Brain-Inspired Hypervector Processing at the Edge of Large Language Models

*Alaaddin Goktug Ayar, Sercan Aygun, M. Hassan Najafi, and Martin Margala

University of Louisiana at Lafayette
School of Computing and Informatics
Lafayette, LA, USA

*alaaddin.ayar1@louisiana.edu
- Proposed Method
- Inference Time: ML vs HDC
- Overall Results
- Conclusion
Proposed Method

If you like original gut wrenching laughter you will like this movie...

(+) Positive Comments

So im not a big fan of Boll's work but then again not many are...

(-) Negative Comments

Conventional domain
If you like original gut wrenching laughter you will like this movie...
(+ Positive Comments)

So im not a big fan of Boll's work but then again not many are...
(- Negative Comments)
Brain-Inspired Hypervector Processing at the Edge of Large Language Models

Proposed Method

I. IMDb

If you like original gut wrenching laughter you will like this movie...
(+): Positive Comments
So im not a big fan of Boll's work but then again not many are...
(-): Negative Comments

II. Encoder Output

<table>
<thead>
<tr>
<th>Encoder Output</th>
<th>E1</th>
<th>E2</th>
<th>E3</th>
<th>...</th>
<th>E768</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>w1</td>
<td>w2</td>
<td>w3</td>
<td>...</td>
<td>w768</td>
</tr>
</tbody>
</table>

from floating point --

Conventional domain

Language Models

BERT

Distil BERT

GPT-2

Attention Mechanism

Trainable Weights

Trained Weights

Mapping domain
Brain-Inspired Hypervector Processing at the Edge of Large Language Models

Proposed Method

I. IMDb

If you like original gut wrenching laughter you will like this movie...

(+) Positive Comments

So I'm not a big fan of Boll's work but then again not many are...

(-) Negative Comments

II. Encoder Output

E1 E2 E3 \cdots E768

\times

w1 w2 w3 \cdots w768

from floating point

Encoder Output

HDC domain

III. Binary Conversion

\begin{align*}
\sum & \sum & \sum & \cdots & \sum \\
\sum & \sum & \sum & \cdots & \sum \\
\sum & \sum & \sum & \cdots & \sum
\end{align*}

sign() \approx \begin{cases} 1 & \text{if } \sum > (Sx) / 2 \\ 0 & \text{else} \end{cases}

\begin{bmatrix}
1 \\
1 \\
0 \\
\cdots \\
1
\end{bmatrix}

Class Hypervector

Conventional domain

Mapping domain
For ARM-based edge devices, the HDC model is 6x, 3x, and 5x faster than SVM, MLP, and Random Forest, respectively. This speed is due to its use of binary data, simplifying inference to just adding and thresholding binary vectors.
### Overall Results

<table>
<thead>
<tr>
<th>Embedding Model</th>
<th>Algorithm</th>
<th>Iso-Accuracy</th>
<th>Training Time (sec.)</th>
<th>Model Size</th>
<th>Inference Time (sec.)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BERT</strong></td>
<td>SVM</td>
<td>81%</td>
<td>946.32</td>
<td>213.51 MB</td>
<td>6.3357</td>
</tr>
<tr>
<td></td>
<td>MLP</td>
<td></td>
<td>300.21</td>
<td>18.05 MB</td>
<td>3.3300</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td></td>
<td>534.59</td>
<td>63.50 MB</td>
<td>5.7673</td>
</tr>
<tr>
<td></td>
<td>HDC</td>
<td></td>
<td>272.94</td>
<td>9KB</td>
<td>1.0180</td>
</tr>
<tr>
<td><strong>DistilBERT</strong></td>
<td>SVM</td>
<td>83%</td>
<td>800.45</td>
<td>95.58 MB</td>
<td>4.5522</td>
</tr>
<tr>
<td></td>
<td>MLP</td>
<td></td>
<td>590.75</td>
<td>1.81 MB</td>
<td>2.9476</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td></td>
<td>519.53</td>
<td>55.76 MB</td>
<td>4.8522</td>
</tr>
<tr>
<td></td>
<td>HDC</td>
<td></td>
<td>269.41</td>
<td>9KB</td>
<td>1.0127</td>
</tr>
<tr>
<td><strong>GPT-2</strong></td>
<td>SVM</td>
<td>77%</td>
<td>750.45</td>
<td>229.67 MB</td>
<td>6.5548</td>
</tr>
<tr>
<td></td>
<td>MLP</td>
<td></td>
<td>343.52</td>
<td>5.42 MB</td>
<td>3.2187</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td></td>
<td>458.48</td>
<td>55.44 MB</td>
<td>4.8012</td>
</tr>
<tr>
<td></td>
<td>HDC</td>
<td></td>
<td>306.41</td>
<td>9KB</td>
<td>1.0290</td>
</tr>
</tbody>
</table>

Result table highlights HDC's efficiency with iso-accurate performance versus other classifiers, achieving high accuracy with a notably compact 9KB model. Unlike traditional algorithms like SVM, MLP, and Random Forest that use larger floating-point representations, HDC's binary encoding greatly reduces model size and complexity.
- HDC pairs effectively with LLMs such as BERT, DistilBERT, and GPT-2 for efficient language processing.

- Achieved high accuracy in binary classification on the IMDb dataset with a compact model size of just 9KB.

- Ensures swift inference times on edge devices, highlighting its suitability for real-time applications.

- Showcased the scalability and cost-effectiveness of HDC for NLP on devices with limited computing power.
Thank you for listening!

alaaddin.ayar1@louisiana.edu