HyperEmbed

Tradeoffs Between Resources and Performance in NLP Tasks with Hyperdimensional Computing enabled embedding of $n$-gram statistics

Pedro Alonso*, Kumar Shridhar*, Denis Kleyko*, Evgeny Osipov, Marcus Liwicki

PML4DC@ICLR 2020

April 2020
Main Idea

Recent advances in Deep Learning have led to a significant performance increase on several NLP tasks, however, the models become more and more computationally demanding.

This paper tackles the domain of computationally efficient algorithms for NLP tasks.
N-Gram Statistics

For each Sample

I cannot find my phone,

Normalize and Tokenize

I cannot find my phone

For each token

Create n-gram vocab (n=3)

cannot

Add to n-gram vocab

#cannot#

can
ann
not
ot#

Trigram Vocabulary

Pick a Vectorizer

Count Vectorizer
TF-IDF Vectorizer
BPE Embedding
Fast Text Vectors
...

Samples

I cannot find my phone

Dataset

Labels

Find Something

I have lost my keys

Find Something

I cannot find my phone

Create n-gram vocab (n=3)

cannot

Add to n-gram vocab

#cannot#

can
ann
not
ot#
Embedding into a HD Vector

1. Every symbol is assigned a $d$-dimensional HD vector.
2. These vectors are stored in matrix $H$.
3. A unique HD vector for each N-gram is formed using binding and permutation operations.
4. All observed N-gram are stored in a single HD vector using the bundling operation.

Three Key operation

01 Bundling $+$ (position wise addition)
02 Binding $\odot$ (position wise multiplication)
03 Permutation $\rho$

\[
\rho^1(H_\#) \odot \rho^2(H_c) \odot \rho^3(H_a)
\]
Dataset

Small Dataset

- Ask Ubuntu : 2 Intents (50 train, 50 test samples per intent)
- Chatbot Corpus : 5 Intents (10 train, 20 test samples per intent)
- WebApplication Corpus : 8 intents (3 train, 7 test samples per intent)

Large Dataset

- 20 News Corpus : 20 Intents (11,500 train, 7,500 test samples per intent)
Empirical Evaluation: ML algorithms

AskUbuntu Corpus. $N=512$

<table>
<thead>
<tr>
<th>Classifier</th>
<th>$F_1$ score</th>
<th>Resources: SH vs. HD</th>
<th>Resources: SH vs. BPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLP</td>
<td>0.92</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>Passive Aggr.</td>
<td>0.92</td>
<td>0.93</td>
<td>0.90</td>
</tr>
<tr>
<td>SGD Classifier</td>
<td>0.89</td>
<td>0.89</td>
<td>0.88</td>
</tr>
<tr>
<td>Ridge Classifier</td>
<td>0.90</td>
<td>0.91</td>
<td>0.90</td>
</tr>
<tr>
<td>KNN Classifier</td>
<td>0.79</td>
<td>0.72</td>
<td>0.82</td>
</tr>
<tr>
<td>Nearest Centroid</td>
<td>0.90</td>
<td>0.89</td>
<td>0.90</td>
</tr>
<tr>
<td>Linear SVC</td>
<td>0.90</td>
<td>0.92</td>
<td>0.90</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.88</td>
<td>0.90</td>
<td>0.86</td>
</tr>
<tr>
<td>Bernoulli NB</td>
<td>0.91</td>
<td>0.92</td>
<td>0.85</td>
</tr>
</tbody>
</table>

On par $F_1$ score with **huge reduction** in space and time complexity.
Empirical Evaluation: ML algorithms

20 News Corpus. $N=512$

<table>
<thead>
<tr>
<th>Classifier</th>
<th>$F_1$ score</th>
<th>Resources: SH vs. HD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SH</td>
<td>HD</td>
</tr>
<tr>
<td>MLP</td>
<td>0.72</td>
<td>0.64</td>
</tr>
<tr>
<td>Passive Aggr.</td>
<td>0.74</td>
<td>0.69</td>
</tr>
<tr>
<td>SGD Classifier</td>
<td>0.70</td>
<td>0.66</td>
</tr>
<tr>
<td>Ridge Classifier</td>
<td>0.16</td>
<td>0.71</td>
</tr>
<tr>
<td>KNN Classifier</td>
<td>0.31</td>
<td>0.31</td>
</tr>
<tr>
<td>Nearest Centroid</td>
<td>0.08</td>
<td>0.15</td>
</tr>
<tr>
<td>Linear SVC</td>
<td>0.75</td>
<td>0.69</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.58</td>
<td>0.26</td>
</tr>
<tr>
<td>Bernoulli NB</td>
<td>0.60</td>
<td>0.15</td>
</tr>
</tbody>
</table>

On par $F_1$ score with **huge reduction** in space and **time** complexity.
Empirical Evaluation: **HyperEmbed** vs. **CountVectorizer** (MLP)

### Chatbot Corpus

#### F1 Score

- **HD vectors**
- **Vectorized n-grams**

#### Speed-up

- **Speed-up baseline**
- **Train speed-up**
- **Test speed-up**

---

### 20 News Corpus

#### F1 Score

- **HD vectors**
- **Vectorized n-grams**

#### Speed-up

- **Speed-up baseline**
- **Train speed-up**
- **Test speed-up**
Thank You!