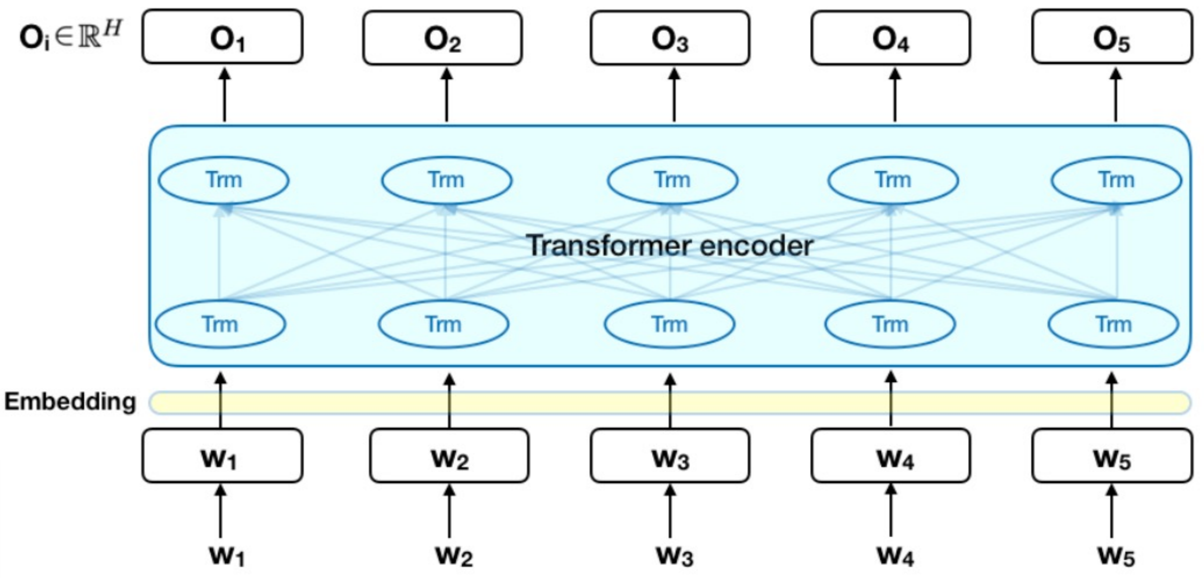
BERTweet: The first Large-scale Pre-trained Language Model for English Tweets

Dat Quoc Nguyen – VinAI Research, Vietnam

VinAI RESEARCH

Motivation

- Language model BERT—Bidirectional Encoder Representations from Transformers (Devlin et al., 2019)—is a recent breakthrough in NLP
  - BERT and its variants, pretrained on large-scale corpora, help improve the state-of-the-art performances of various NLP research & application tasks
  - Represent words by embedding vectors which encode the contexts where the words appear, i.e. contextualized word embeddings



https://www.lyrn.ai/wp-content/uploads/2018/11/transformer.png

Motivation

- Tweet data:
  - Widely-used and real-time source of information in various important analytic tasks (Ghani et al., 2019)
  - Typical short length and frequent use of informal grammar as well as irregular vocabulary e.g. abbreviations, typographical errors and hashtags

Existing language models pre-trained on large-scale conventional text corpora (Wikipedia, news and books) with formal grammar and regular vocabulary

- *To the best of our knowledge, there is not an existing language model pre-trained on a large-scale corpus of English Tweets*

Pre-training BERTweet

- Pre-training corpus:
  - A large-scale corpus of 850M English Tweets (80GB)
  - Use TweetTokenizer to tokenize raw Tweets and the *emoji* package to translate emotion icons into text strings
  - Replace user mentions and URLs by tokens “@USER” and “HTTPURL”, respectively
  - Segment all Tweets with subword units

- BERTweet pre-training procedure is based on RoBERTa (Liu et. al., 2019) which optimizes BERT for more robust performance
  - Remove the next sentence prediction task
  - Use dynamic masking

