WPI

Hyper-HaBERTor: a light-weight pretrained hatespeech language detector using hypercomplex space

Thanh Tran, Kyumin Lee

Department of Computer Science and Data Science

Worcester Polytechnic Institute

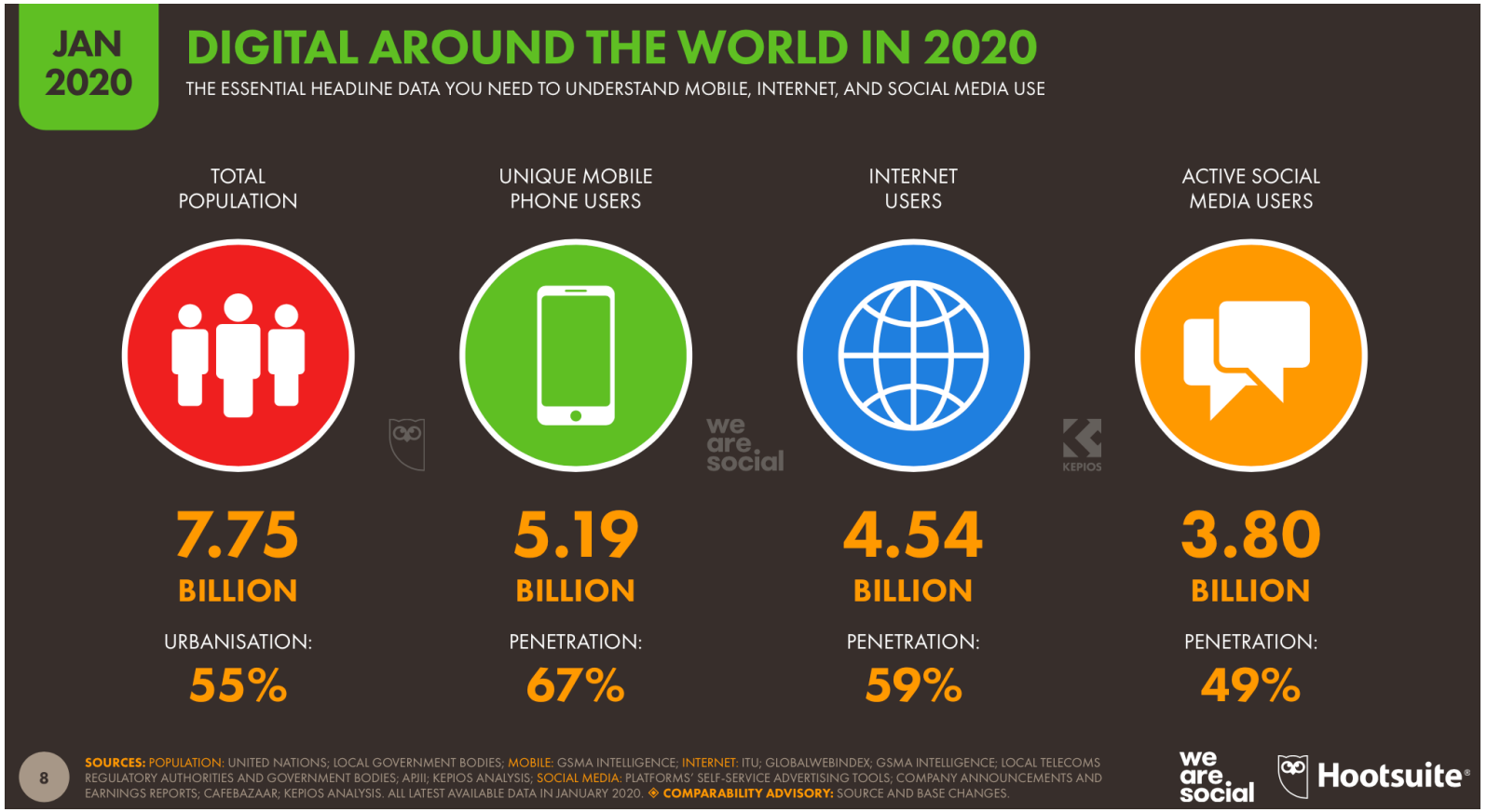
MA, USA

2020

What is hatespeech?

