Constrained Non-Affine Alignment of Embeddings

Yuwei Wang, Yan Zheng, Yanqing Peng, Michael Yeh, Zhongfang Zhuang, Das Mahashweta, Bendre Mangesh, Feifei Li, Wei Zhang, Jeff M. Phillips

University of Utah, Visa Research
Embeddings

What are embeddings?
An embedding is moderate-dimensional vector representation of an entity where many features can be captured.

Text

Word embedding methods

woman: [0, 1.2, 4.7, 9, ..., 2.4]
girl:    [0.5, 2, 4.2, 3, ..., 1.1]
teacher: [8, 5.4, 1.1, 6.5, ..., 0.5]
...


Embeddings

Embeddings preserve the information of entities by placing similar entities close together in the embedding space, as measured by cosine distance.
Embeddings

However, it is hard to tease apart features since all the features are entangled together, and there is no simple mapping between the dimensions and the features.

How can we measure the significance of a feature (e.g. polarity)?
Feature Measurement

Our scenario

• We can access a subset of features in the original data.
• We measure how significantly an embedding is affected by a known categorical feature.
• We quantify the significance using a classifier.
Feature Measurement

We have dataset $D$ and a balanced feature $F$ with $M$ labels. For fields using numerical values, we set thresholds to label the values into different categories.

Consider an embedding generator $E : D \to \mathbb{R}^d$ and a balanced feature $F : D \to \{0, 1, ..., M - 1\}$. For a family of classifiers $\mathcal{C}$ on the embedding space, and a positive value $\varepsilon$ with the following probability:

$$\max_{C \in \mathcal{C}} \text{Prob}_{x \in D}[C(E(x)) = F(x)] > \frac{1}{M} + \varepsilon$$

We say that $E$ embeds $F$ with weight $\varepsilon$. 
Feature Measurement

We have dataset $D$ and a balanced feature $F$ with $M$ labels. For fields using numerical values, we set thresholds to label the values into different categories.

Consider an embedding generator $E : \mathcal{D} \to \mathbb{R}^d$ and a balanced binary feature $F : \mathcal{D} \to \{0, 1\}$. For a family of classifiers $C$ on the embedding space, and a positive value $\varepsilon$ with the following probability:

$$\max_{C \in C} \text{Prob}_{x \in \mathcal{D}} [C(E(x)) = F(x)] > 50\% + \varepsilon$$

We say that $E$ embeds $F$ with weight $\varepsilon$. 
Embedding Attenuation and Retention

Given an undesired binary feature \( F \) on dataset \( D \), two subsets \( A \) and \( B \) with corresponding embeddings, how can we align them by removing the effect of the \( F \)?
Embedding Attenuation and Retention

Our proposed method UCAN (Unsupervised Constrained Alignment that is None-Affine)
Embedding Attenuation and Retention

Our proposed method UCAN (Unsupervised Constrained Alignment that is None-Affine)
Embedding Attenuation and Retention

Our proposed method UCAN (Unsupervised Constrained Alignment that is None-Affine)
Embedding Attenuation and Retention

Our proposed method UCAN (Unsupervised Constrained Alignment that is None-Affine)
Embedding Attenuation and Retention

Our proposed method UCAN (Unsupervised Constrained Alignment that is None-Affine)
Embedding Attenuation and Retention

Our proposed method UCAN (Unsupervised Constrained Alignment that is None-Affine)

structure preservation component
Experiments

Synthetic datasets
Real-world datasets
Applications
Experiments

Real datasets
We use three real-world datasets and choose MUSE and UMWE as our baselines.
Experiments

Real datasets
- airport embedding dataset
- multi-language embeddings dataset
- merchant embedding dataset
Experiments

Multi-language embeddings dataset
The unsupervised word vectors were trained using FastText. The languages focused on in our experiments are English (en), Spanish (es), French (fr), German (de), Russian (ru) and Italian (it).
We use the standard K-nearest neighbor (NN) and Cross-Lingual Similarity Scaling (CSLS) as the evaluation approaches. We measure how many times one of the correct translations of a source word are retrieved, and report the precisions for $K = 1, 5, 10$. 
Experiments

