Big Bird, Transformers for Longer Sequences

(NeurIPS 2020)

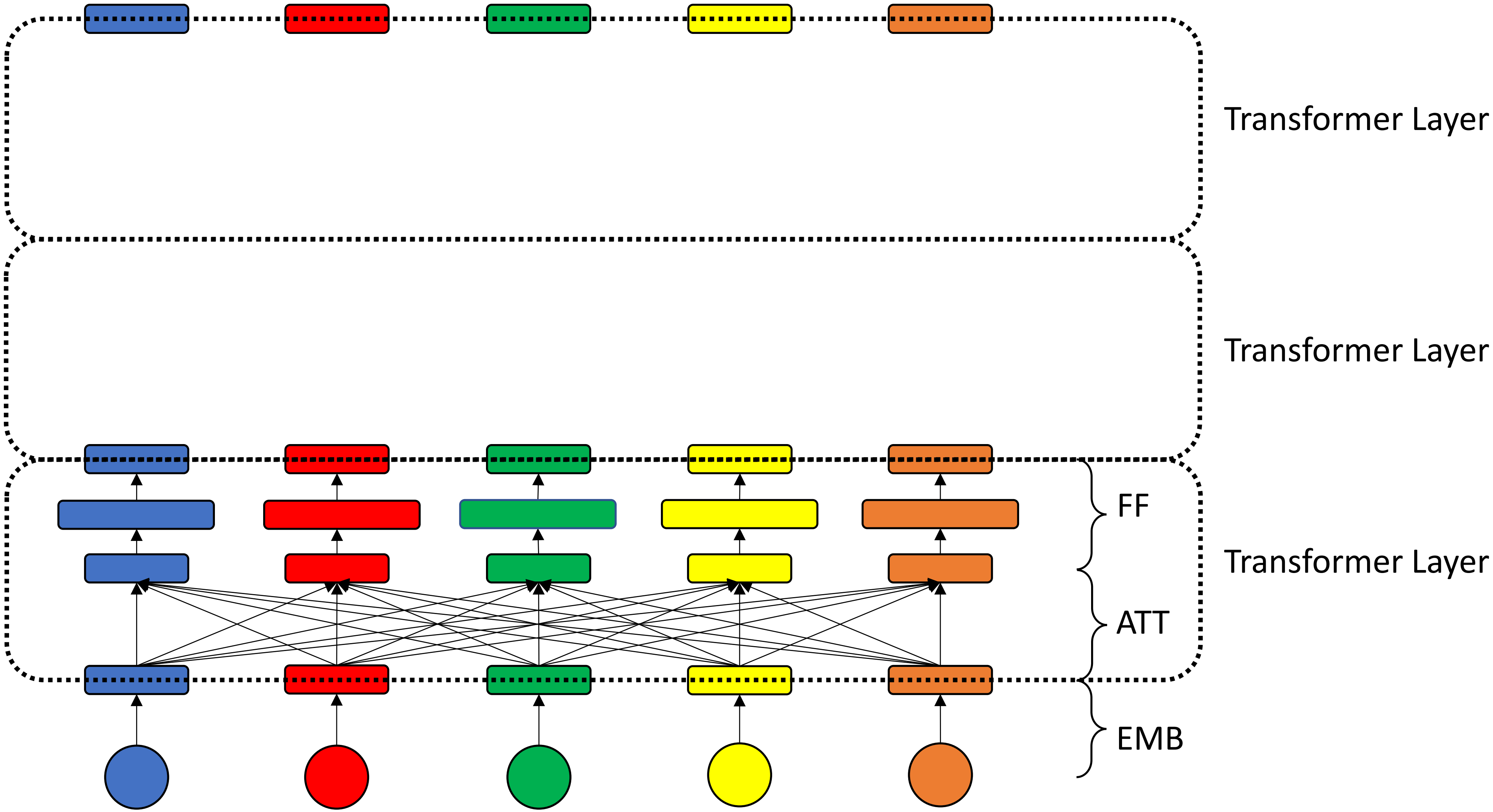
2022/04/22

Peng-Hsuan Li (slides)

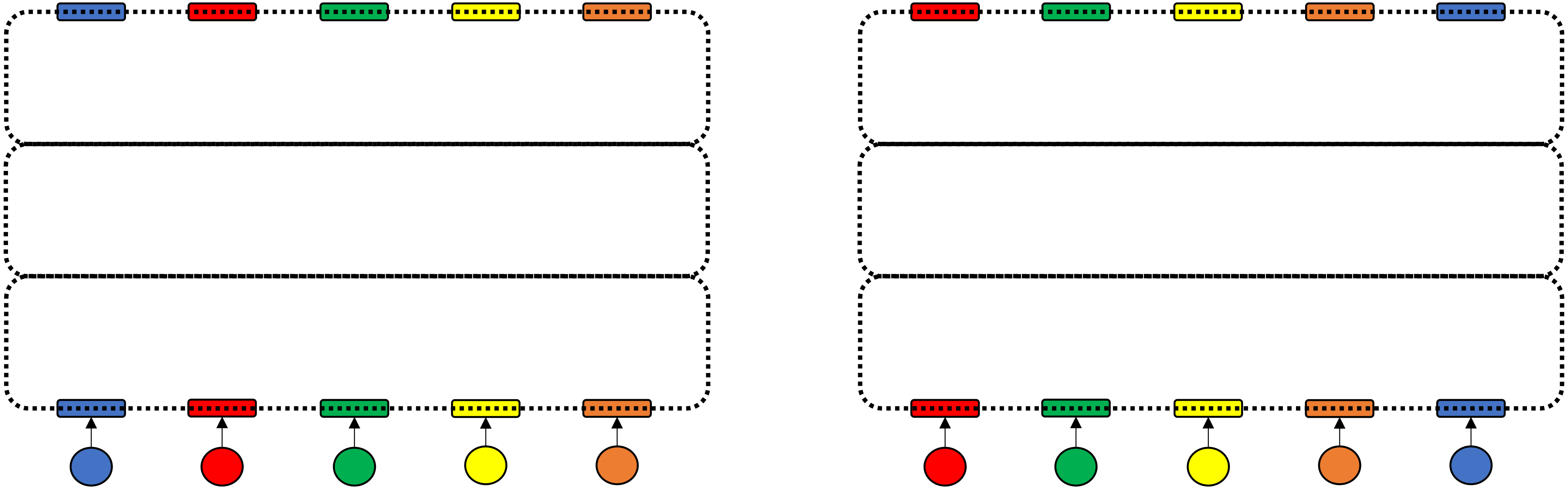
Agenda

- Transformers
  - Universality, Turing Completeness
- Sparse Transformers
  - Universality, Turing Completeness, Graph Theory
- Applications
  - NLP, Genomics

