Analytics Vidhya
Learn everything about analytics

DATAHACK SUMMIT 2019

State of Transfer Learning in NLP

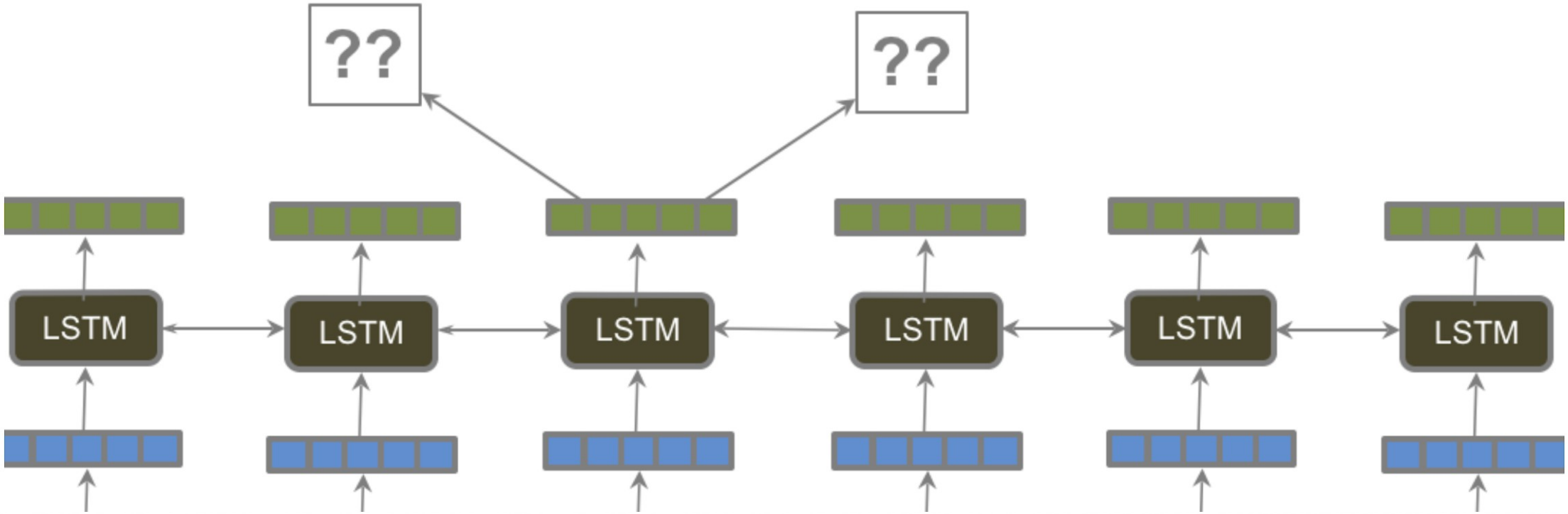
Sudalai Rajkumar

H2O.ai

13 - 16 November 2019 Bengaluru

12 JULY 2018 / NATURAL LANGUAGE PROCESSING

NLP's ImageNet moment has arrived



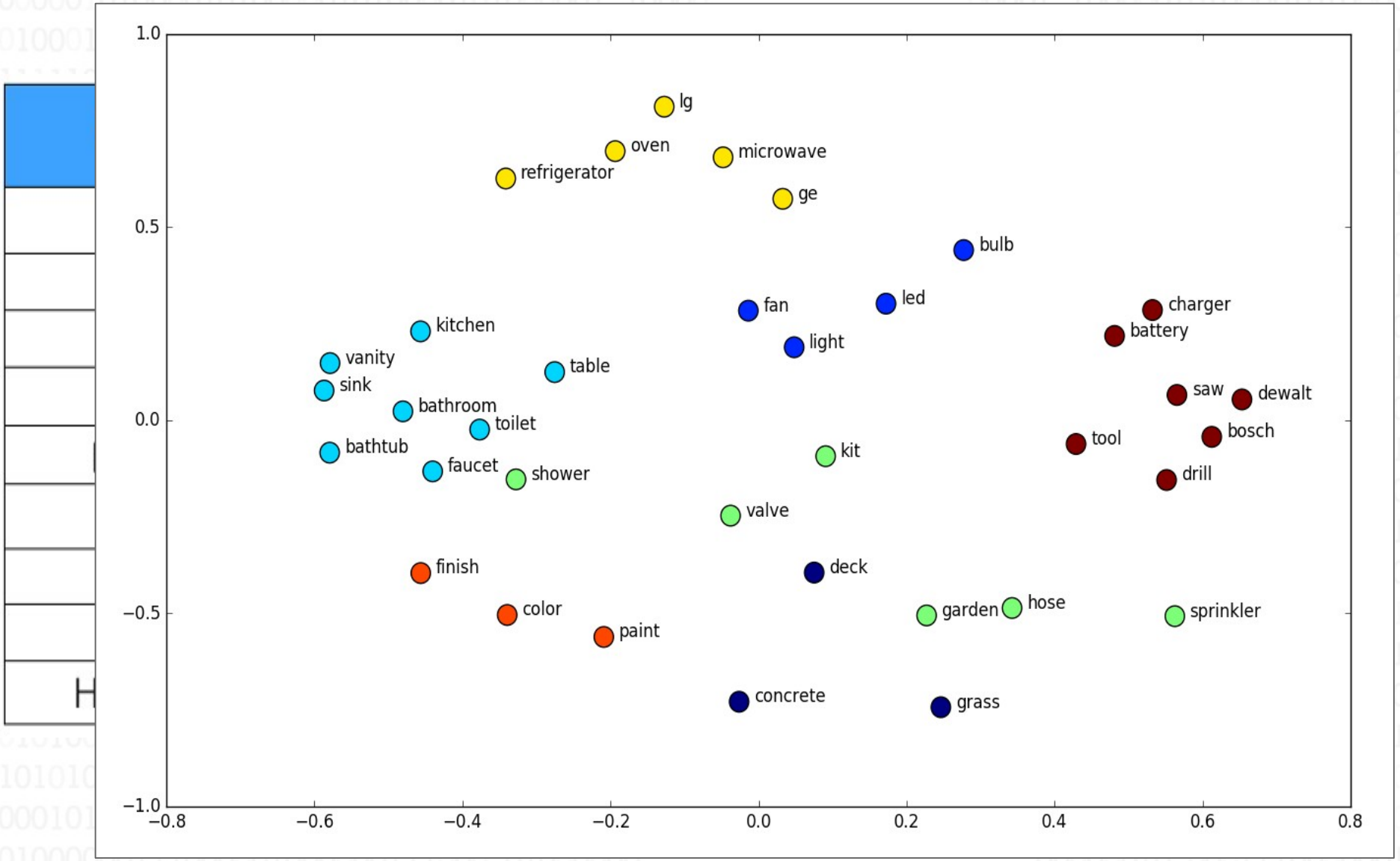
https://ruder.io/nlp-imagenet/

Count / TFIDF

Word	TF		IDF	TF*IDF	
	A	B		A	B
The	1/7	1/7	log(2/2) = 0	0	0
Car	1/7	0	log(2/1) = 0.3	0.043	0
Truck	0	1/7	log(2/1) = 0.3	0	0.043
Is	1/7	1/7	log(2/2) = 0	0	0
Driven	1/7	1/7	log(2/2) = 0	0	0
On	1/7	1/7	log(2/2) = 0	0	0
The	1/7	1/7	log(2/2) = 0	0	0
Road	1/7	0	log(2/1) = 0.3	0.043	0
Highway	0	1/7	log(2/1) = 0.3	0	0.043