Multi-language embeddings dataset
Our goal is to remove the “language’” feature, and retain the meanings of the words from different languages.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MUSE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K@1</td>
<td>69.07</td>
<td>64.40</td>
<td>53.96</td>
<td>61.6</td>
<td>61.02</td>
<td>51.40</td>
<td>24.20</td>
<td>32.00</td>
<td>56.45</td>
<td>59.33</td>
</tr>
<tr>
<td>K@5</td>
<td>82.93</td>
<td>78.40</td>
<td>65.08</td>
<td>75.93</td>
<td>71.19</td>
<td>66.60</td>
<td>44.20</td>
<td>50.00</td>
<td>66.13</td>
<td>73.73</td>
</tr>
<tr>
<td>K@10</td>
<td>86.87</td>
<td>81.8</td>
<td>68.25</td>
<td>80.60</td>
<td>72.88</td>
<td>71.93</td>
<td>51.40</td>
<td>56.20</td>
<td>69.35</td>
<td>81.13</td>
</tr>
<tr>
<td><strong>UWME</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K@1</td>
<td>64.13</td>
<td>61.73</td>
<td>62.47</td>
<td>61.27</td>
<td>52.40</td>
<td>51.73</td>
<td>27.47</td>
<td>39.13</td>
<td>56.73</td>
<td>57.27</td>
</tr>
<tr>
<td>K@5</td>
<td>76.20</td>
<td>75.80</td>
<td>75.73</td>
<td>75.67</td>
<td>74.27</td>
<td>67.33</td>
<td>52.33</td>
<td>57.67</td>
<td>71.67</td>
<td>71.27</td>
</tr>
<tr>
<td>K@10</td>
<td>79.47</td>
<td>79.60</td>
<td>79.60</td>
<td>79.47</td>
<td>79.13</td>
<td>71.87</td>
<td>60.87</td>
<td>63.53</td>
<td>76.33</td>
<td>75.33</td>
</tr>
<tr>
<td><strong>UCAN</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K@1</td>
<td>68.93</td>
<td>72.00</td>
<td>69.60</td>
<td>69.00</td>
<td>60.27</td>
<td>61.80</td>
<td>25.73</td>
<td>46.60</td>
<td>61.33</td>
<td>63.07</td>
</tr>
<tr>
<td>K@5</td>
<td>81.26</td>
<td>84.60</td>
<td>82.93</td>
<td>83.13</td>
<td>79.33</td>
<td>76.26</td>
<td>51.33</td>
<td>65.73</td>
<td>77.93</td>
<td>79.07</td>
</tr>
<tr>
<td>K@10</td>
<td>85.00</td>
<td>88.40</td>
<td>85.80</td>
<td>87.20</td>
<td>83.27</td>
<td>80.80</td>
<td>59.80</td>
<td>70.26</td>
<td>82.87</td>
<td>82.53</td>
</tr>
</tbody>
</table>
Experiments

Multi-language embeddings dataset

We also demonstrate the results on languages which are less similar to English, including Greek (el), Vietnamese (vi), Arabic (ar), Czech (cs) and Dutch (nl).

<table>
<thead>
<tr>
<th></th>
<th>MUSE</th>
<th>UWME</th>
<th>UCAN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>K@1</td>
<td>K@5</td>
<td>K@10</td>
</tr>
<tr>
<td></td>
<td>en-el</td>
<td>el-en</td>
<td>en-vi</td>
</tr>
<tr>
<td>K@1</td>
<td>13.87</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>K@5</td>
<td>31.13</td>
<td>0.00</td>
<td>0.13</td>
</tr>
<tr>
<td>K@10</td>
<td>38.33</td>
<td>0.00</td>
<td>0.33</td>
</tr>
<tr>
<td>K@1</td>
<td>18.73</td>
<td>26.27</td>
<td>3.27</td>
</tr>
<tr>
<td>K@5</td>
<td>36.47</td>
<td>44.80</td>
<td>8.40</td>
</tr>
<tr>
<td>K@10</td>
<td>43.27</td>
<td>50.53</td>
<td>10.73</td>
</tr>
<tr>
<td>K@1</td>
<td>28.20</td>
<td>41.47</td>
<td>12.13</td>
</tr>
<tr>
<td>K@5</td>
<td>46.60</td>
<td>59.93</td>
<td>23.27</td>
</tr>
<tr>
<td>K@10</td>
<td>52.53</td>
<td>65.27</td>
<td>27.67</td>
</tr>
</tbody>
</table>
Experiments