Pre-training BERTweet

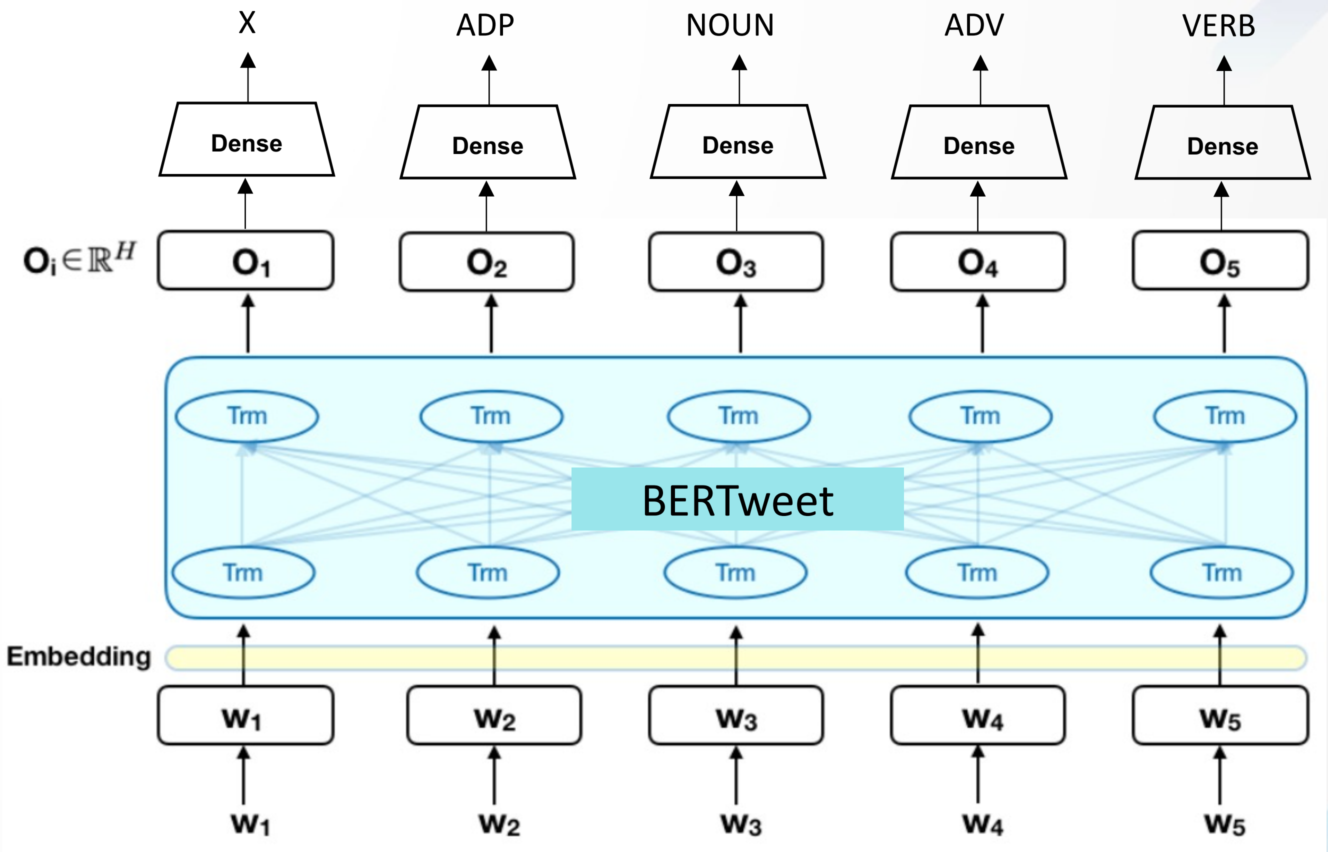
- BERTweet-base (135M parameters)
  - Pre-trained using 8 GPUs V100 32GB memory each
  - Released: 05/2020
- BERTweet-large (355M parameters)
  - Pre-trained using 8 GPUs A100 40GB memory each*
  - Released: 08/2021
- Publicly released under MIT license: https://github.com/VinAIResearch/BERTweet
- BERTweet can be used with popular open-source libraries: ***transformers*** and ***fairseq***

**With FP16 mixed-precision: we find that A100 is 2.5x speedup compared to V100*

Downstream task evaluation

- Part-of-Speech (POS) tagging: To assign a lexical category tag to each word in a text
  - Use a linear prediction layer on top of the BERTweet architecture

ID	Form	POS tag
1	#openfollow	X
2	for	ADP
3	kpopers	NOUN
4	just	ADV
5	retweet	VERB

