>MAE</td>
<td>2.1</td>
<td>4.2</td>
</tr>
</tbody>
</table>

- CorrNet’s ability to self reconstruct is sacrificed in the interest of learning correlated representations
## Improvements using Deep CorrNet

<table>
<thead>
<tr>
<th>Model</th>
<th>Left to Right</th>
<th>Right to Left</th>
<th>Sum Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCA</td>
<td>66.13</td>
<td>66.71</td>
<td>17.05</td>
</tr>
<tr>
<td>KCCA</td>
<td>70.68</td>
<td>70.83</td>
<td>30.58</td>
</tr>
<tr>
<td>MAE</td>
<td>68.69</td>
<td>72.54</td>
<td>24.40</td>
</tr>
<tr>
<td>CorrNet-50</td>
<td>76.60</td>
<td>79.51</td>
<td>45.47</td>
</tr>
<tr>
<td>CorrNet-500-50</td>
<td>77.68</td>
<td>77.95</td>
<td><strong>47.21</strong></td>
</tr>
<tr>
<td>DCCA-500-50</td>
<td>66.41</td>
<td>64.65</td>
<td>33.00</td>
</tr>
<tr>
<td>CorrNet-500-300-50</td>
<td><strong>80.46</strong></td>
<td><strong>80.47</strong></td>
<td>45.63</td>
</tr>
<tr>
<td>DCCA-500-300-50</td>
<td>70.06</td>
<td>72.43</td>
<td>33.77</td>
</tr>
</tbody>
</table>

- The Deeper versions perform better than the shallow CorrNet
Outline

• Related Work
• Proposed Model: CorrNet
• Analysis of the model
• **Application to cross language learning**
• Beyond 2 views – Bridge CorrNet
• Results
Cross Language Learning

• Train a document classifier using labeled data in language L1

• Use this classifier to classify documents in L2
Setup

- **Language Pairs**: English<-->French, English<-->German, English<-->Spanish

- **Parallel Data**: Europarl corpus used for learning common representations

- **Labeled Data**: Reuters RCV1/RCV2 corpus with 4 classes (CCAT, ECAT, GCAT, MCAT)

- **Hyperparameters**: $\text{dim}=40$, 20 epochs, stochastic gradient descent with mini-batch size 20
Nearest cross language neighbors

<table>
<thead>
<tr>
<th>january</th>
<th>president</th>
<th>said</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>de</td>
<td>en</td>
</tr>
<tr>
<td>january</td>
<td>januar</td>
<td>president</td>
</tr>
<tr>
<td>march</td>
<td>märz</td>
<td>i</td>
</tr>
<tr>
<td>october</td>
<td>oktober</td>
<td>mr</td>
</tr>
<tr>
<td>july</td>
<td>juli</td>
<td>presidents</td>
</tr>
<tr>
<td>december</td>
<td>dezember</td>
<td>thank</td>
</tr>
<tr>
<td>1999</td>
<td>jahres</td>
<td>president-in-office</td>
</tr>
<tr>
<td>june</td>
<td>juni</td>
<td>report</td>
</tr>
<tr>
<td>month</td>
<td>1999</td>
<td>voted</td>
</tr>
</tbody>
</table>

- Nearest neighbors are either translations or semantically related words
# Results