- **Hatespeech**: refer to the speech that **conveys hateful or discriminatory or stereotyped ideas** on the basis of factors such as: **race, gender, religion, sexual orientation**.

Why detecting hatespeech?



Easy to spread hatespeech to a large number of online users.

Why detecting hatespeech?

Hatespeech hurts targeting people:

- **Restrict** their **freedom of speech** on social medias.
- **Interfere** with **civil discourse**.
- **Turn good people away** from saying out loud their opinions.

What's wrong with existing hatespeech detectors?

- Some of the first works design features manually and then feed into traditional machine learning models (i.e. SVM, Logistic Regression):
  - → **Limited by the quality and quantity of the human-crafted features.**
- Recent works used deep neural network:
  - CNN based models: [1], [2].
  - RNN based models: [3]
  - → **Those models lack the ability of Language Understanding and the results are sensitive with weight initialization methods.**

[1] Gamback et al., Using convolutional neural networks to classify hatespeech, ACL 2017 workshop.
[2] Park et al., One-step and twostep classification for abusive language detection on twitter, ACL 2017.
[3] Badjatiya et al., Deep learning for hate speech detection in tweets, WWW 2017.

What's wrong with existing hatespeech detectors?

- Recent pretrained language model (BERT-base [1]) has shown its superior in NLP tasks by pretraining a language model on a vast amount of texts.
  - While hatespeech texts have **unique properties compared to normal texts** (i.e. using a lot of asterisks like: **f*ck**, …; or letters are often replaced by numbers: i → 1, g → 9, …)
  - → **Language model like BERT-base model is still limited in two points:**
    - **Lack of hatespeech language understanding** because they are pretrain on non-hatespeech (formal) corpus like Wiki, Book corpus.
    - **Has a lot number of parameters**, which cause a lot of (GPU) memory to train.
- Recent efforts aim to reduce BERT-base complexity by knowledge distillation method,
  - Still lack of **hatespeech language understanding**, and performed **worse** than BERT-base.

[1] Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, EMNLP 2019.

Motivation

Can we build a pretrained Language Model that

- Has a **smaller number of parameters ?**
  - **Solution: Using Quaternion representations and Quaternion feed forward transformation.**

- With a **better or equivalent** performance on hatespeech detection task?
  - **Solution**: Pretrain from scratch a language model on hatespeech related corpus to equip the model with some hatespeech language understanding.

What is Quaternion representations?

- Quaternion number:

$$X = r + a\,\boldsymbol{i} + b\,\boldsymbol{j} + c\,\boldsymbol{k}$$

$$(ijk = i^2 = j^2 = k^2 = -1; ij = k; jk = i; ki = j; ji = -k; kj = -i; ik = -j)$$