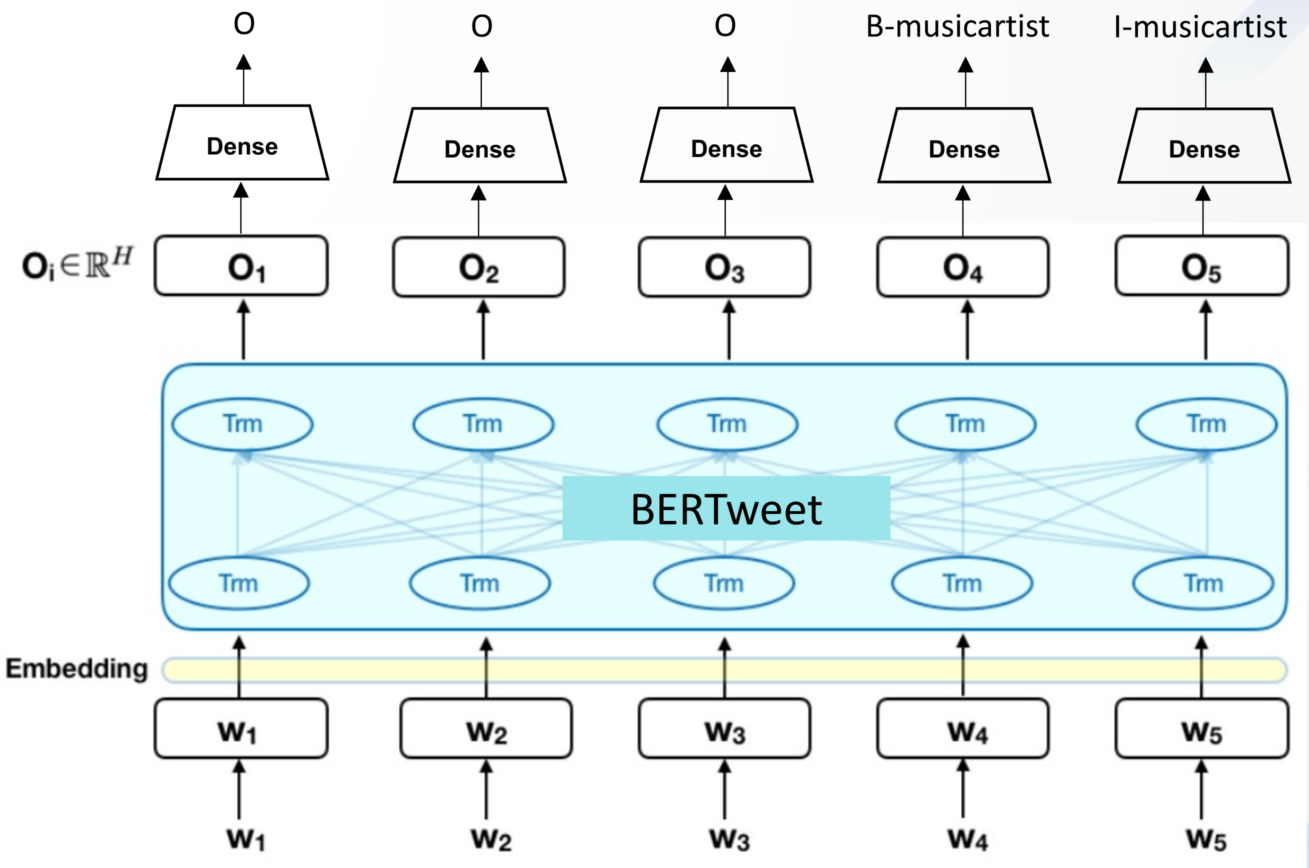


Drawn based on https://www.lyrn.ai/wp-content/uploads/2018/11/transformer.png

Downstream task evaluation

- Named entity recognition (NER): To identify locations, organizations,…
  - Use a linear prediction layer on top of the BERTweet architecture

ID	Form	NER tag
1	oldskool	O
2	night	O
3	wiith	O
4	dj	B-musicartist
5	finese	I-musicartist