Real datasets
- airport embedding dataset
- multi-language embeddings dataset
- merchant embedding dataset
Experiments

Merchant embedding dataset
The merchant embedding is generated from a real-world transaction dataset involving 70 million merchants and 26-million customers. We focus on two features: the location and the merchant category code (MCC).
Experiments

Merchant embedding dataset

Our goal is to retain the merchant category code feature (F1) and remove the location feature (F2).

<table>
<thead>
<tr>
<th>ORG</th>
<th>UCAN</th>
<th>MUSE</th>
<th>UWME</th>
</tr>
</thead>
<tbody>
<tr>
<td>O_F1</td>
<td>O_F2</td>
<td>O_F1/O_F2</td>
<td>C_F1</td>
</tr>
<tr>
<td>LA→SF</td>
<td>0.64</td>
<td>0.88</td>
<td>0.73</td>
</tr>
<tr>
<td>SF→LA</td>
<td>0.64</td>
<td>0.88</td>
<td>0.73</td>
</tr>
<tr>
<td>LA→CHI</td>
<td>0.58</td>
<td>0.95</td>
<td>0.61</td>
</tr>
<tr>
<td>CHI→LA</td>
<td>0.58</td>
<td>0.95</td>
<td>0.61</td>
</tr>
<tr>
<td>SF→CHI</td>
<td>0.65</td>
<td>0.93</td>
<td>0.70</td>
</tr>
<tr>
<td>CHI→SF</td>
<td>0.65</td>
<td>0.93</td>
<td>0.70</td>
</tr>
<tr>
<td>MAN→LA</td>
<td>0.61</td>
<td>0.90</td>
<td>0.68</td>
</tr>
<tr>
<td>LA→MAN</td>
<td>0.61</td>
<td>0.90</td>
<td>0.68</td>
</tr>
<tr>
<td>MAN→SF</td>
<td>0.66</td>
<td>0.88</td>
<td>0.75</td>
</tr>
<tr>
<td>SF→MAN</td>
<td>0.66</td>
<td>0.88</td>
<td>0.75</td>
</tr>
<tr>
<td>MAN→CHI</td>
<td>0.62</td>
<td>0.93</td>
<td>0.67</td>
</tr>
<tr>
<td>CHI→MAN</td>
<td>0.62</td>
<td>0.93</td>
<td>0.67</td>
</tr>
</tbody>
</table>
Experiments

Synthetic datasets
Real-world datasets
Applications
Experiments

Application: on the merchant embedding dataset

• False Merchant Identity Detection
• Cross-City Restaurant Recommendations
Experiments

False Merchant Identity Detection
Check whether a merchant's self-reported merchant category is within the top-kth most likely merchant. If it is not, the merchant will be reported as a suspicious merchant.
We trained our model with Los Angeles's merchants and tested the model on San Francisco's merchants.
Experiments

False Merchant Identity Detection
Experiments

Application

• False Merchant Identity Detection
• Cross-City Restaurant Recommendations
Experiments

Cross-City Restaurant Recommendations
Our goal is to recommend Los Angeles (LA) restaurants to the customers who usually have restaurant transactions in San Francisco (SF).
Experiments

Cross-City Restaurant Recommendations
Our original Word2Vec embeddings give a score of 60.36%. The score increases to 63.60% after UCAN removes the location feature from the embeddings, while MUSE does not improve the score.
Thank you

ywang@cs.utah.edu
Video Format
• Video resolution: 1280p Frame Width x 720p Frame Height
• Frame Rate: 25 Frames / Second
• File Type: mp4
• We suggest to authors to use for the creation of their video the ZOOM recording feature that by default support the aforementioned specs