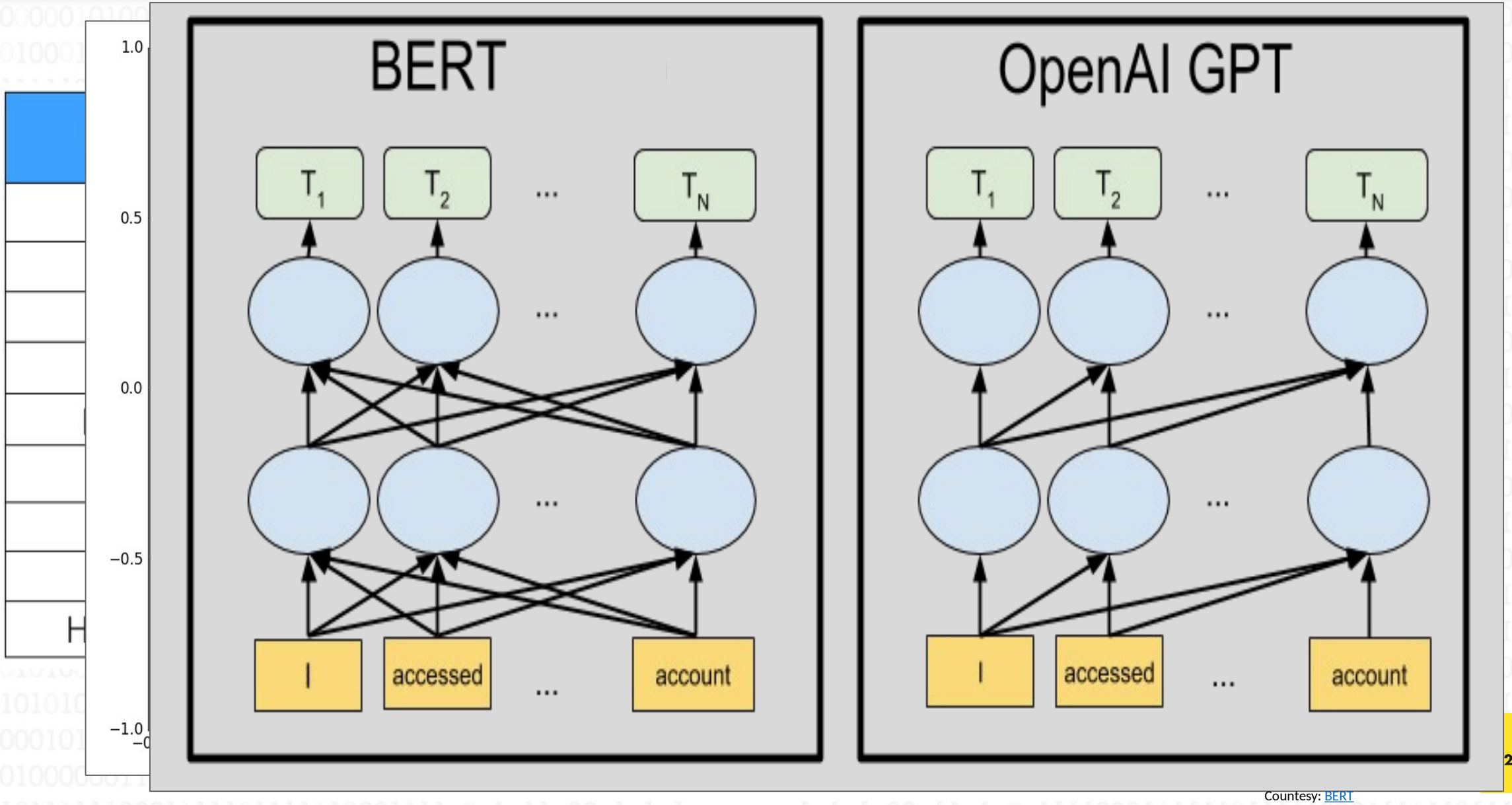
Countesy: Freecodecamp

Word Embeddings



Countesy: Shane Lynn

Pretrained Language Models

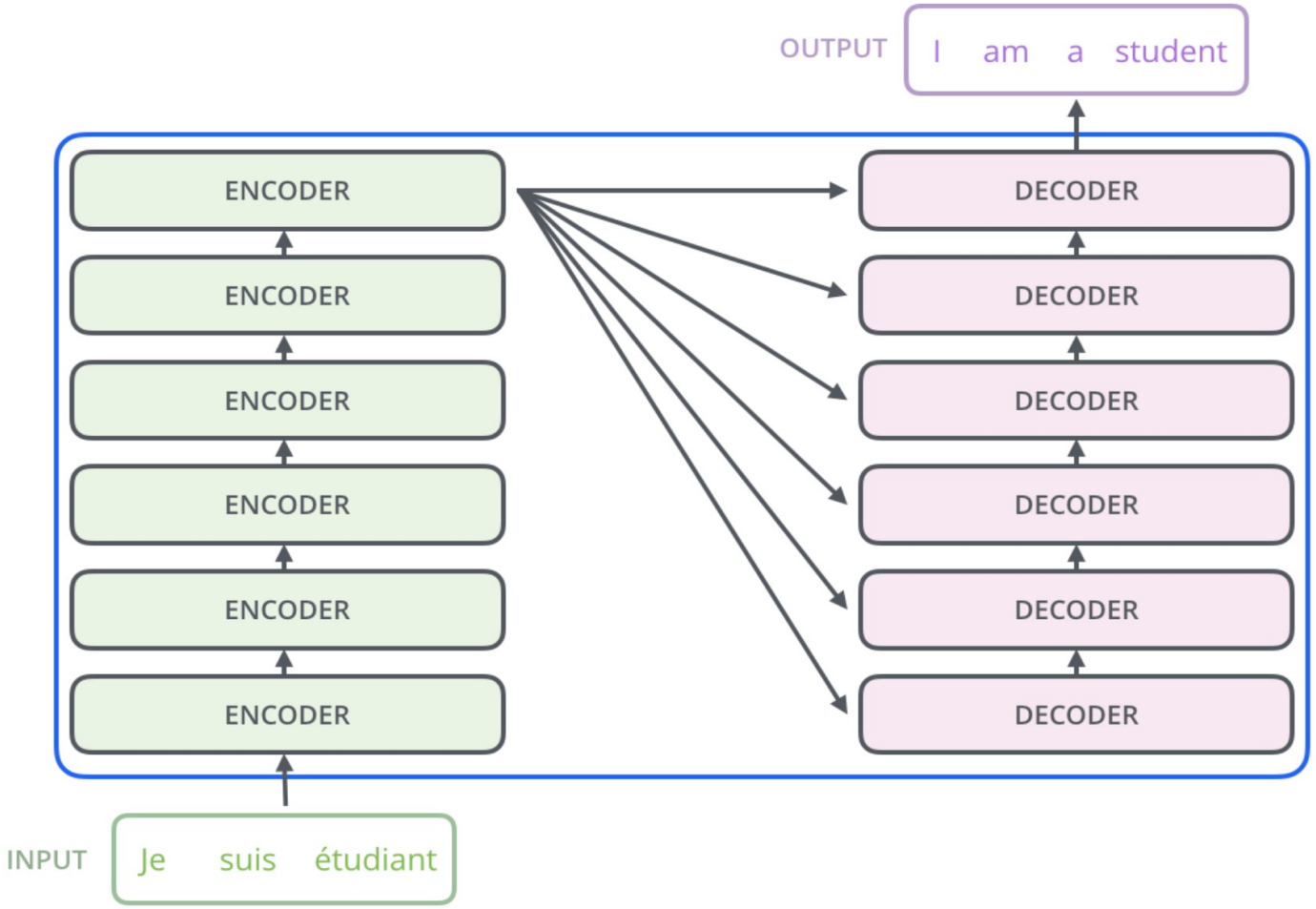


Countesy: BERT

Pretrained Language Models

- BERT
- GPT2
- XLNet
- XLM
- RoBERTa
- DistilBERT
- ULMFit
- ELMo

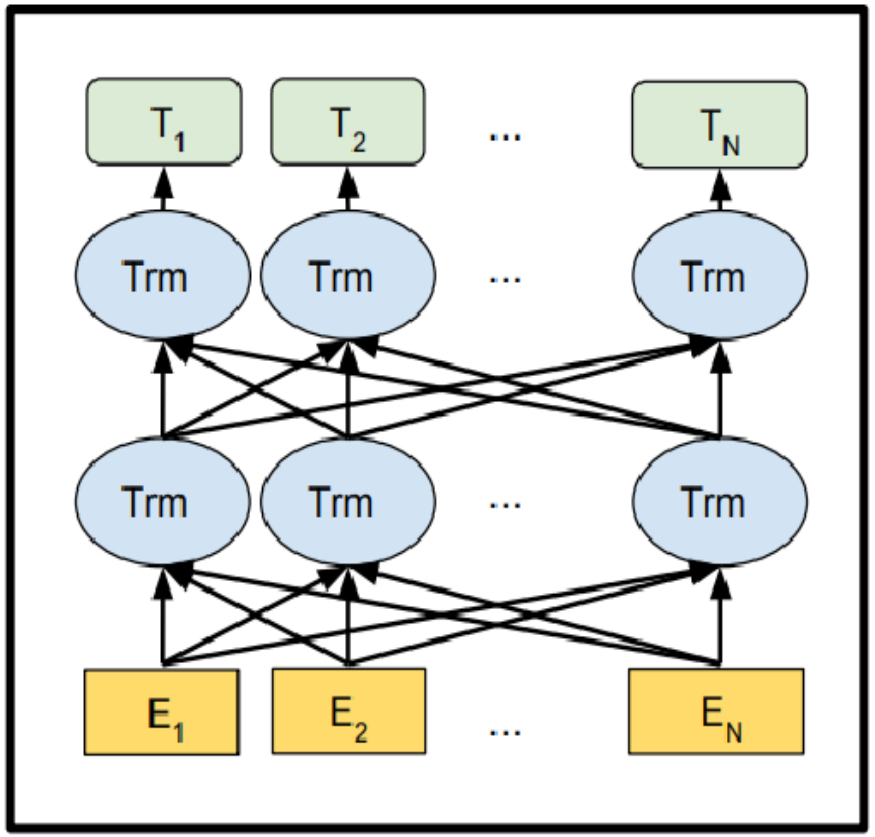
Transformer Architecture



http://jalammar.github.io/illustrated-transformer/

BERT

Bidirectional Encoder Representations for Transformers

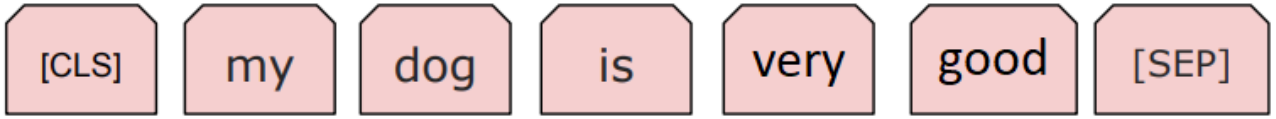


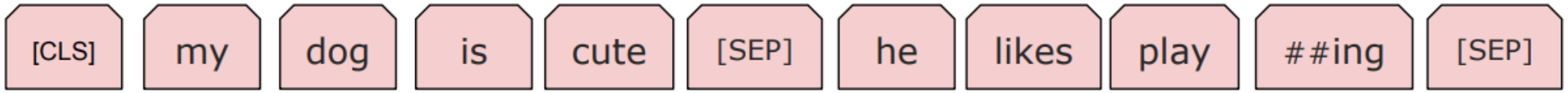
BERT Pre-training tasks

- Masked Language Modeling
  - 15% of the tokens are chosen for masking
    - Replaced with [MASK] - 80% of the times
    - Replaced with random token - 10% of the times
    - Left unchanged - 10% of the times
- Next sentence prediction
  - Given a pair of sentences, the task is to predict whether the second sentence is the actual next sentence of the first sentence
  - Binary classification task

BERT Tokenization

- Tokenization - Wordpiece





- Attention Mask - (optional) a sequence of 1s and 0s, with 1s for all input tokens and 0s for all padding tokens.