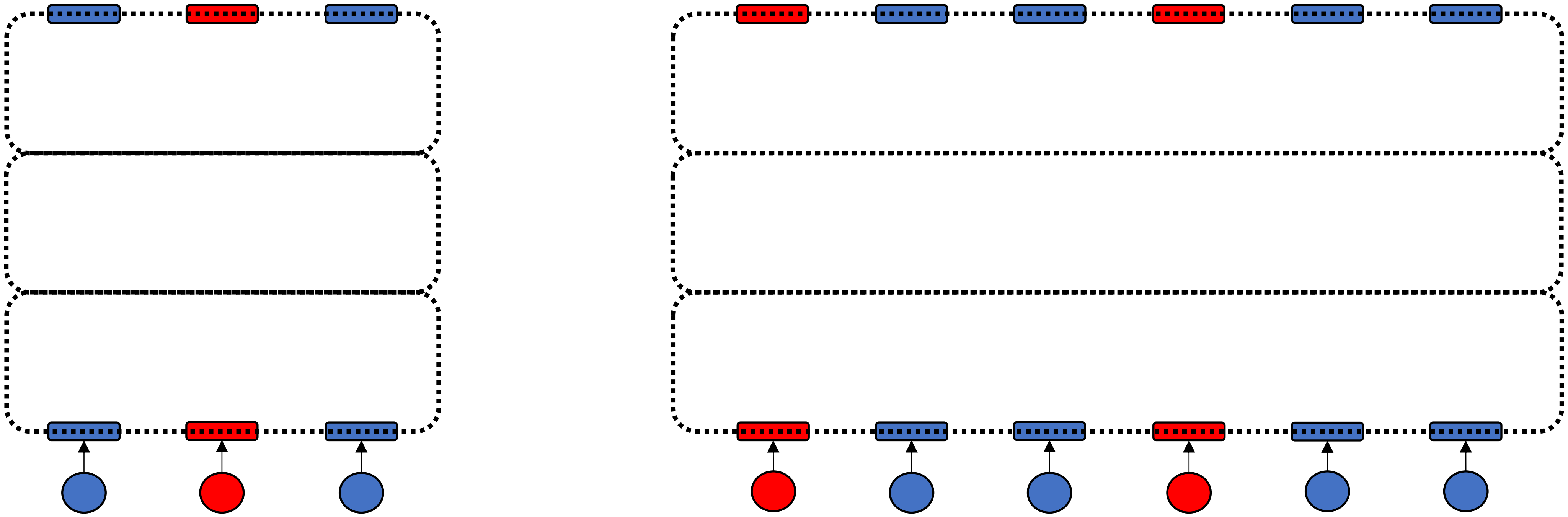
Transformers



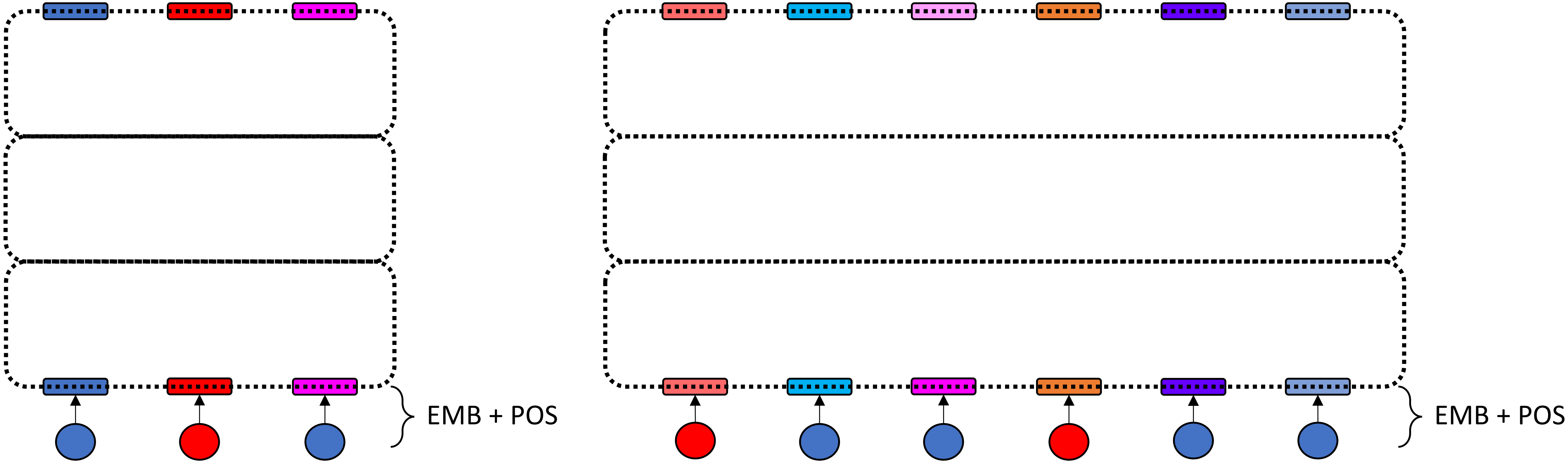
Transformers: Permutation Invariant



Transformers: Proportion Invariant



Transformers: Sequence Modeling



Transformers: Universality

Are Transformers Universal Approximators of Sequence-to-sequence Functions? (ICLR 2020)

Transformers: Universality

- Theorem

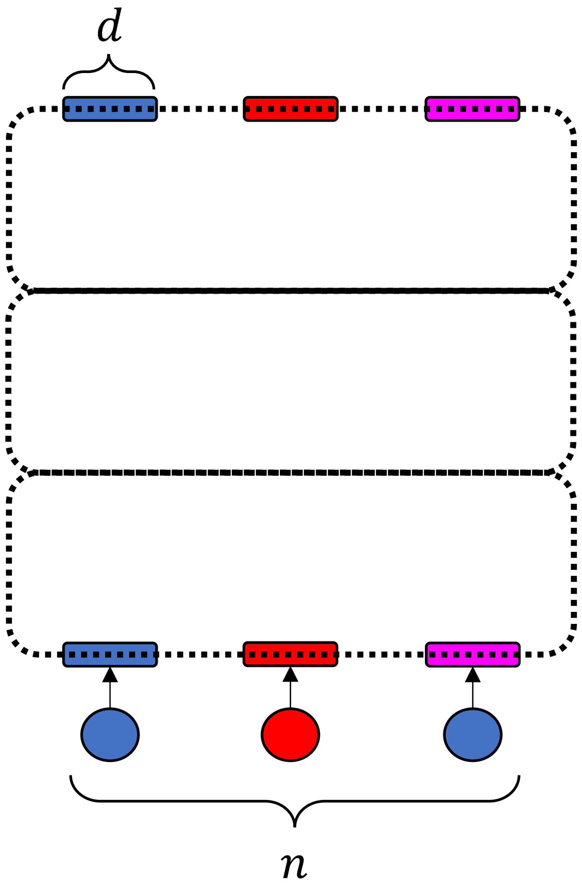
  For every $f: R^{n\times d} \to R^{n\times d}$ with a compact support,

  there exists a transformer t

  s.t. $d(f, t)$ is as small as desired

(Distance between functions)

$$\mathrm{d}_p(f_1, f_2) := \left(\int \|f_1(\boldsymbol{X}) - f_2(\boldsymbol{X})\|_p^p \, d\boldsymbol{X} \right)^{1/p}$$



Transformers: Universality

- Key proposition 1

  $\forall C \subset R^{n \times d}$, C *compact, there exists a transformer* t,

  s.t. $\forall U, V \in C$, $t(U)_i \neq t(V)_j$ *if* $U_i \neq V_j$ *or* $U \neq_p V$

(by constructing specific ATT)

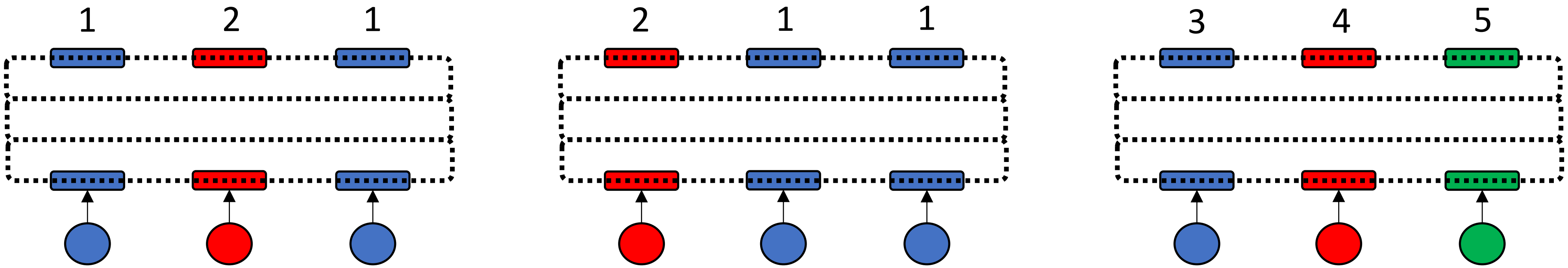
($\neq_p$: not proportionally equivalent)

Transformers: Universality

- Key proposition 1

$\forall C \subset R^{n\times d}$, *$C$ compact, there exists a transformer t,*

s.t. $\forall U, V \in C,\ t(U)_i \neq t(V)_j$ *if* $U_i \neq V_j$ *or* $U \neq_p V$



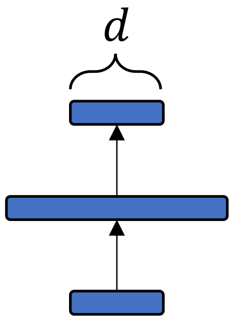
Transformers: Universality

- Key proposition 2

For every $f: R^d \to R^d$ with a compact support,

there exists a feed-forward network t

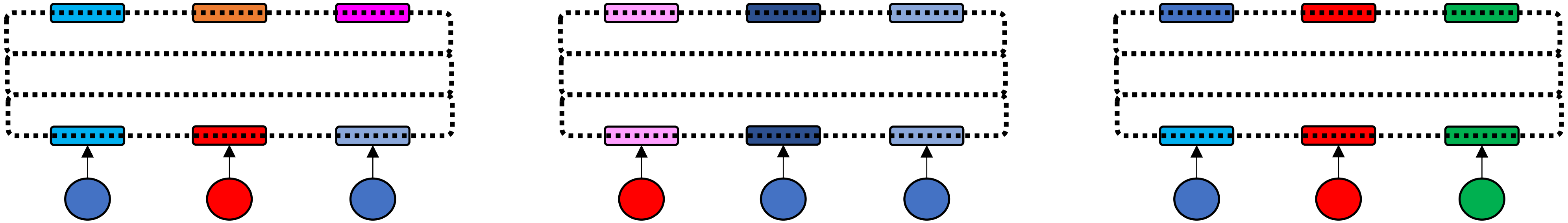
s.t. $d(f, t)$ is as small as desired



Transformers: Universality

- Transformers are universal sequence models

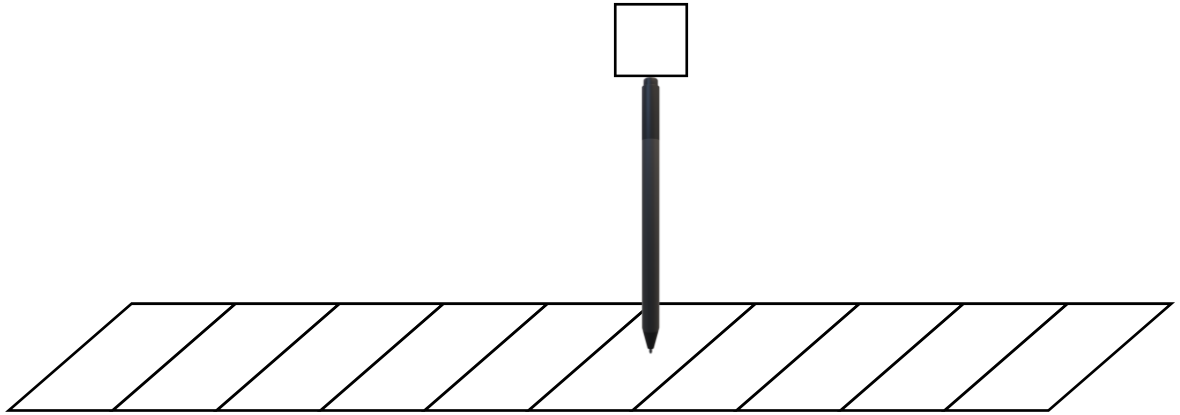
(through the cooperation of POS + ATT + FF)