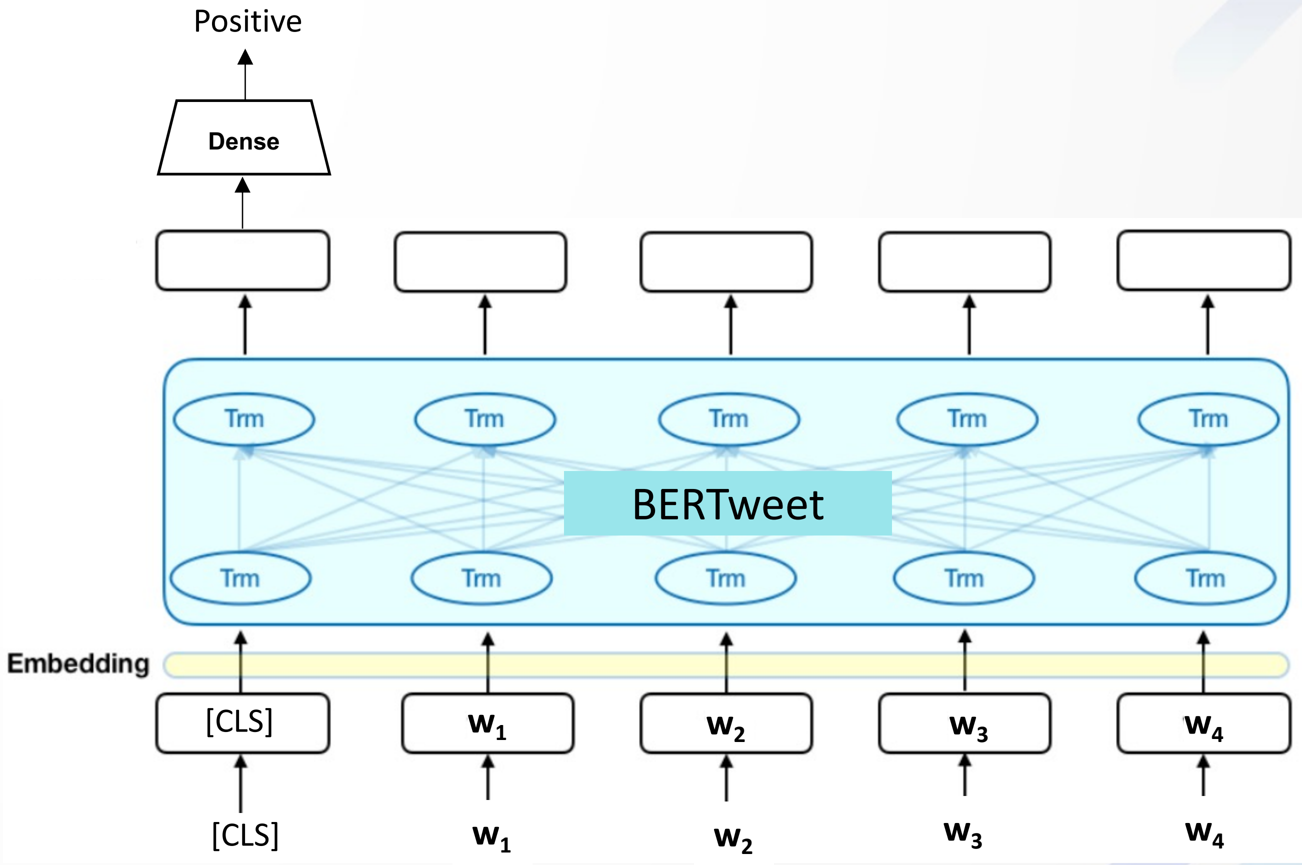


Drawn based on https://www.lyrn.ai/wp-content/uploads/2018/11/transformer.png

Downstream task evaluation

- Text classification: Sentiment analysis (positive, negative or neutral), Irony detection (ironic or not-ironic)
  - Use a linear prediction layer on top of the BERTweet output for the classification token [CLS]—the first token of the input sequence

Tweet	@USER I saw you in Milan, May 9th and it was absolutely incredible
Label	positive



Drawn based on https://www.lyrn.ai/wp-content/uploads/2018/11/transformer.png

Downstream task evaluation

- Benchmark datasets:
  - *POS tagging*: Ritter11-T-POS (Ritter et al., 2011), ARK-Twitter (Gimpelet al., 2011) and Tweebank-v2 (Liu et al., 2018)
  - *NER*: WNUT16 NER shared task (Strauss et al., 2016) and WNUT17 shared task on novel and emerging entity recognition (Derczynski et al., 2017)
  - *Text classification*: SemEval2017 sentiment analysis task 4A (Rosenthal et al., 2017) and SemEval2018 irony detection task 3A (Van Hee et al., 2018)

- “soft” normalization strategy
  - Translate word tokens of user mentions and web/url links into the special tokens “@USER” and “HTTPURL”
  - Convert emotion icon tokens into corresponding strings

Downstream task evaluation

- Pre-trained language model baselines:
  - *RoBERTa*: pre-trained on 160GB of texts covering books, Wikipedia, CommonCrawl news, CommonCrawl stories, and web text contents
  - *XLM-R*: a cross-lingual variant of RoBERTa, pre-trained on a 2.5TB multilingual corpus that contains 300GB of English CommonCrawl texts

Downstream task evaluation

- POS tagging accuracy results on the Ritter11-T-POS (Ritter11), ARK-Twitter (ARK) and Tweebank-v2 (TB-v2) test sets

