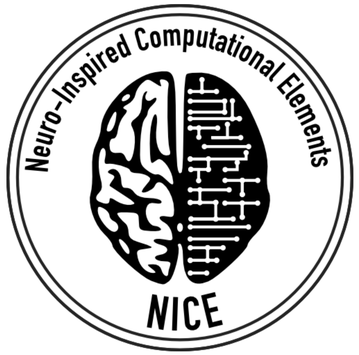


Brain-Inspired Hypervector Processing at the Edge of Large Language Models

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

University of Louisiana at Lafayette

School of Computing and Informatics

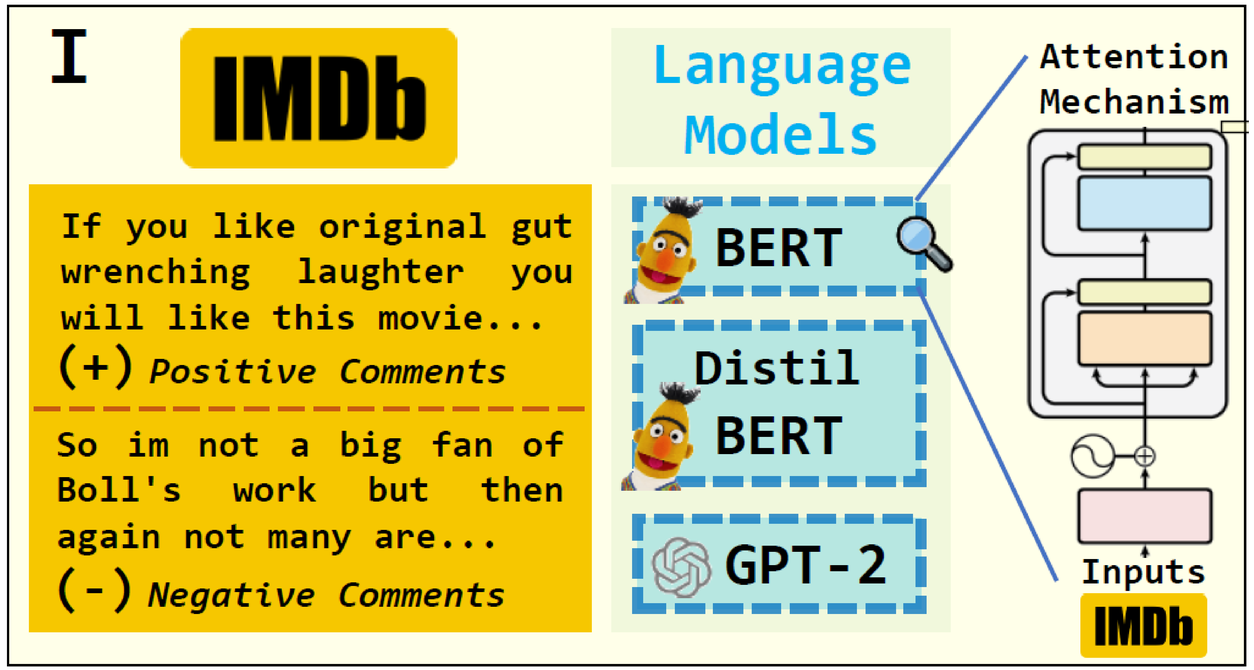
Lafayette, LA, USA

*alaaddin.ayar1@louisiana.edu

Outline

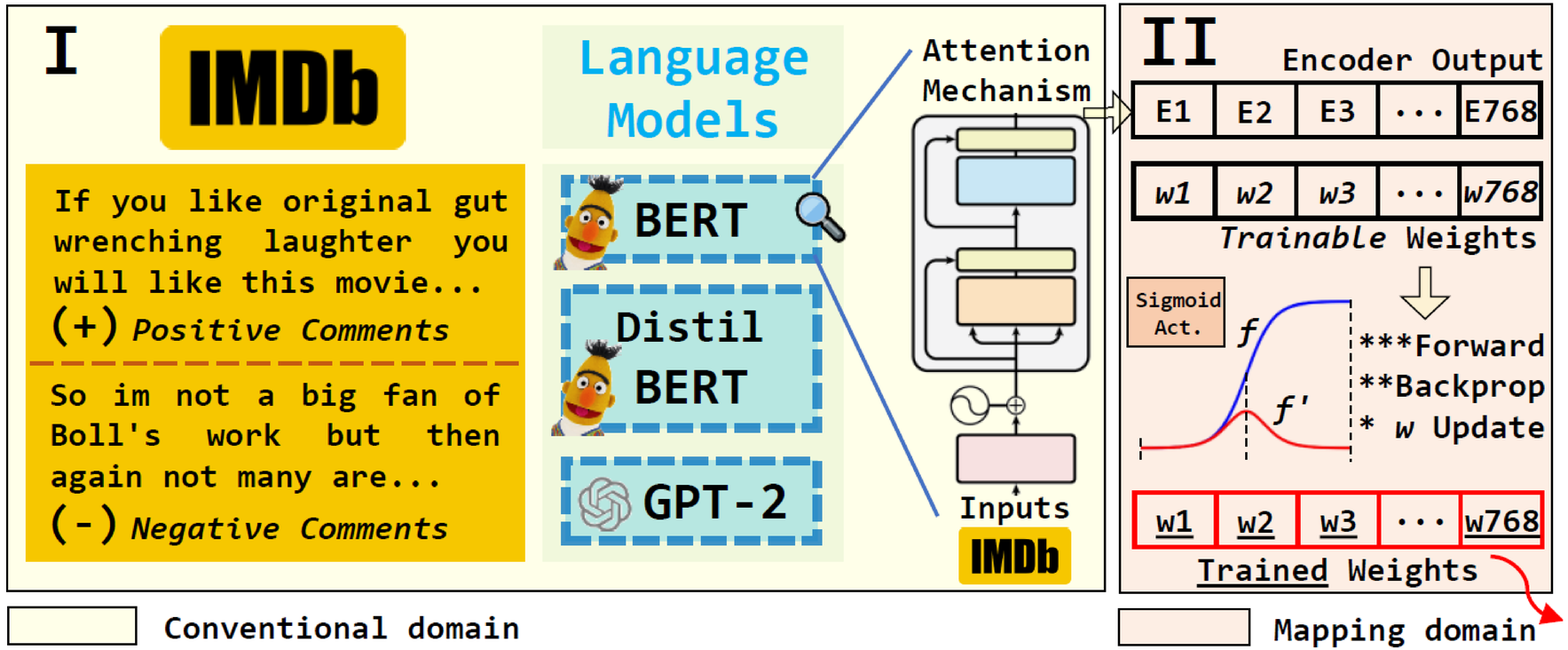
- Proposed Method
- Inference Time : ML vs HDC
- Overall Results
- Conclusion