Transformers: Turing Completeness

- A Turing machine

$$\delta : \underset{\substack{\text{current}\\ \text{state}}}{Q} \times \underset{\substack{\text{current}\\ \text{memory}\\ \text{symbol}}}{\Sigma} \rightarrow \underset{\substack{\text{next}\\ \text{state}}}{Q} \times \underset{\substack{\text{symbol}\\ \text{to}\\ \text{write}}}{\Sigma} \times \underset{\substack{\text{move}\\ \text{left/right}}}{\{L, R\}}$$



Transformers: Turing Completeness

- Church-Turing Thesis

Any effectively calculable function can be realized by some Turing machines

-> A computer is either as powerful as Turing machines or less powerful

-> Those who are as powerful are said to be *Turing-complete*

Transformers: Turing Completeness

On the Turing Completeness of Modern Neural Network Architectures

(ICLR 2019)

- Theorem

  The class of transformers is Turing-complete

Transformers: Turing Completeness

- The key proposition

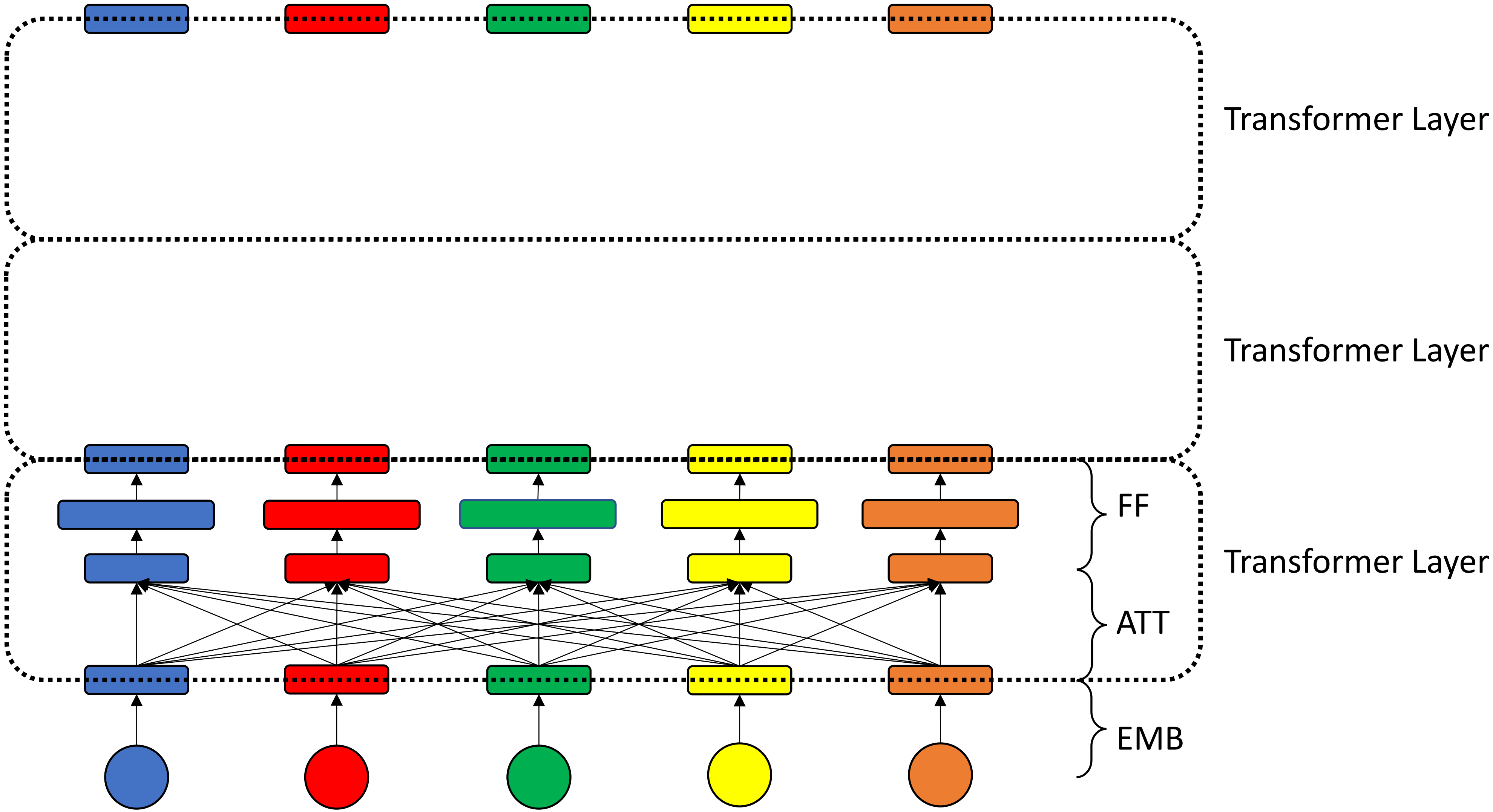
  Every Turing machine can be directly realized (state transition, memory read/write, move left/right) by a sequence-to-sequence transformer with