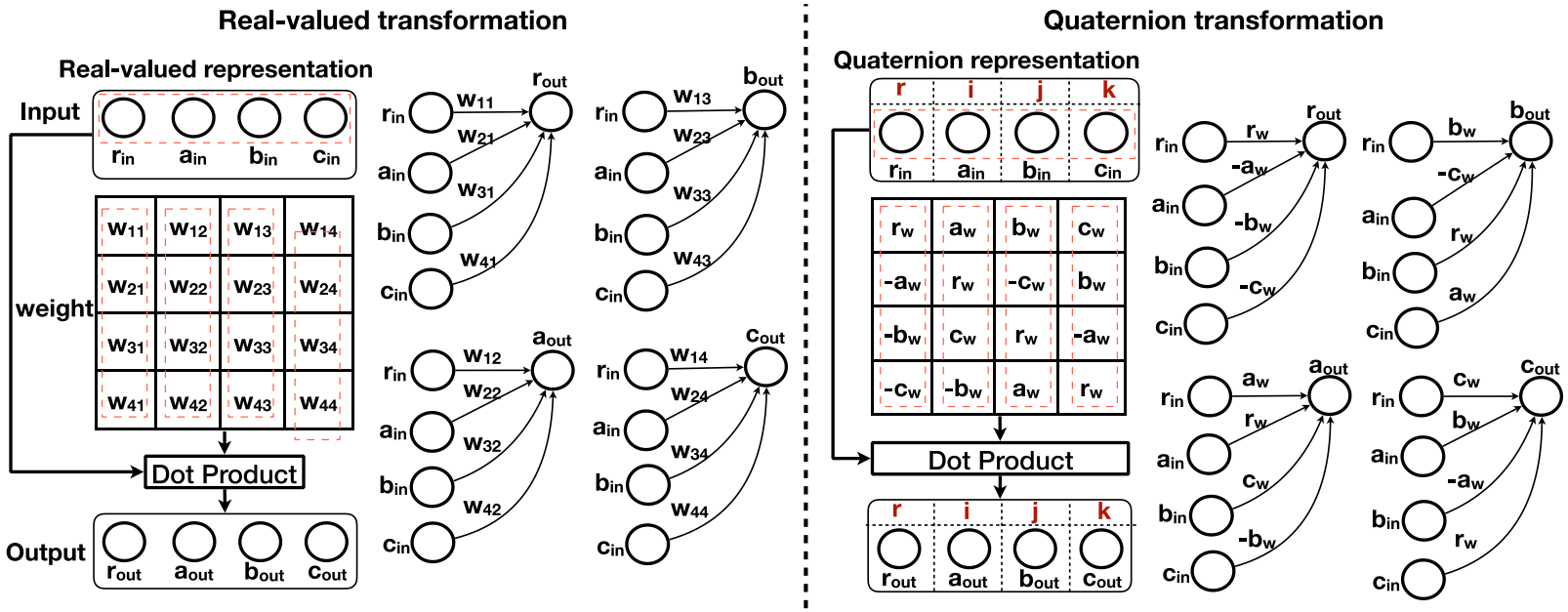
For example: X = 5 + 2**i** + 3**j** + 4**k**

- Hamilton product between two Quaternions X and Y ($X \otimes Y$):

$$\begin{aligned} X \otimes Y = \; & r_X r_Y - a_X a_Y - b_X b_Y - c_X c_Y + \\ & (r_X a_Y + a_X r_Y + b_X c_Y - c_X b_Y)\, i + \\ & (r_X b_Y - a_X c_Y + b_X r_Y + c_X a_Y) j + \\ & (r_X c_Y + a_X b_Y - b_X a_Y + c_X r_Y) k \end{aligned}$$

$$X \otimes Y = [r_X, a_X, b_X, c_X] \begin{bmatrix} r_Y & a_Y & b_Y & c_Y \\ -a_Y & r_Y & -c_Y & b_Y \\ -b_Y & c_Y & r_Y & -a_Y \\ -c_Y & -b_Y & a_Y & r_Y \end{bmatrix}$$

Why Quaternion representation?