<table>
<thead>
<tr>
<th>Model</th>
<th>En→De</th>
<th>De→En</th>
<th>En→Fr</th>
<th>Fr→En</th>
<th>En→Es</th>
<th>Es→En</th>
</tr>
</thead>
<tbody>
<tr>
<td>CorrNet</td>
<td>91.8</td>
<td>74.2</td>
<td>84.6</td>
<td>74.2</td>
<td>49.0</td>
<td>64.4</td>
</tr>
<tr>
<td>Machine Translation</td>
<td>68.1</td>
<td>67.4</td>
<td>76.3</td>
<td>71.1</td>
<td>52.0</td>
<td>58.4</td>
</tr>
<tr>
<td>Klementiev et.al. 2012</td>
<td>77.6</td>
<td>71.1</td>
<td>74.5</td>
<td>61.9</td>
<td>31.3</td>
<td>63.0</td>
</tr>
<tr>
<td>Hermann and Blunsom, 2014</td>
<td>88.1</td>
<td>79.1</td>
<td>N.A.</td>
<td>N.A.</td>
<td>N.A.</td>
<td>N.A.</td>
</tr>
<tr>
<td>Majority Class</td>
<td>46.8</td>
<td>46.8</td>
<td>22.5</td>
<td>25.0</td>
<td>15.3</td>
<td>22.2</td>
</tr>
</tbody>
</table>

![Graph showing accuracy vs. number of documents](image.png)
Outline

• Related Work
• Proposed Model
• Analysis of the model
• Application to cross language learning
  • Beyond 2 views – Bridge CorrNet
• Results
So far we have assumed that there are two views and parallel data exists between them

What if this is not the case?

What if there are $n$ views with parallel data missing between many of them?
Parallel data is available between English and French.

Can we learn a common representation for English, French, German?

Can we learn a common representation for English, French, German?
Parallel data is available between English and French.

Can we learn a common representation for English, French, German & Spanish?

and so on.....
Given n views with parallel data available only between each of these views and a pivot view, can we learn a common representation for all views?

Yes, we can!

(Bridge Correlational Neural Networks).....
Recall

for each view $x, y, z$ find non-linear projections $f_x(\cdot), f_y(\cdot), f_z(\cdot)$ such that

(i) minimize self reconstruction error

(ii) minimize cross reconstruction error

(iii) High correlation even for non-pivot views

Optimize to learn projection matrices $W_x, W_y, W_z$
\[
\begin{align*}
\text{minimize}_{w=r} & \quad \sum_{i=1}^{N_1} (g_x(f_x(x_i)) - x_i)^2 \\
& + \sum_{i=1}^{N_1} (g_y(f_x(x_i)) - y_i)^2 \\
& + \sum_{i=1}^{N_1} (g_z(f_x(x_i)) - z_i)^2 \\
& + \sum_{i=1}^{N_2} (g_x(f_y(y_i)) - x_i)^2 \\
& + \sum_{i=1}^{N_2} (g_y(f_y(y_i)) - y_i)^2 \\
& + \sum_{i=1}^{N_2} (g_z(f_y(y_i)) - z_i)^2 \\
\end{align*}
\]

\[h(x) = f_x(x) = (W_x x + b)\]
\[h(y) = f_y(y) = (W_y y + b)\]
\[h(z) = f_z(z) = (W_z z + b)\]

\[x' = g_x(h(\cdot)) = (W'_x h(\cdot) + b')\]
\[y' = g_y(h(\cdot)) = (W'_y h(\cdot) + b')\]
\[z' = g_z(h(\cdot)) = (W'_z h(\cdot) + b')\]

\[\text{corr}(h(\bar{X}), h(\bar{Y}))\]
\[\text{corr}(h(\bar{Z}), h(\bar{Y}))\]
Outline

• Related Work
• Proposed Model
• Analysis of the model
• Application to cross language learning
• Beyond 2 views – Bridge CorrNet
• Results
Cross Language Learning

Input

\[ X \]

\[ Y_1 \]

\[ N_1 \]

\[ N_2 \]

\[ Y_2 \]

\[ Z \]

Train Bridge CorrNet

Output

Classify test documents in L2

Project Labeled data to common space
Multilingual TED corpus

• One pivot language - English
• 11 non-pivot languages
• All 11 languages have labeled train/valid/test documents
• Cross language learning experiments with $\binom{11}{2}$ source-target language pairs
# Comparison with Hermann and Blunsom (2014)