  a 1-layer encoder,

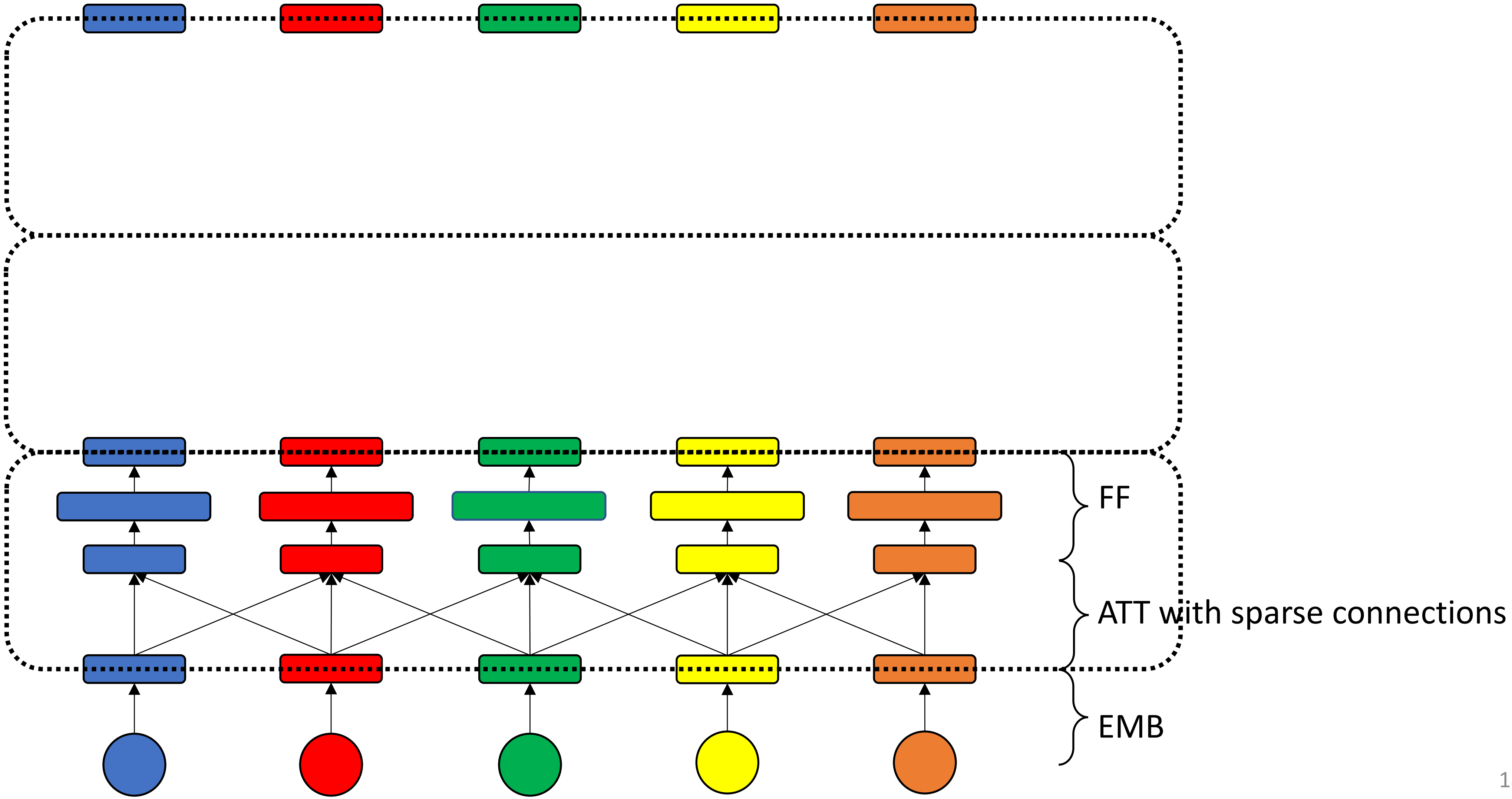
  a 3-layer decoder,

  a vector dimension of $2|Q| + 4|\Sigma| + 11$

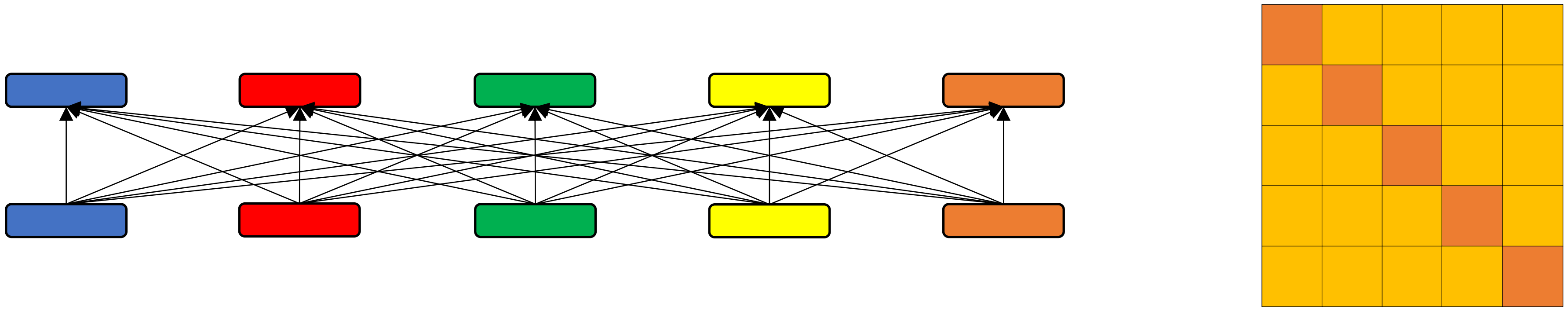
Transformers

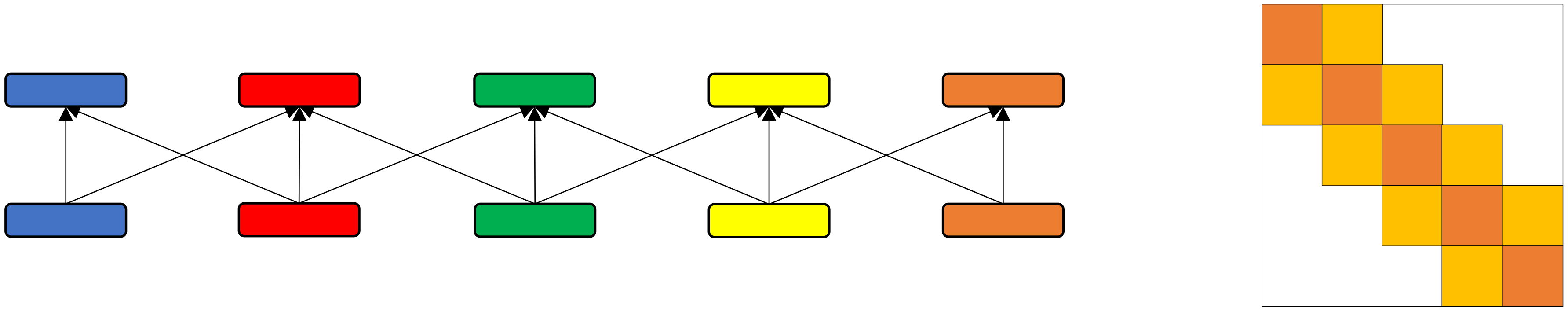


Sparse Transformers

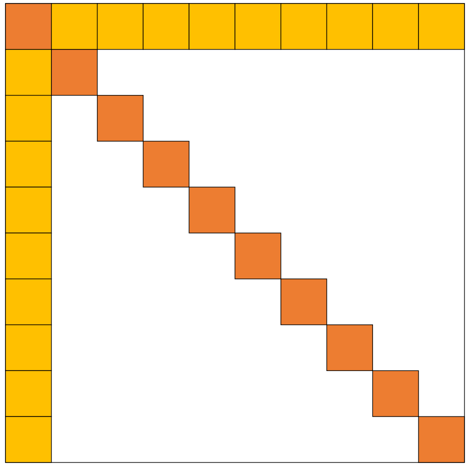


Sparse Transformers: Adjacency Matrices

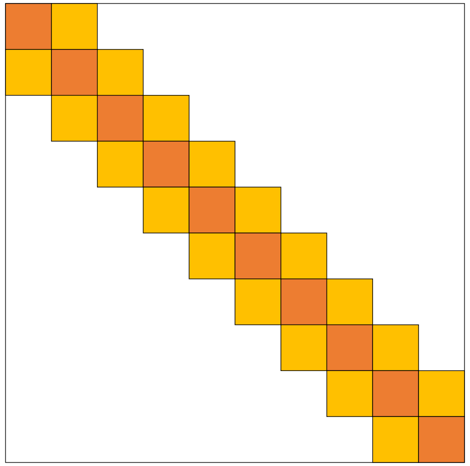




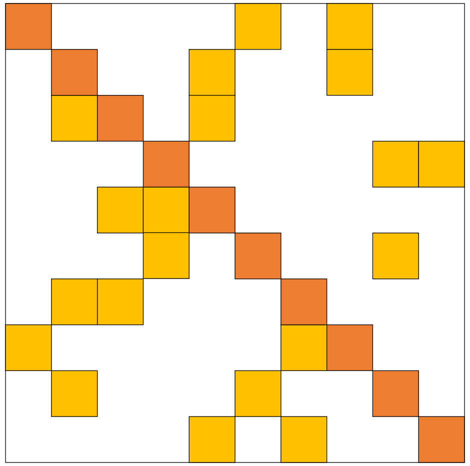
Sparse Transformers: Adjacency Matrices



global

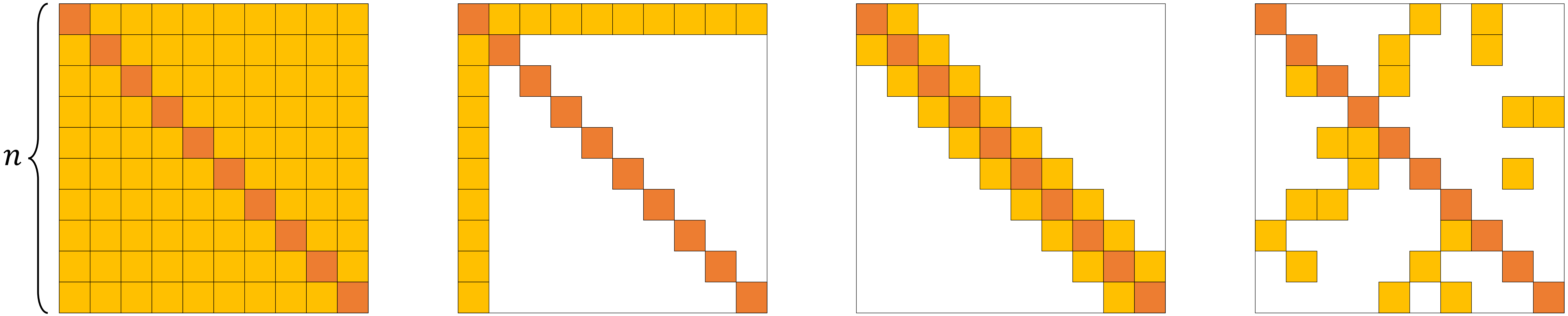


window
$w = 3$



random
r = 2

Sparse Transformers: Complexity



$O(n^2)$ $O(n)$ $O(n)$ $O(n)$

Sparse Transformers: Universality

- Theorem

  For every $f: R^{n\times d} \rightarrow R^{n\times d}$ *with a compact support,*

  there exists a sparse transformer t *with a global adjacency matrix*

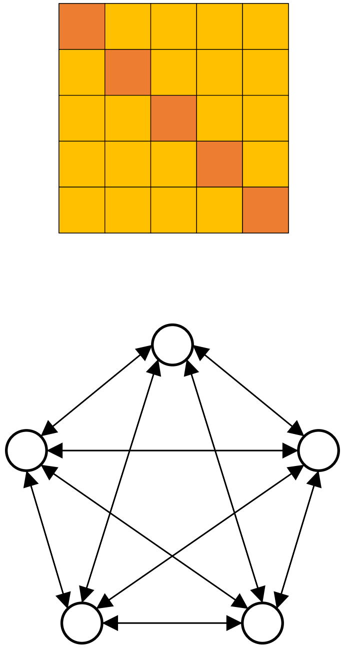
  s.t. $d(f, t)$ *is as small as desired*