• Video durations are specified in the following table:
• 7+5QA
Word Embeddings

One hot vectors or bag of words embeddings

• 100K or more dimensional vectors
• sparse but inefficient
• including useful semantic and syntactic information
Word Embeddings

Word2Vec, GloVe, FastText

- distributed embeddings
- about ~300 dimensions
- additional useful linear relationships

![Diagram showing semantic relationships between words](diagram.png)
Word Embeddings

ELMo, BERT
• contextual embeddings
• about 3000 (3*1024) dimensions
• distinguishable word senses
What is embeddings?
An embedding is a relatively low-dimensional space into which you can translate high-dimensional vectors.
Experiments

Synthetic datasets
Datasets have two dimensions – the color feature (the x-axis) and the lightness feature (the y-axis) and each corresponds to one type of binary feature. In this experiment, we aim to remove the color feature (F2) and retain the lightness feature (F1).
Experiments

Synthetic datasets
- balanced dataset
- unbalanced dataset
Experiments

Synthetic datasets
- balanced dataset
- unbalanced dataset
Experiments

Real dataset

- airport embedding dataset
- multi-language embeddings dataset
- merchant embedding dataset
Experiments

Airport embedding dataset
The air-traffic networks dataset includes three air-traffic networks [3] from three countries. We consider two features: the level of activity (F1) with 4 classes, and country location (F2) with 3 classes.
Experiments

Airport embedding dataset
We aim to retain the level of activity (F1) and remove country location (F2).

<table>
<thead>
<tr>
<th></th>
<th>O_F1</th>
<th>O_F2</th>
<th>O_F1/O_F2</th>
<th>C_F1</th>
<th>C_F2</th>
<th>C_F1/C_F2</th>
<th>M_F1</th>
<th>M_F2</th>
<th>M_F1/M_F2</th>
<th>U_F1</th>
<th>U_F2</th>
<th>U_F1/U_F2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazil→Europe</td>
<td>0.80</td>
<td>1</td>
<td>0.80</td>
<td>0.77</td>
<td>0.84</td>
<td>0.92</td>
<td>0.76</td>
<td>0.97</td>
<td>0.78</td>
<td>0.57</td>
<td>0.71</td>
<td>NA</td>
</tr>
<tr>
<td>Europe→Brazil</td>
<td>0.80</td>
<td>1</td>
<td>0.80</td>
<td>0.79</td>
<td>0.86</td>
<td>0.92</td>
<td>0.82</td>
<td>0.95</td>
<td>0.86</td>
<td>0.57</td>
<td>0.80</td>
<td>NA</td>
</tr>
<tr>
<td>Brazil→USA</td>
<td>0.85</td>
<td>1</td>
<td>0.85</td>
<td>0.85</td>
<td>0.61</td>
<td>1.39</td>
<td>0.83</td>
<td>0.98</td>
<td>0.85</td>
<td>0.5</td>
<td>1</td>
<td>NA</td>
</tr>
<tr>
<td>USA→Brazil</td>
<td>0.85</td>
<td>1</td>
<td>0.85</td>
<td>0.88</td>
<td>0.91</td>
<td>0.97</td>
<td>0.87</td>
<td>0.93</td>
<td>0.94</td>
<td>0.52</td>
<td>0.51</td>
<td>NA</td>
</tr>
<tr>
<td>USA→Europe</td>
<td>0.84</td>
<td>1</td>
<td>0.84</td>
<td>0.85</td>
<td>0.90</td>
<td>0.94</td>
<td>0.83</td>
<td>0.92</td>
<td>0.90</td>
<td>0.54</td>
<td>0.53</td>
<td>NA</td>
</tr>
<tr>
<td>Europe→USA</td>
<td>0.84</td>
<td>1</td>
<td>0.84</td>
<td>0.83</td>
<td>0.62</td>
<td>1.34</td>
<td>0.80</td>
<td>0.99</td>
<td>0.81</td>
<td>0.49</td>
<td>0.97</td>
<td>NA</td>
</tr>
</tbody>
</table>
Feature Measurement

Imbalanced dataset and AUC

• “Accuracy Paradox”
• We use average one vs. all AUC as the metric to measure if a feature is embedded for both balanced and imbalanced data.