<table>
<thead>
<tr>
<th>Arabic</th>
<th>German</th>
<th>Spanish</th>
<th>French</th>
<th>Italian</th>
<th>Dutch</th>
<th>Polish</th>
<th>Portuguese</th>
<th>Romanian</th>
<th>Russian</th>
<th>Turkish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>German</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>French</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Italian</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dutch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polish</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portuguese</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Romanian</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russian</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Turkish</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Comparison with Hermann and Blunsom (2014)

<table>
<thead>
<tr>
<th>Arabic</th>
<th>German</th>
<th>Spanish</th>
<th>French</th>
<th>Italian</th>
<th>Dutch</th>
<th>Polish</th>
<th>Portuguese</th>
<th>Romanian</th>
<th>Russian</th>
<th>Turkish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>German</td>
<td>55.25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>French</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Italian</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dutch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polish</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portuguese</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Romanian</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russian</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Turkish</td>
<td>-6.96</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Comparison with Hermann and Blunsom (2014)

<table>
<thead>
<tr>
<th>Language</th>
<th>Arabic</th>
<th>German</th>
<th>Spanish</th>
<th>French</th>
<th>Italian</th>
<th>Dutch</th>
<th>Polish</th>
<th>Portuguese</th>
<th>Romanian</th>
<th>Russian</th>
<th>Turkish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>28.35</td>
<td>21.85</td>
<td>6.96</td>
<td>12.66</td>
<td>7.27</td>
<td>23.73</td>
<td>20.27</td>
<td>20.99</td>
<td>16.08</td>
<td>24.69</td>
<td></td>
</tr>
<tr>
<td>German</td>
<td>55.25</td>
<td>-1.47</td>
<td>4.54</td>
<td>23.18</td>
<td>6.65</td>
<td>25.57</td>
<td>8.98</td>
<td>13.64</td>
<td>7.87</td>
<td>15.38</td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td>31.31</td>
<td>10.97</td>
<td>4.75</td>
<td>7.27</td>
<td>12.16</td>
<td>14.01</td>
<td>10.29</td>
<td>14.69</td>
<td>12.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>French</td>
<td>37.78</td>
<td>19.18</td>
<td>-0.29</td>
<td>4.35</td>
<td>13.06</td>
<td>14.19</td>
<td>10.02</td>
<td>2.05</td>
<td>3.16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Italian</td>
<td>30.28</td>
<td>-0.47</td>
<td>-0.87</td>
<td>2.75</td>
<td>6.28</td>
<td>4.50</td>
<td>8.70</td>
<td>0.49</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dutch</td>
<td>46.95</td>
<td>1.65</td>
<td>10.67</td>
<td>10.62</td>
<td>6.28</td>
<td>4.50</td>
<td>7.49</td>
<td>0.49</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polish</td>
<td>17.40</td>
<td>9.95</td>
<td>10.56</td>
<td>6.19</td>
<td>5.94</td>
<td>5.23</td>
<td>-1.86</td>
<td>1.53</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portuguese</td>
<td>21.82</td>
<td>19.18</td>
<td>4.50</td>
<td>7.66</td>
<td>4.66</td>
<td>5.93</td>
<td>-1.84</td>
<td>3.72</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Romanian</td>
<td>57.41</td>
<td>2.14</td>
<td>0.05</td>
<td>11.54</td>
<td>14.56</td>
<td>-6.96</td>
<td>7.83</td>
<td>0.14</td>
<td>-3.57</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Nearest cross language neighbors