- Segment Mask - (optional) a sequence of 1s and 0s used to identify whether the input is one sentence or two sentences long.

Countesy: Yashu Seth

Downstream NLP Tasks

- Sequence classification
  - Sentiment analysis
  - Document classification
- Sentence Pair classification
  - Textual similarity
- Question Answering
- Single Sentence Tagging
  - Named Entity Recognition
- Natural Language Generation

Performance Comparison

	MNLI	QNLI	QQP	RTE	SST	MRPC	CoLA	STS	WNLI	Avg
Single-task single models on dev										
$\text{BERT}_{\text{LARGE}}$	86.6/-	92.3	91.3	70.4	93.2	88.0	60.6	90.0	-	-
$\text{XLNet}_{\text{LARGE}}$	89.8/-	93.9	91.8	83.8	95.6	89.2	63.6	91.8	-	-
RoBERTa	**90.2/90.2**	**94.7**	**92.2**	**86.6**	**96.4**	**90.9**	**68.0**	**92.4**	**91.3**	-

From RoBERTa paper

Performance Comparison

	BERT	RoBERTa	DistilBERT	XLNet
Size (millions)	**Base**: 110 **Large**: 340	**Base**: 110 **Large**: 340	**Base:** 66	**Base**: ~110 **Large**: ~340
Training Time	**Base**: 8 x V100 x 12 days* **Large:** 64 TPU Chips x 4 days (or 280 x V100 x 1 days*)	**Large:** 1024 x V100 x 1 day; 4-5 times more than BERT.	**Base**: 8 x V100 x 3.5 days; 4 times less than BERT.	**Large:** 512 TPU Chips x 2.5 days; 5 times more than BERT.