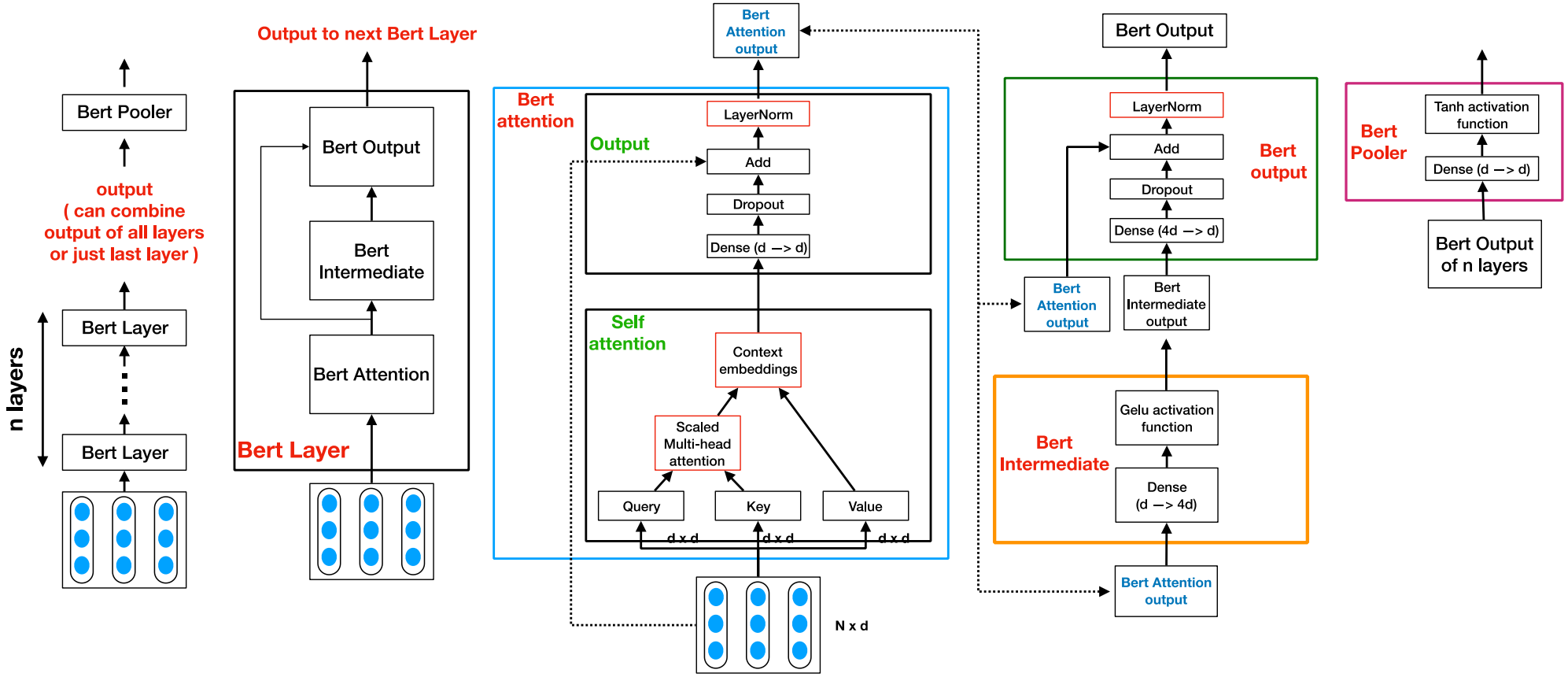


- **Benefits:**
  - Provide a **better inter-dependencies interaction coding** due to the weight sharing in Hamilton product.
  - **Reduce 75% of the number of parameters** over real-valued representations in Euclidean space.
  - Recent works have shown its great performance in NLP and computer vision [1,2,3].

[1] Tay et al., "*Lightweight and efficient neural natural language processing with quaternion networks*", ACL 2019.
[2] Parcollet et al., "Quaternion Recurrent Neural Network ", ICLR 2019.
[3] Yu et al., "*Quaternion-based sparse representation of color image*", ICME 2013.

Our model: Hyper-HaBERTor = Quaternion + BERT + pretrain on hatespeech corpus.