Sparse Transformers: Turing Completeness

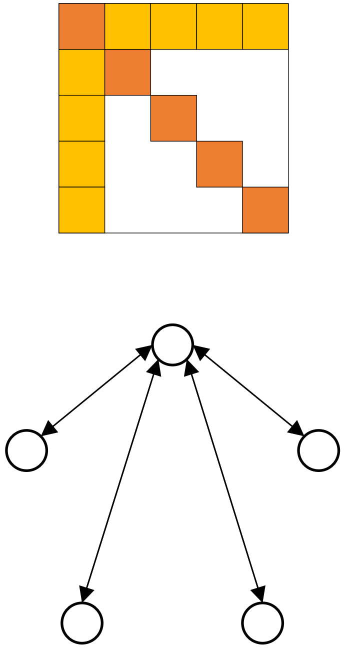
- Theorem

  The class of the sparse transformers with $O(n)$ adjacency matrices is Turing-complete

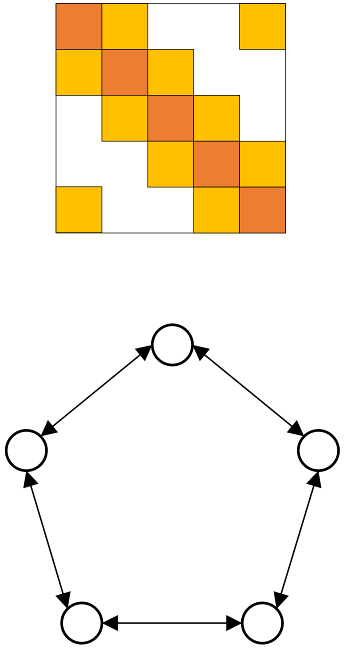
Sparse Transformers: Graph Theory



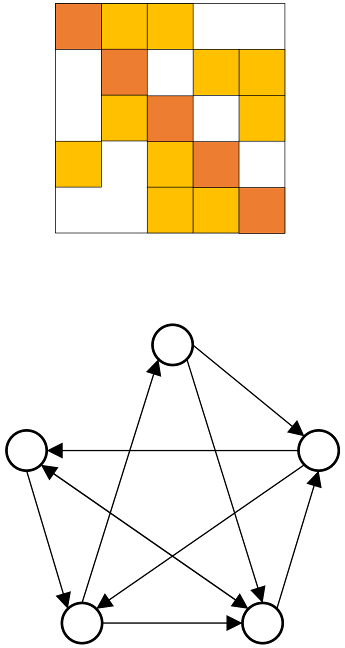
complete



star



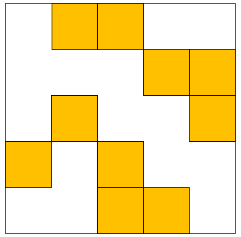
circular



Erdős-Rényi
$ER(5,2)$

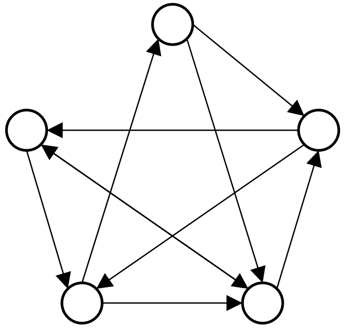
Sparse Transformers: Graph Theory

The Average Distances in Random Graphs with Given Expected Degrees (PNAS 2002)



- Theorem 1

  The average distance (shortest paths between nodes) in $ER(n, d)$ are almost surely in $O(\log n / \log d)$

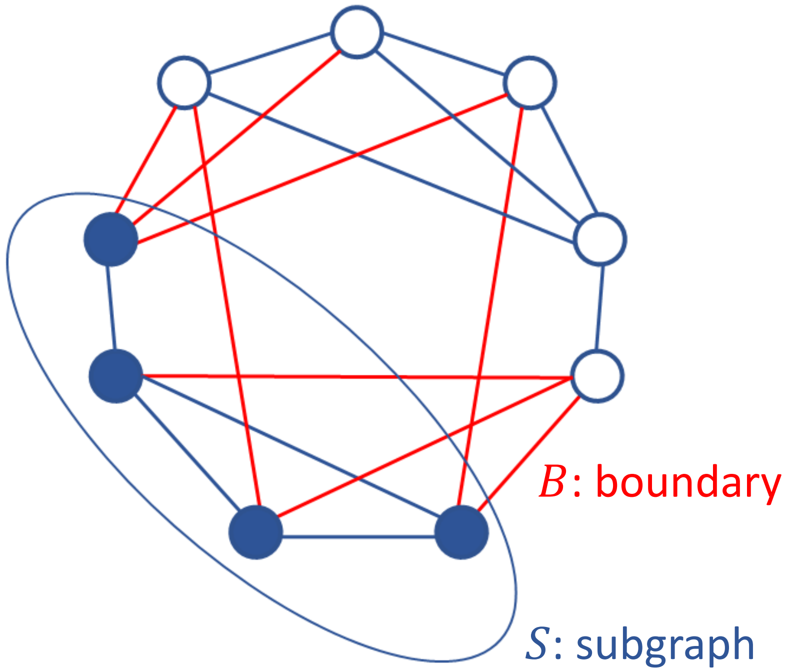


Erdős-Rényi
$ER(5, 2)$

Sparse Transformers: Graph Theory

- Graph Expansion

$$\text{Expansion}(G) \coloneqq \min_{S} \frac{|B|}{|S|}$$

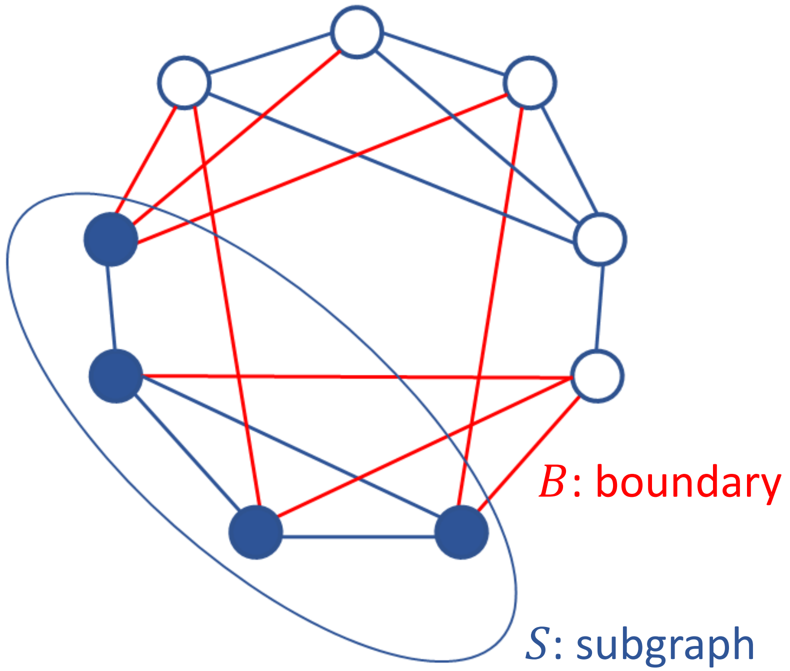


Sparse Transformers: Graph Theory

- Theorem 2.1

The expansion of a regular graph is bounded by $\lambda_1 - \lambda_2$, where λ_i is the i^{th} largest eigenvalue of the adjacency matrix