<table>
<thead>
<tr>
<th>English word</th>
<th>Spanish</th>
<th>French</th>
<th>Italian</th>
<th>Dutch</th>
<th>Polish</th>
<th>Pt-Br</th>
<th>Rom’n</th>
<th>Russian</th>
<th>Turkish</th>
</tr>
</thead>
<tbody>
<tr>
<td>mercado</td>
<td>marché</td>
<td>mercato</td>
<td>markt</td>
<td>mercado</td>
<td>rynku</td>
<td>piaţă</td>
<td>рынок</td>
<td>пazar</td>
<td>пazar</td>
</tr>
<tr>
<td>market</td>
<td>marché</td>
<td>marché</td>
<td>marché</td>
<td>marché</td>
<td>lançadas</td>
<td>rynek</td>
<td>пiaţă</td>
<td>рынка</td>
<td>пiyasa</td>
</tr>
<tr>
<td>place</td>
<td>boursier</td>
<td>vendita</td>
<td>marktonderzoek</td>
<td>marktaandeel</td>
<td>mercados</td>
<td>gieldzie</td>
<td>piaţa</td>
<td>рынке</td>
<td>пazara</td>
</tr>
<tr>
<td>comercializan</td>
<td>markeier</td>
<td>azionario</td>
<td>marktaandeel</td>
<td>marktaandeel</td>
<td>timbuktu</td>
<td>targ</td>
<td>piaţa</td>
<td>рынок, пiaţар</td>
<td>пazarн</td>
</tr>
<tr>
<td>petróleo</td>
<td>petróleo</td>
<td>petróleo</td>
<td>petróleo</td>
<td>petróleo</td>
<td>ropy</td>
<td>petrol</td>
<td>нефти</td>
<td>петрол</td>
<td>петролн</td>
</tr>
<tr>
<td>aceite</td>
<td>l’huile</td>
<td>olio</td>
<td>olieprijzen</td>
<td>olieramp</td>
<td>óleo</td>
<td>petrol</td>
<td>нефть</td>
<td>петролю</td>
<td>петролю</td>
</tr>
<tr>
<td>petroleros</td>
<td>d’huile</td>
<td>l’olio</td>
<td>oliereijzen</td>
<td>derrammento</td>
<td>azeite</td>
<td>ropé</td>
<td>нефть</td>
<td>петролю</td>
<td>петролю</td>
</tr>
<tr>
<td>crudo</td>
<td>petrölères</td>
<td>dell’olio</td>
<td>olieramp</td>
<td>derrammento</td>
<td>oleju</td>
<td>ulei</td>
<td>maslo</td>
<td>пало</td>
<td>петролед</td>
</tr>
<tr>
<td>casa</td>
<td>maison</td>
<td>casa</td>
<td>thuis</td>
<td>casa</td>
<td>domu</td>
<td>acasă</td>
<td>домой</td>
<td>evе</td>
<td>evе</td>
</tr>
<tr>
<td>hogar</td>
<td>foyer</td>
<td>tramandare</td>
<td>huis</td>
<td>lar</td>
<td>dom</td>
<td>acasa</td>
<td>дома</td>
<td>evе</td>
<td>evе</td>
</tr>
<tr>
<td>casas</td>
<td>domicile</td>
<td>dimora</td>
<td>woning</td>
<td>casas</td>
<td>domem</td>
<td>наса</td>
<td>дом</td>
<td>ev</td>
<td>ev</td>
</tr>
<tr>
<td>hogares</td>
<td>rentre</td>
<td>casetta</td>
<td>thuisblijven</td>
<td>lares</td>
<td>domach</td>
<td>casa</td>
<td>доме</td>
<td>evе</td>
<td>evе</td>
</tr>
<tr>
<td>historia</td>
<td>l’histoire</td>
<td>storia</td>
<td>geschiedenis</td>
<td>historiа</td>
<td>historii</td>
<td>istoria</td>
<td>история</td>
<td>tarih</td>
<td>tarih</td>
</tr>
<tr>
<td>historial</td>
<td>d’histoire</td>
<td>qualcuno</td>
<td>wereldgeschiedenis</td>
<td>histórico</td>
<td>istorie</td>
<td>istore</td>
<td>историю</td>
<td>tarihîn</td>
<td>tarihîn</td>
</tr>
<tr>
<td>contado</td>
<td>histoire</td>
<td>dell’umanità</td>
<td>history</td>
<td>historia</td>
<td>historia</td>
<td>istoriei</td>
<td>история</td>
<td>tarihî</td>
<td>tarihî</td>
</tr>
<tr>
<td>participó</td>
<td>historique</td>
<td>popolo</td>
<td>historie</td>
<td>britannica</td>
<td>dziejach</td>
<td>poveste</td>
<td>история</td>
<td>tarihînde</td>
<td>tarihînde</td>
</tr>
</tbody>
</table>
Correction: We actually build one single model...