Proposed Method

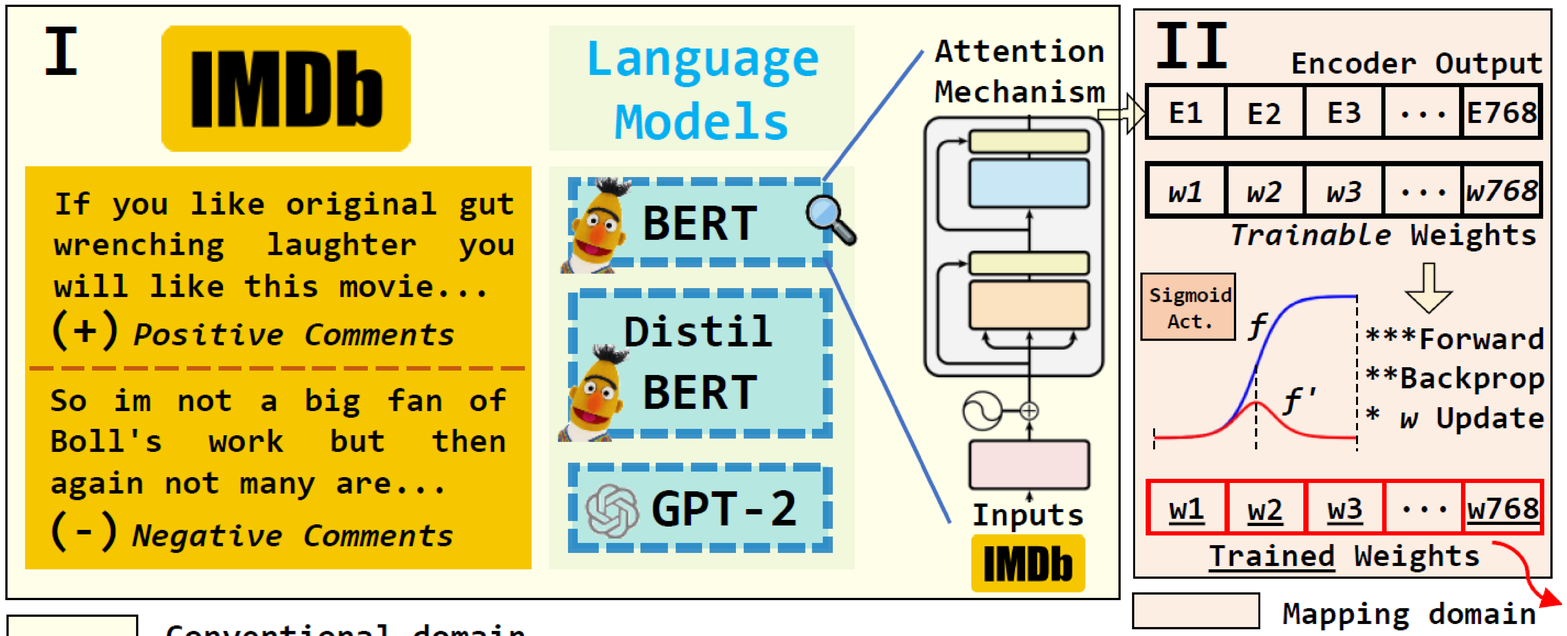


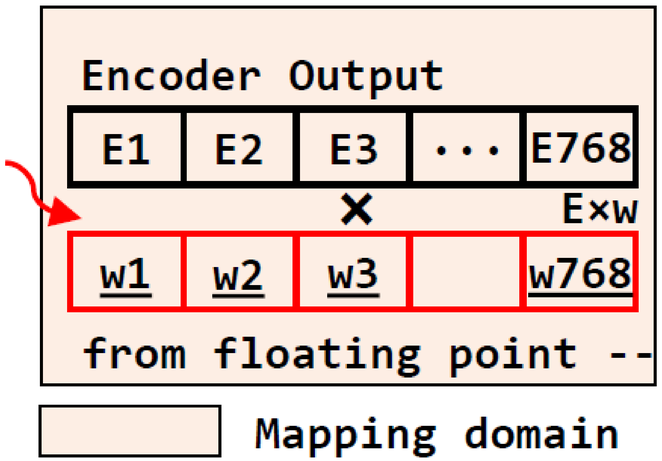
Conventional domain

Proposed Method

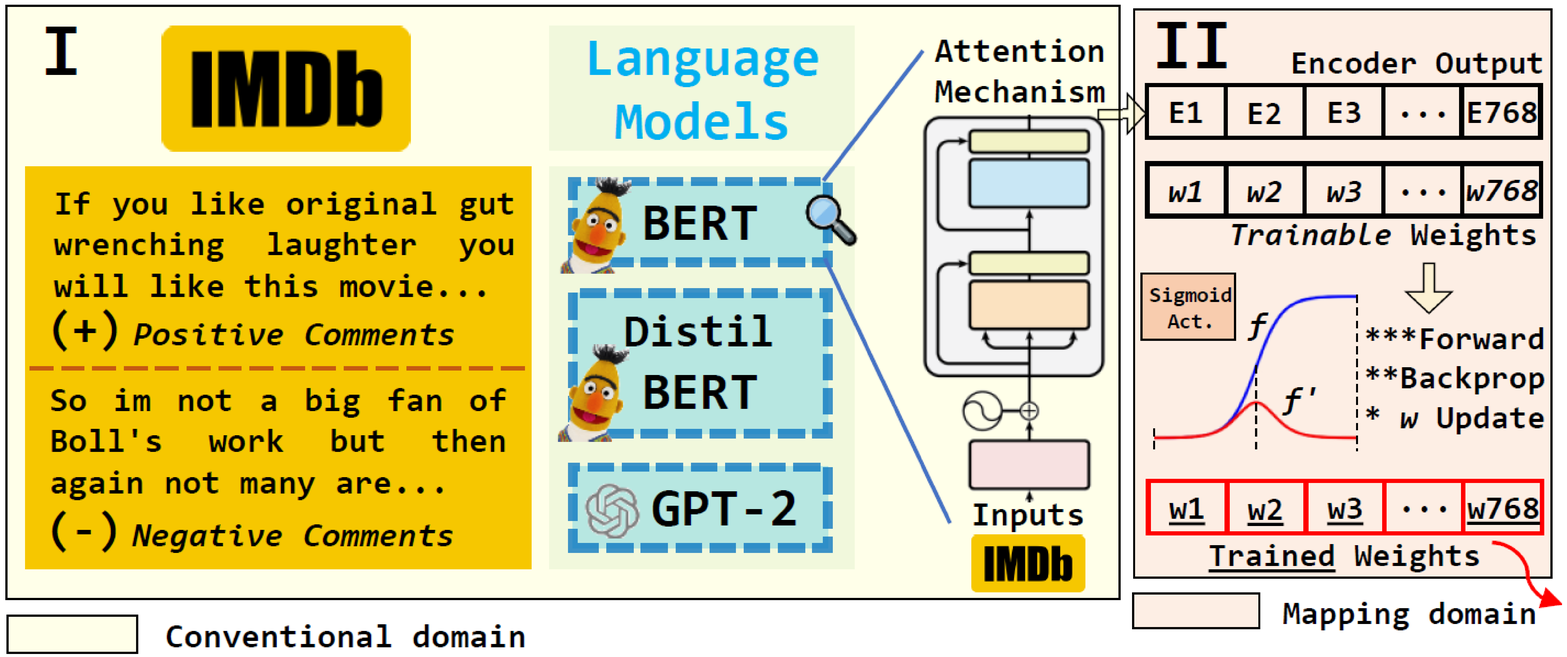