-> Expansion of a graph is related to its spectral properties

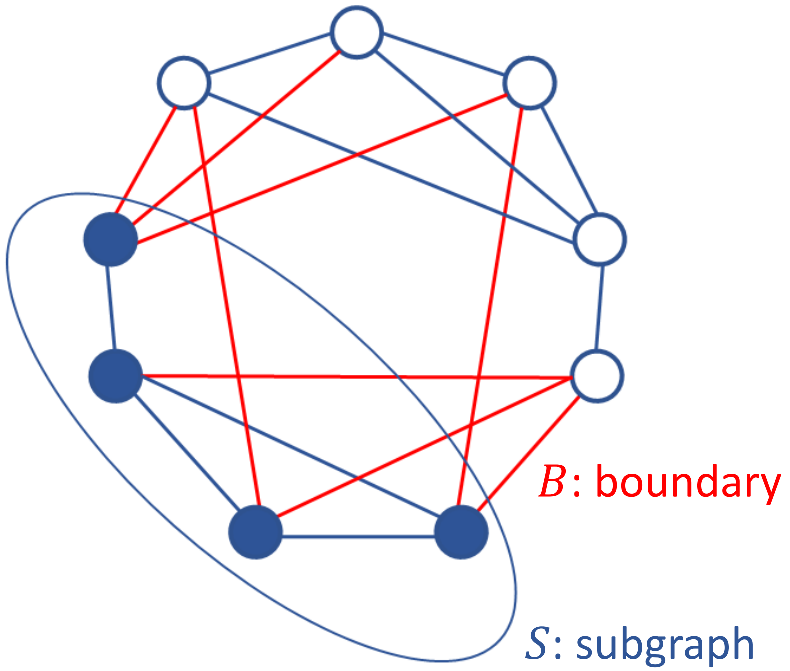


Sparse Transformers: Graph Theory

- Theorem 2.2

  An Erdős-Rényi graph approximates its corresponding complete graph spectrally

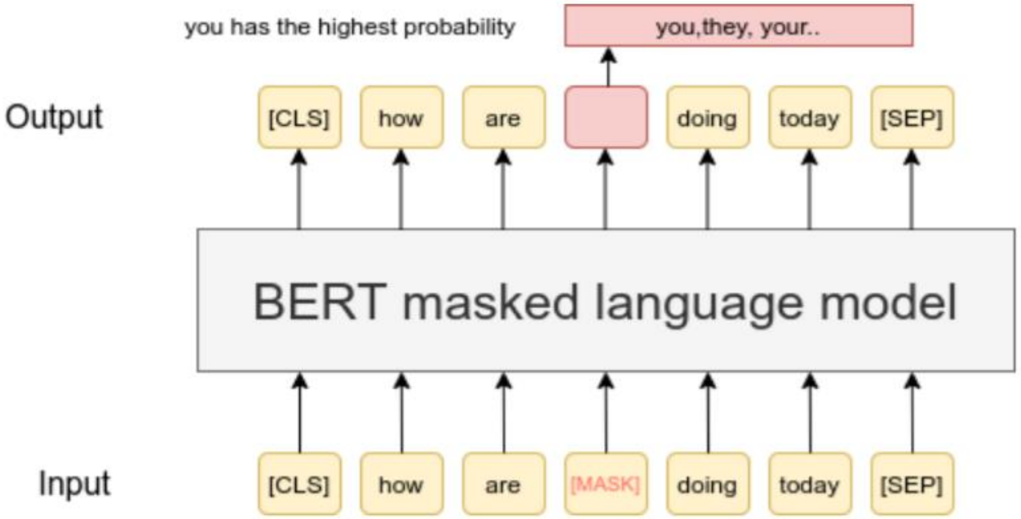
  -> Sparse transformers expand contexts fast like full transformers



Applications: NLP

- Masked Language Modeling (MLM)

Self-supervised learning to give each word a contextualized embedding



Applications: NLP

- Masked Language Modeling (MLM)

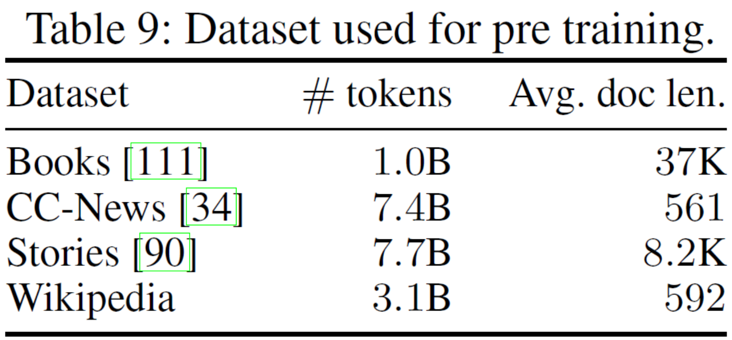


Table 9: Dataset used for pre training.

Dataset	# tokens	Avg. doc len.
Books [111]	1.0B	37K
CC-News [34]	7.4B	561
Stories [90]	7.7B	8.2K
Wikipedia	3.1B	592

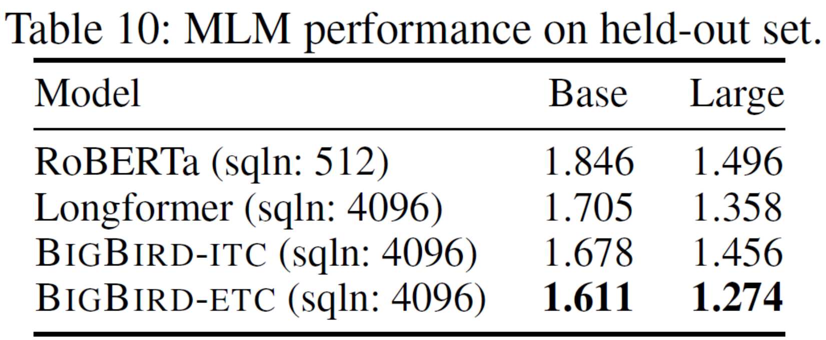


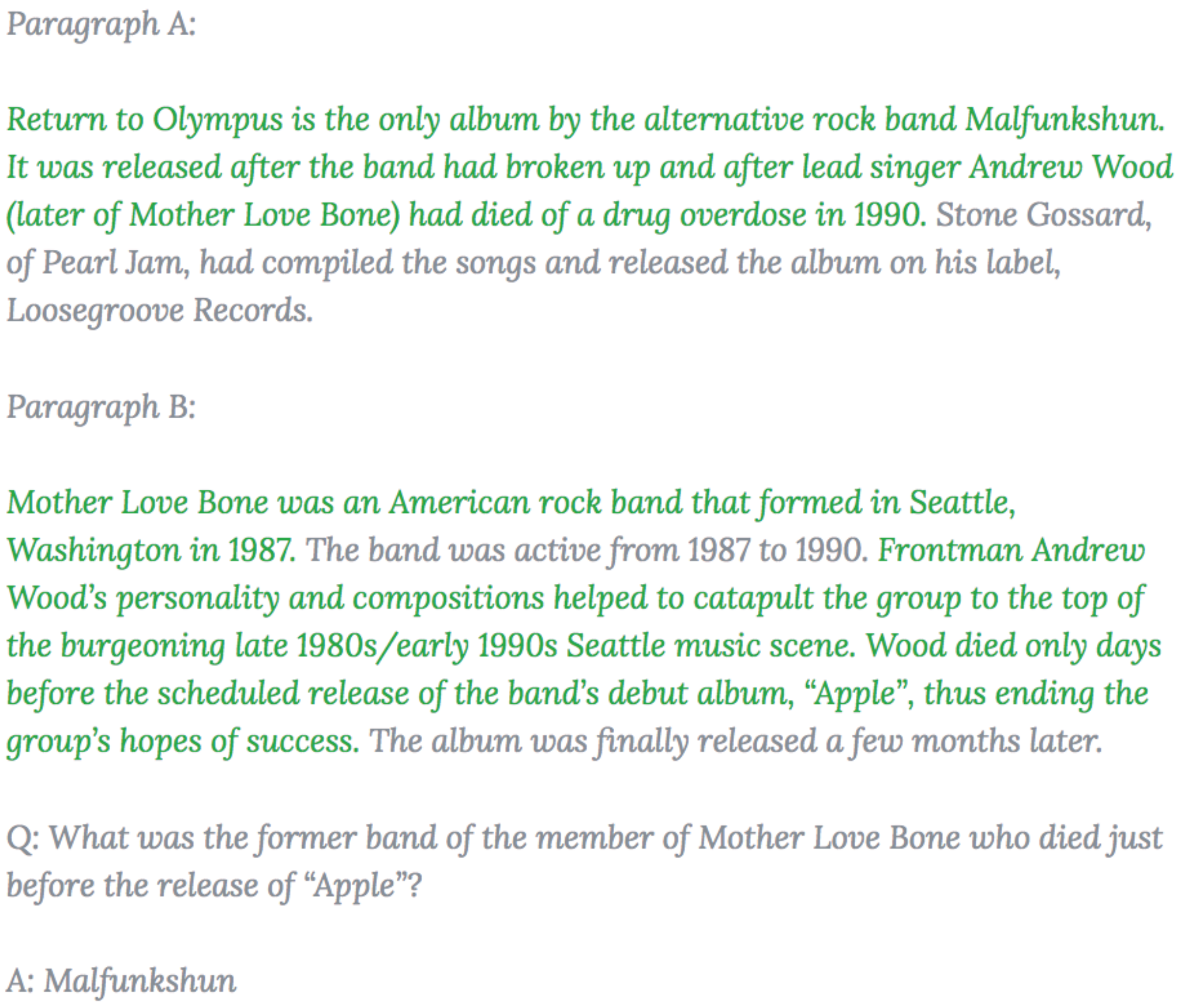
Table 10: MLM performance on held-out set.

Model	Base	Large
RoBERTa (sqln: 512)	1.846	1.496
Longformer (sqln: 4096)	1.705	1.358
BIGBIRD-ITC (sqln: 4096)	1.678	1.456
BIGBIRD-ETC (sqln: 4096)	**1.611**	**1.274**

Application: NLP

- Question Answering (QA)

  Find the answer and its supporting evidence in a paragraph, a document, or multiple documents



Paragraph A:

Return to Olympus is the only album by the alternative rock band Malfunkshun. It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990. Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

Paragraph B:

Mother Love Bone was an American rock band that formed in Seattle, Washington in 1987. The band was active from 1987 to 1990. Frontman Andrew Wood's personality and compositions helped to catapult the group to the top of the burgeoning late 1980s/early 1990s Seattle music scene. Wood died only days before the scheduled release of the band's debut album, "Apple", thus ending the group's hopes of success. The album was finally released a few months later.

Q: What was the former band of the member of Mother Love Bone who died just before the release of "Apple"?