...instead of $\binom{n}{2}$ pairwise Bridge CorrNets...
What about doing this across modalities?

Say, languages and images?
Cross modal access

Task 1: Given a French caption retrieve relevant images

Task 2: Given an image retrieve relevant French captions
Solution

1. Project given image (query) into this common space

2. Project all French captions (documents) into this common space

3. Rank nearest neighbors based on Euclidean distance

Learn a common representation for English, French and Images using Bridge CorrNet
Data

**Task 1:** 1000 images (queries) and 5000 French/German captions (documents)

**Task 2:** 5000 French/German captions (queries) and 1000 images (documents)
<table>
<thead>
<tr>
<th>Model</th>
<th>Captions</th>
<th>Recall @50</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ImageToCaption</td>
<td>CaptionToImage</td>
</tr>
</tbody>
</table>
### Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Captions</th>
<th>Recall @50</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>En-Image CorrNet</td>
<td>English</td>
<td>0.452</td>
<td>0.532</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Captions</th>
<th>Recall @50</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ImageToCaption</td>
<td>CaptionToImage</td>
</tr>
<tr>
<td>En-Image CorrNet</td>
<td>English</td>
<td>0.452</td>
</tr>
<tr>
<td>BridgeMAE</td>
<td>French</td>
<td>0.069</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Captions</th>
<th>Recall @50</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ImageToCaption</td>
<td>CaptionToImage</td>
</tr>
<tr>
<td>En-Image CorrNet</td>
<td>English</td>
<td>0.452</td>
<td>0.532</td>
</tr>
<tr>
<td>BridgeMAE</td>
<td>French</td>
<td>0.069</td>
<td>0.063</td>
</tr>
<tr>
<td>2-CorrNet</td>
<td>French</td>
<td>0.109</td>
<td>0.205</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Captions</th>
<th>Recall @50</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ImageToCaption</td>
<td>CaptionToImage</td>
<td></td>
</tr>
<tr>
<td>En-Image CorrNet</td>
<td>English</td>
<td>0.452</td>
<td>0.532</td>
<td></td>
</tr>
<tr>
<td>BridgeMAE</td>
<td>French</td>
<td>0.069</td>
<td>0.063</td>
<td></td>
</tr>
<tr>
<td>2-CorrNet</td>
<td>French</td>
<td>0.109</td>
<td>0.205</td>
<td></td>
</tr>
<tr>
<td>Bridge CorrNet</td>
<td>French</td>
<td>0.327</td>
<td>0.232</td>
<td></td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Captions</th>
<th>Recall @50</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ImageToCaption</td>
<td>CaptionToImage</td>
<td></td>
</tr>
<tr>
<td>En-Image CorrNet</td>
<td>English</td>
<td>0.452</td>
<td>0.532</td>
<td></td>
</tr>
<tr>
<td>BridgeMAE</td>
<td>French</td>
<td>0.069</td>
<td>0.063</td>
<td></td>
</tr>
<tr>
<td>2-CorrNet</td>
<td>French</td>
<td>0.109</td>
<td>0.205</td>
<td></td>
</tr>
<tr>
<td>Bridge CorrNet</td>
<td>French</td>
<td>0.327</td>
<td>0.232</td>
<td></td>
</tr>
<tr>
<td>CorrNet + MT</td>
<td>French</td>
<td>0.430</td>
<td>0.416</td>
<td></td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Captions</th>
<th>Recall @50</th>
<th>ImageToCaption</th>
<th>CaptionToImage</th>
</tr>
</thead>
<tbody>
<tr>
<td>En-Image CorrNet</td>
<td>English</td>
<td>0.452</td>
<td></td>
<td>0.532</td>
</tr>
<tr>
<td>BridgeMAE</td>
<td>French</td>
<td>0.069</td>
<td>0.063</td>
<td></td>
</tr>
<tr>
<td>2-CorrNet</td>
<td>French</td>
<td>0.109</td>
<td>0.205</td>
<td></td>
</tr>
<tr>
<td>Bridge CorrNet</td>
<td>French</td>
<td>0.327</td>
<td>0.232</td>
<td></td>
</tr>
<tr>
<td>CorrNet + MT</td>
<td>French</td>
<td>0.430</td>
<td>0.416</td>
<td></td>
</tr>
<tr>
<td>BridgeMAE</td>
<td>German</td>
<td>0.053</td>
<td>0.058</td>
<td></td>
</tr>
<tr>
<td>2-CorrNet</td>
<td>German</td>
<td>0.065</td>
<td>0.098</td>
<td></td>
</tr>
<tr>
<td>Bridge CorrNet</td>
<td>German</td>
<td>0.298</td>
<td>0.183</td>
<td></td>
</tr>
<tr>
<td>CorrNet + MT</td>
<td>German</td>
<td>0.431</td>
<td>0.343</td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td></td>
<td>0.050</td>
<td>0.047</td>
<td></td>
</tr>
</tbody>
</table>
Zwei Pferde stehen auf einem sandigen Strand nahe dem Ocean.
(Two horses standing on a sandy beach near the ocean)