- **Architecture:**



Comparision between BERT-base and Hyper-HaBERTor

- **How many parameters can we reduce?**

Embedding size	intermediate size	#layers	Vocab size
768	3,072	12	32,000

Component			From	To	#params
Vocab					24,576,000
1 encoder layer	Self-attention	Transform key	768	768	589,824
		Transform query	768	768	589,824
		Transform value	768	768	589,824
		output	768	768	589,824
	Intermediate		768	3,072	2,359,296
	Output		3,072	768	2,359,296
1 encoder layer					**7,077,888**
12 encoder layers					84,934,656
Pooling			768	768	589,824
Total					110,100,480

BERT-base

Component			From	To	#params
Vocab					24,576,000
1 encoder layer	Self-attention	Transform key	768	768	147,456
		Transform query	768	768	147,456
		Transform value	768	768	147,456
		output	768	768	147,456
	Intermediate		768	3,072	589,824
	Output		3,072	768	589,824
1 encoder layer					**1,769,472**
12 encoder layers					21,233,664
Pooling			768	768	147,456
Total					**45,957,120**

Hyper-HaBERTor

Experiments

- **Configuration**:
- **Modification of HyperBert:**
  - Using SentencePiece to learn subword vocabs.
  - Real-valued transformation → Quaternion transformation.
  - Real-valued embeddings → Quaternion embeddings.
  - Multi-head scaled dot attention: concatenate 4 components of Quaternion and do the same logic.
  - 10 training examples for masked token prediction/instance as similar to RoBERTa.
  - Adding **targeted adversarial learning** for fine-tuning pha
- **Comparison**:
  - Baseline: BERT-base (110M), DistillBert (66M), TinyBert-4layers (14.5M), RoBERTa-base (125M, 50k vocab size).
  - Downstream Task: hatespeech detection.
  - Data: Yahoo, Wiki, Twitter.

	BERT-base	Hyper-HaBERTor
#layers	12	6
#embedding size	768	384
#intermediate size	3,072	1,536
#attention heads	12	6
#vocabs	32,000	40,000
#parameters	110M	18M
Pretraining corpus	3300M words (Wiki + Book)	33M words from Yahoo comments.

	Yahoo	Twitter	Wiki
#Total	1.4M	16K	115K
#Hateful (%)	100K (7%)	5K (31%)	13K (12%)

Experiments

- **Results on fine-tuning hate-speech task:**

	Yahoo News			Yahoo Finance			Twitter			Wiki		
	AUC	**AP**	**F1**	**AUC**	**AP**	**F1**	**AUC**	**AP**	**F1**	**AUC**	**AP**	**F1**
TinyBert-4Layers	93.13	71.25	64.69	94.12	60.56	58.01	92.23	83.88	78.33	97.10	87.64	79.70
DistilBert	93.13	71.25	64.69	94.12	60.56	58.01	92.13	80.21	77.89	97.23	88.16	80.21
BERT-base	**93.56**	**71.65**	**65.30**	94.60	**62.34**	**59.72**	**93.21**	**86.67**	**79.68**	**97.75**	**89.23**	**80.73**
RoBERTa	92.63	70.15	63.76	93.83	59.73	57.74	90.63	84.36	76.30	95.71	84.48	76.74
HyperBERT - adv	93.55	**71.94**	**65.54**	95.20	62.17	59.65	91.88	84.23	78.05	97.06	87.24	79.46
HyperBERT + adv	**93.71**	**72.59**	**66.08**	95.11	**62.88**	**59.88**	93.26	**86.81**	**80.21**	97.24	88.01	80.27

HyperBERT get better results with adversarial learning on fine-tuning downstream task.
HyperBERT + adv work best for Yahoo and Twitter, but less than BERT-base in wiki with a small amount.
Reason: BERT-base is pretrained on Wiki --> had advantage.

Conclusion

In this talk:

- Utilizing Quaternion space in building a **hatespeech** language model using **Quaternion** representations for the **hatespeech detection task**.
  - **First work using Quaternion representations** on a pretrained language model.
  - **First work applies adversarial learning** on Quaternion space.

- Hyper-HaBERTor obtained, on average, **slightly better F1 scores** on several hatespeech datasets compared to BERT-base model, while reduce **6 times number** of parameters.

Thank You and Questions?