A: Malfunkshun

Applications: NLP

- Question Answering (QA)

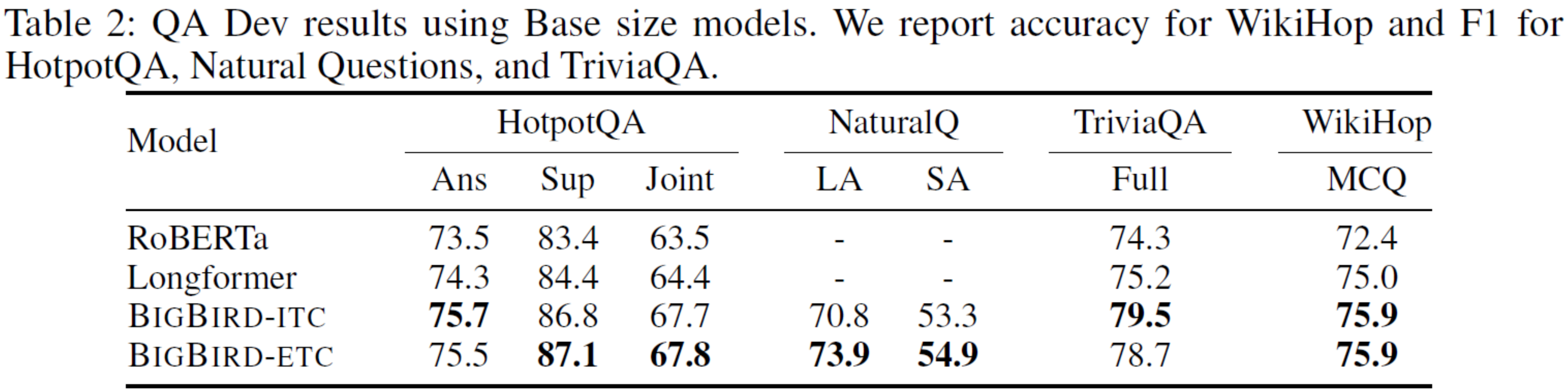


Table 2: QA Dev results using Base size models. We report accuracy for WikiHop and F1 for HotpotQA, Natural Questions, and TriviaQA.

Model	HotpotQA			NaturalQ		TriviaQA	WikiHop
	Ans	Sup	Joint	LA	SA	Full	MCQ
RoBERTa	73.5	83.4	63.5	-	-	74.3	72.4
Longformer	74.3	84.4	64.4	-	-	75.2	75.0
BigBird-ITC	**75.7**	86.8	67.7	70.8	53.3	**79.5**	**75.9**
BigBird-ETC	75.5	**87.1**	**67.8**	**73.9**	**54.9**	78.7	**75.9**

Applications: NLP

- Question Answering (QA)

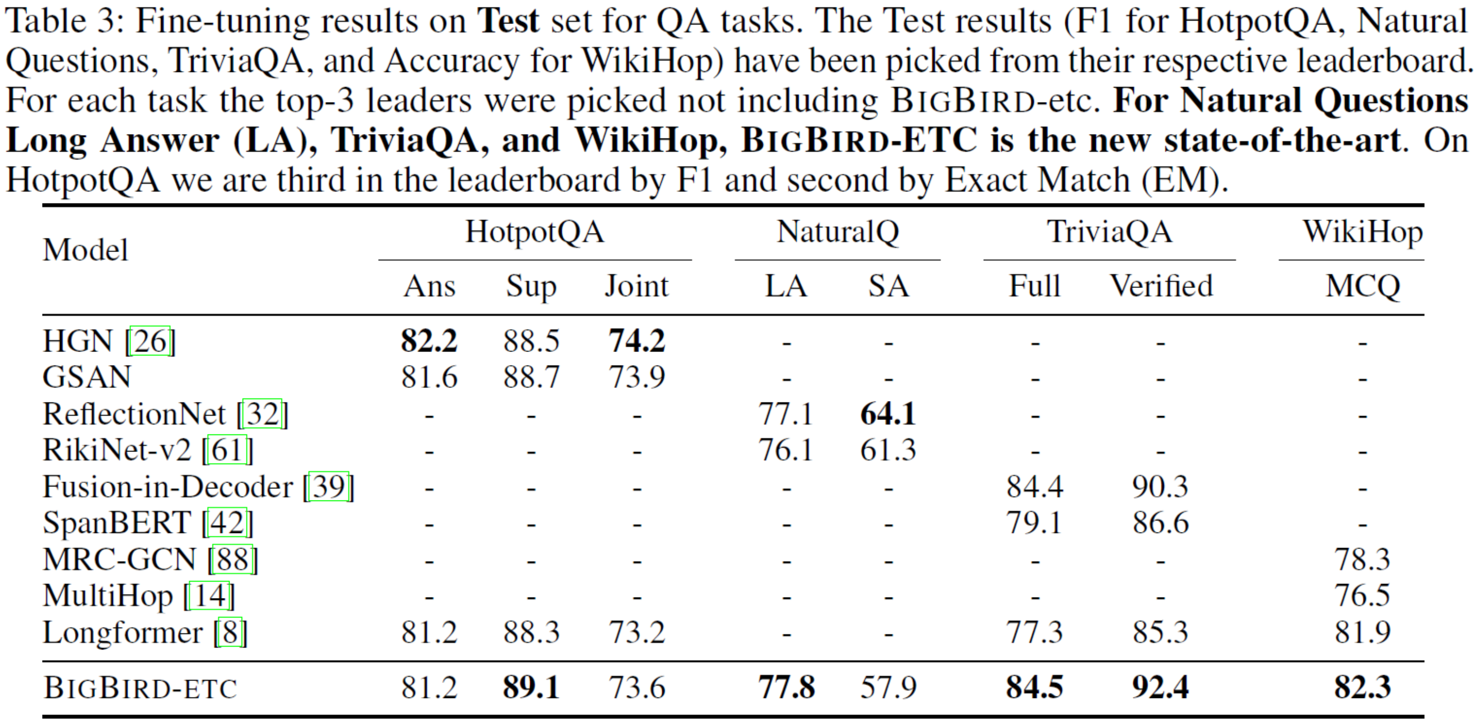


Table 3: Fine-tuning results on **Test** set for QA tasks. The Test results (F1 for HotpotQA, Natural Questions, TriviaQA, and Accuracy for WikiHop) have been picked from their respective leaderboard. For each task the top-3 leaders were picked not including BigBird-etc. **For Natural Questions Long Answer (LA), TriviaQA, and WikiHop, BigBird-ETC is the new state-of-the-art**. On HotpotQA we are third in the leaderboard by F1 and second by Exact Match (EM).

Model	HotpotQA			NaturalQ		TriviaQA		WikiHop
	Ans	Sup	Joint	LA	SA	Full	Verified	MCQ
HGN [26]	**82.2**	88.5	**74.2**	-	-	-	-	-
GSAN	81.6	88.7	73.9	-	-	-	-	-
ReflectionNet [32]	-	-	-	77.1	**64.1**	-	-	-
RikiNet-v2 [61]	-	-	-	76.1	61.3	-	-	-
Fusion-in-Decoder [39]	-	-	-	-	-	84.4	90.3	-
SpanBERT [42]	-	-	-	-	-	79.1	86.6	-
MRC-GCN [88]	-	-	-	-	-	-	-	78.3
MultiHop [14]	-	-	-	-	-	-	-	76.5
Longformer [8]	81.2	88.3	73.2	-	-	77.3	85.3	81.9
BigBird-ETC	81.2	**89.1**	73.6	**77.8**	57.9	**84.5**	**92.4**	**82.3**

Applications: NLP

- Classification

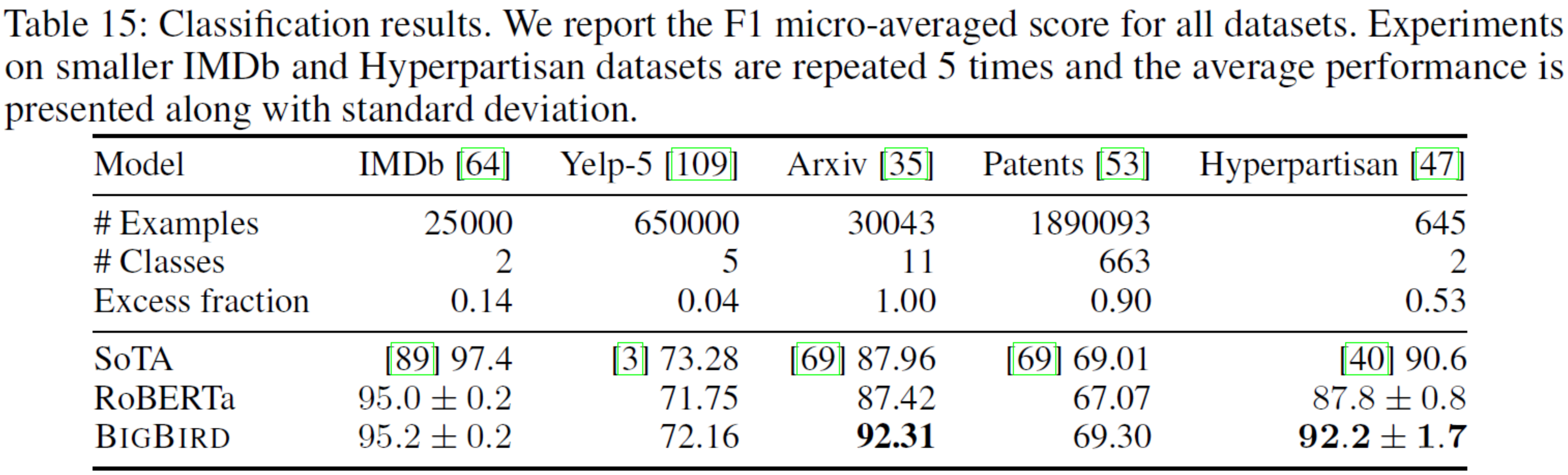


Table 15: Classification results. We report the F1 micro-averaged score for all datasets. Experiments on smaller IMDb and Hyperpartisan datasets are repeated 5 times and the average performance is presented along with standard deviation.