grasende Pferde auf einer trockenen Weide bei einem Flughafen. (Horses grazing in a dry pasture by an airport.)

ein Elefant, Wasser auf seinen Rückend sprühend, in einem staubigen Bereich neben einem Baum. (A elephant spraying water on its back in a dirt area next to tree.)

ein braunes pferd ißt hohes gras neben einem behälter mit wasser. (Brown horses eating tall grass beside a body of water.)

vier Pferde grasen auf ein Feld mit braunem gras. (Four horses are grazing through a field of brown grass.)
Top-5 nearest French captions

Un homme portant une batte de baseball à deux mains lors d'un jeu de balle professionnel. (A man holding a baseball bat with two hands at a professional ball game.)

un joueur de tennis balance une raquette à une balle. (A tennis player swinging a racket at a ball.)

un garçon qui est de frapper une balle avec une batte de baseball. (A boy that is hitting a ball with a baseball bat.)

une équipe de joueurs de baseball jouant un jeu de base-ball. (A team of baseball players playing a game of baseball.)

un garçon se prépare à frapper une balle de tennis avec une raquette. (A boy prepares to hit a tennis ball with a racquet.)
German Query

Speisen und Getränke auf einem Tisch mit einer Frau essen im Hintergrund.

(Food and beverages set on a table with a woman eating in the background.)

Top-5 nearest images

- a woman is seen eating at an outdoor table with a bowl of soup and a roll and a beer in the foreground in front of her
- a girl playing with a toy on a room floor
- a woman sitting in front of a cake holding her hands
- a man and woman cutting a cake with a large knife
- a baby girl sitting in front of a small cupcake with a candle
French Query

de ski en se tenant debout dans la neige.

(People wearing ski equipment while standing in snow.)

Top-5 nearest images

two people wearing skis standing next to each other

a man that is on skis standing in the snow.

a toddler is on a small surfboard at the shore.

a man riding a snow board down a snow covered slope.

a man riding a snow board down a snow covered slope.
Thank You!!


http://www.cse.iitm.ac.in/~ravi
http://cmsrv.iitm.ac.in/ilds/home