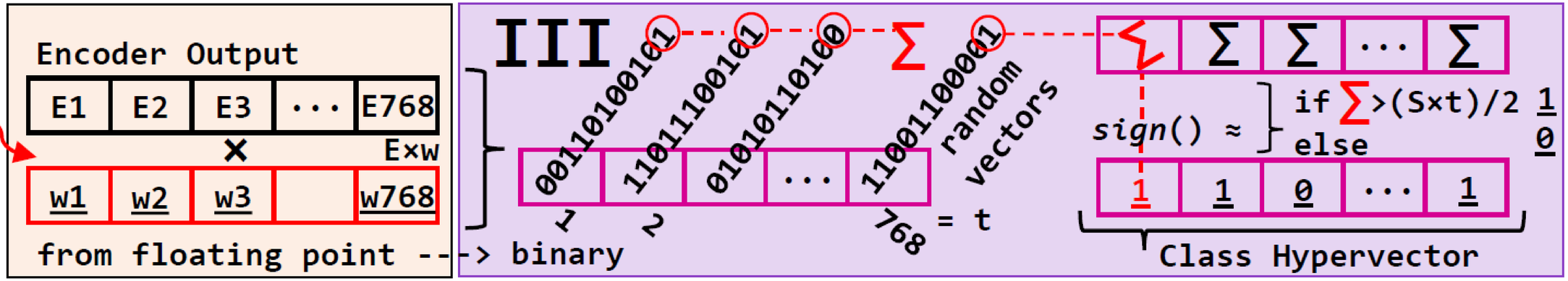
Proposed Method



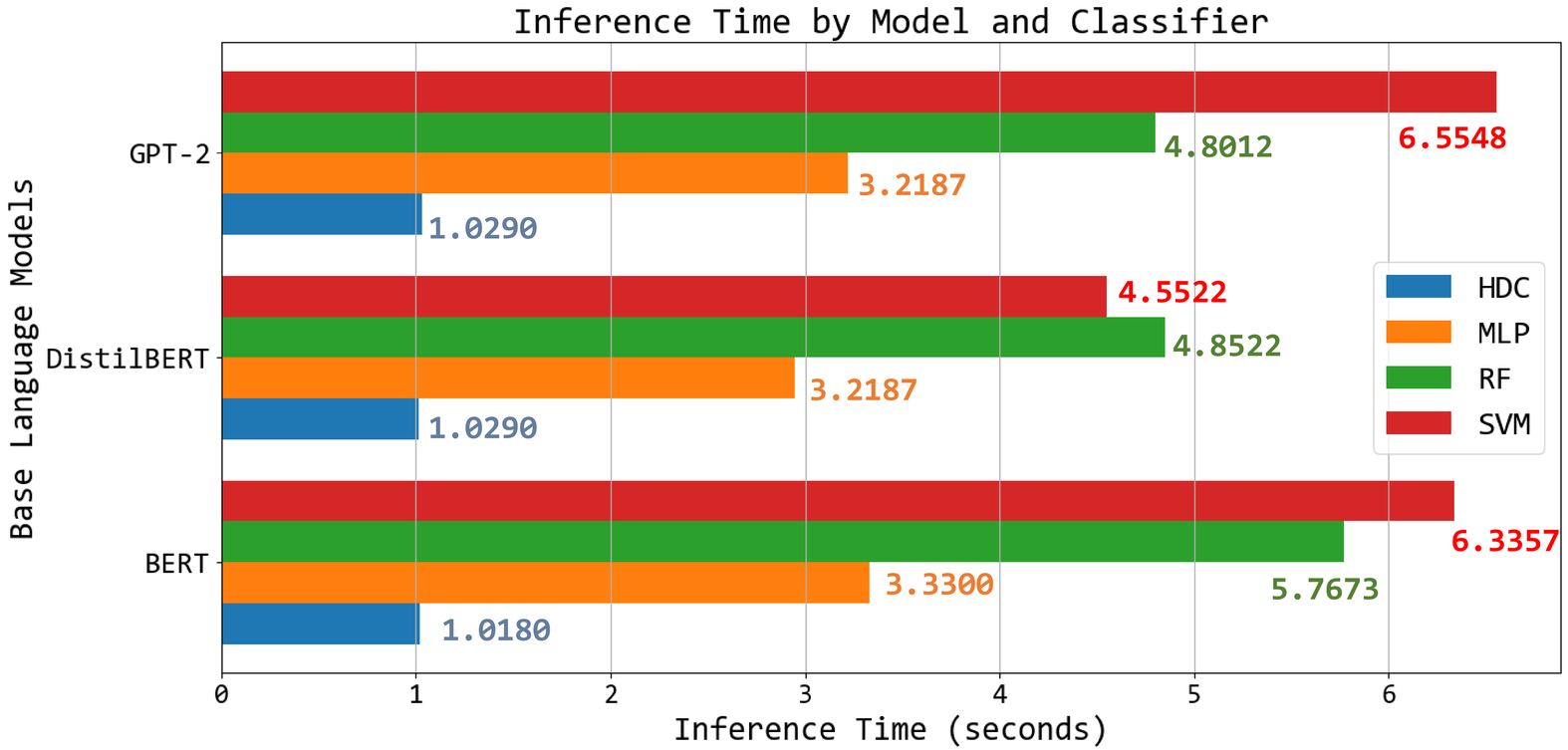


Proposed Method





Inference Time : ML vs HDC



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

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

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.

Conclusion

- 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