	Model	Ritter11	ARK	TB-v2
Our results	RoBERTa-large	91.7	93.7	94.9
	XLM-R-large	92.6	94.2	95.5
	BERTweet-large	**92.8**	**95.0**	**95.8**
	RoBERTa-base	88.7	91.8	93.7
	XLM-R-base	**90.4**	92.8	94.7
	BERTweet-base	90.1	**94.1**	**95.2**
	DCNN (Gui et al., 2018) [∗]	91.2	92.4	–
	TPANN (Gui et al., 2017) [∗]	90.9	92.8	–
	ARKtagger [∗]	90.4	93.2	94.6
	BiLSTM-CNN-CRF [∗]	–	–	92.5

[✻] denotes the use of additional training data

Downstream task evaluation

- Entity-level F1 scores on the WNUT16 and WNUT17 test sets

	Model		WNUT16	WNUT17
Our results	RoBERTa-large		55.4	56.9
	XLM-R-large		55.8	57.1
	BERTweet-large		**56.7**	**59.8**
	RoBERTa-base		49.7	52.2
	XLM-R-base		49.9	53.5
	BERTweet-base		**52.1**	**56.5**
CambridgeLTL		[∗]	52.4	–
DATNet (Zhou et al.)		[∗]	53.0	42.3
Aguilar et al. (2017)			–	41.9

[✻] denotes the use of additional training data

Downstream task evaluation

- Text classification test results

	Model	AvgRec	F_1^{NP}	Accuracy
Our results	RoBERTa-large	72.5	72.0	70.7
	XLM-R-large	71.7	71.1	70.7
	BERTweet-large	**73.3**	**73.1**	**72.2**
	RoBERTa-base	71.6	71.2	71.6
	XLM-R-base	70.3	69.4	69.3
	BERTweet-base	**73.2**	**72.8**	**71.7**
	Cliche (2017)	68.1	68.5	65.8
	Baziotis et al. (2017)	68.1	67.7	65.1

SemEval2017 sentiment analysis task 4A

	Model	F_1^{pos}	Accuracy
Our results	RoBERTa-large	73.2	76.5
	XLM-R-large	70.8	74.2
	BERTweet-large	**76.3**	**80.3**
	RoBERTa-base	71.0	74.0
	XLM-R-base	66.6	70.8
	BERTweet-base	**74.6**	**78.2**
	Wu et al. (2018)	70.5	73.5
	Baziotis et al. (2018)	67.2	73.2

SemEval2018 irony detection task 3A

Downstream task evaluation

- Apply a "hard" strategy by further applying lexical normalization dictionaries (Aramaki, 2010; Liu et al., 2012; Han et al., 2012) to normalize word tokens in Tweets
  - Lexical normalization on Tweets is a lossy translation task (Owoputi et al., 2013)

Model	Ritter11		ARK		TB-v2	
	soft	hard	soft	hard	soft	hard
RoBERTa-base	88.7	88.3	91.8	91.6	93.7	93.5
XLM-R-base	**90.4**	**90.3**	92.8	92.6	94.7	94.3
BERTweet-base	90.1	89.5	**94.1**	**93.4**	**95.2**	**94.7**

Model	WNUT16		WNUT17	
	soft	hard	soft	hard
RoBERTa-base	49.7	49.2	52.2	52.0
XLM-R-base	49.9	49.4	53.5	53.0
BERTweet-base	**52.1**	**51.3**	**56.5**	**55.6**

Model	AvgRec		F_1^{NP}		Accuracy	
	soft	hard	soft	hard	soft	hard
RoBERTa-base	71.6	71.8	71.2	71.2	71.6	70.9
XLM-R-base	70.3	70.3	69.4	69.6	69.3	69.7
BERTweet-base	**73.2**	**72.8**	**72.8**	**72.5**	**71.7**	**72.0**

SemEval2017 sentiment analysis task 4A

Model	F_1^{pos}		Accuracy	
	soft	hard	soft	hard
RoBERTa-base	71.0	71.2	74.0	74.0
XLM-R-base	66.6	66.2	70.8	70.8
BERTweet-base	**74.6**	**74.3**	**78.2**	**78.2**

SemEval2018 irony detection task 3A

Key takeaways

- BERTweet is the first public large-scale monolingual language model pre-trained for English Tweets

- BERTweet produces state-of-the-art performances on 3 downstream Tweet NLP tasks: POS tagging, NER, and text classification (i.e. sentiment analysis & irony detection)
  - Outperform its baselines (i.e. RoBERTa and XLM-R) and previous models
  - Effectiveness of a large-scale and domain-specific pre-trained language model for English Tweets

- BERTweet can serve as a strong baseline for future research and applications of Tweet analytic tasks: https://github.com/VinAIResearch/BERTweet

Thanks for your attention!