Performance	Outperforms state-of-the-art in Oct 2018	2-20% improvement over BERT	3% degradation from BERT	2-15% improvement over BERT
Data	16 GB BERT data (Books Corpus + Wikipedia). 3.3 Billion words.	160 GB (16 GB BERT data + 144 GB additional)	16 GB BERT data. 3.3 Billion words.	**Base**: 16 GB BERT data **Large**: 113 GB (16 GB BERT data + 97 GB additional). 33 Billion words.

Countesy: Suleiman Khan

Hack Session

Transformers

Model Performance on Fashion Reviews

Model Name	AUC	Training Time	Prediction Time
BERT (Base uncased)	0.9644	300	40
BERT (Base cased)	0.9608	300	40
XLNet (Base cased)	0.9632	404	55
RoBERTa	0.9635	324	36
DistilBERT	0.9572	181	24

References

- https://ruder.io/nlp-imagenet/
- http://jalammar.github.io/illustrated-transformer/
- https://yashuseth.blog/2019/06/12/bert-explained-faqs-understand-bert-working/
- https://towardsdatascience.com/bert-roberta-distilbert-xlnet-which-one-to-use-3d5ab82ba5f8

Analytics Vidhya
Learn everything about analytics

DATAHACK SUMMIT 2019

Thank you!