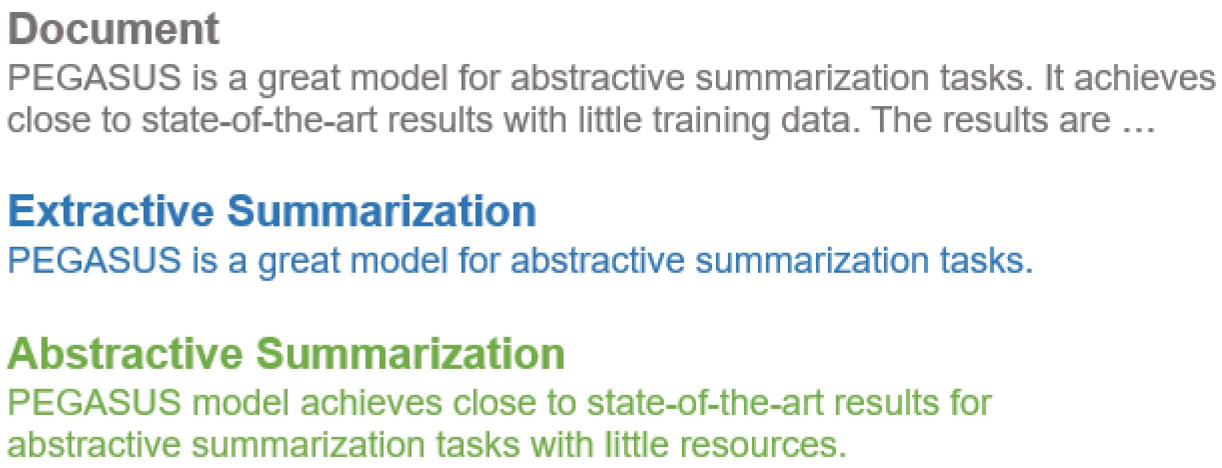
Model	IMDb [64]	Yelp-5 [109]	Arxiv [35]	Patents [53]	Hyperpartisan [47]
# Examples	25000	650000	30043	1890093	645
# Classes	2	5	11	663	2
Excess fraction	0.14	0.04	1.00	0.90	0.53
SoTA	[89] 97.4	[3] 73.28	[69] 87.96	[69] 69.01	[40] 90.6
RoBERTa	95.0 ± 0.2	71.75	87.42	67.07	87.8 ± 0.8
BIGBIRD	95.2 ± 0.2	72.16	**92.31**	69.30	**92.2 ± 1.7**

*Excess fraction: proportion of samples longer than 512 words

Applications: NLP

- Summarization

Abstractive summarization via seq2seq learning



Document

PEGASUS is a great model for abstractive summarization tasks. It achieves close to state-of-the-art results with little training data. The results are …

Extractive Summarization

PEGASUS is a great model for abstractive summarization tasks.

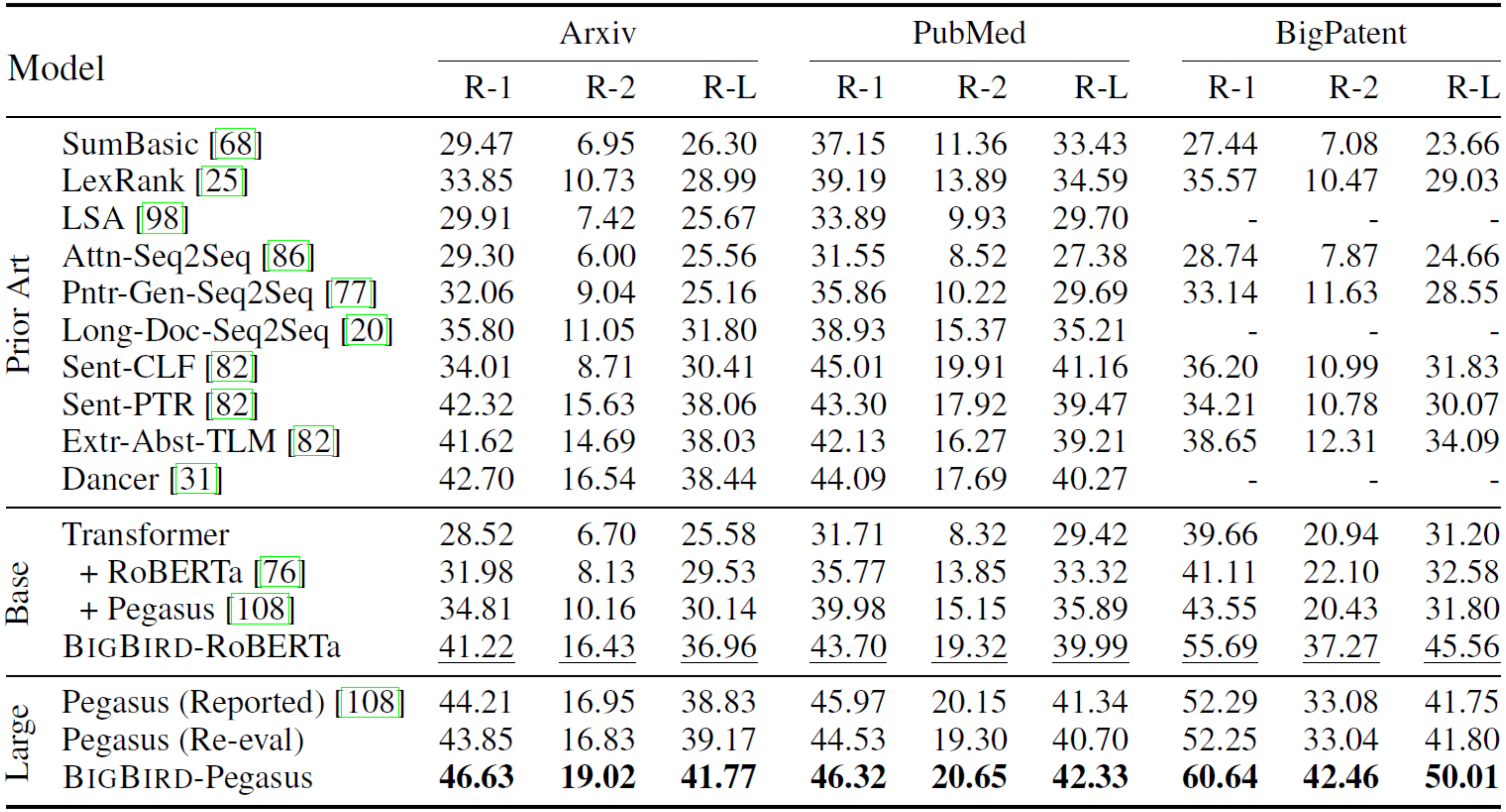
Abstractive Summarization

PEGASUS model achieves close to state-of-the-art results for abstractive summarization tasks with little resources.

Applications: NLP

- Summarization

Table 4: Summarization ROUGE score for long documents.



	Model	Arxiv			PubMed			BigPatent		
		R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Prior Art	SumBasic [68]	29.47	6.95	26.30	37.15	11.36	33.43	27.44	7.08	23.66
	LexRank [25]	33.85	10.73	28.99	39.19	13.89	34.59	35.57	10.47	29.03
	LSA [98]	29.91	7.42	25.67	33.89	9.93	29.70	-	-	-
	Attn-Seq2Seq [86]	29.30	6.00	25.56	31.55	8.52	27.38	28.74	7.87	24.66
	Pntr-Gen-Seq2Seq [77]	32.06	9.04	25.16	35.86	10.22	29.69	33.14	11.63	28.55
	Long-Doc-Seq2Seq [20]	35.80	11.05	31.80	38.93	15.37	35.21	-	-	-
	Sent-CLF [82]	34.01	8.71	30.41	45.01	19.91	41.16	36.20	10.99	31.83
	Sent-PTR [82]	42.32	15.63	38.06	43.30	17.92	39.47	34.21	10.78	30.07
	Extr-Abst-TLM [82]	41.62	14.69	38.03	42.13	16.27	39.21	38.65	12.31	34.09
	Dancer [31]	42.70	16.54	38.44	44.09	17.69	40.27	-	-	-
Base	Transformer	28.52	6.70	25.58	31.71	8.32	29.42	39.66	20.94	31.20
	+ RoBERTa [76]	31.98	8.13	29.53	35.77	13.85	33.32	41.11	22.10	32.58
	+ Pegasus [108]	34.81	10.16	30.14	39.98	15.15	35.89	43.55	20.43	31.80
	BIGBIRD-RoBERTa	41.22	16.43	36.96	43.70	19.32	39.99	55.69	37.27	45.56
Large	Pegasus (Reported) [108]	44.21	16.95	38.83	45.97	20.15	41.34	52.29	33.08	41.75
	Pegasus (Re-eval)	43.85	16.83	39.17	44.53	19.30	40.70	52.25	33.04	41.80
	BIGBIRD-Pegasus	**46.63**	**19.02**	**41.77**	**46.32**	**20.65**	**42.33**	**60.64**	**42.46**	**50.01**

Applications: Genomics

- DNA MLM

Self-supervised learning to give each *DNA word* a contextualized embedding according to its *DNA sentence*

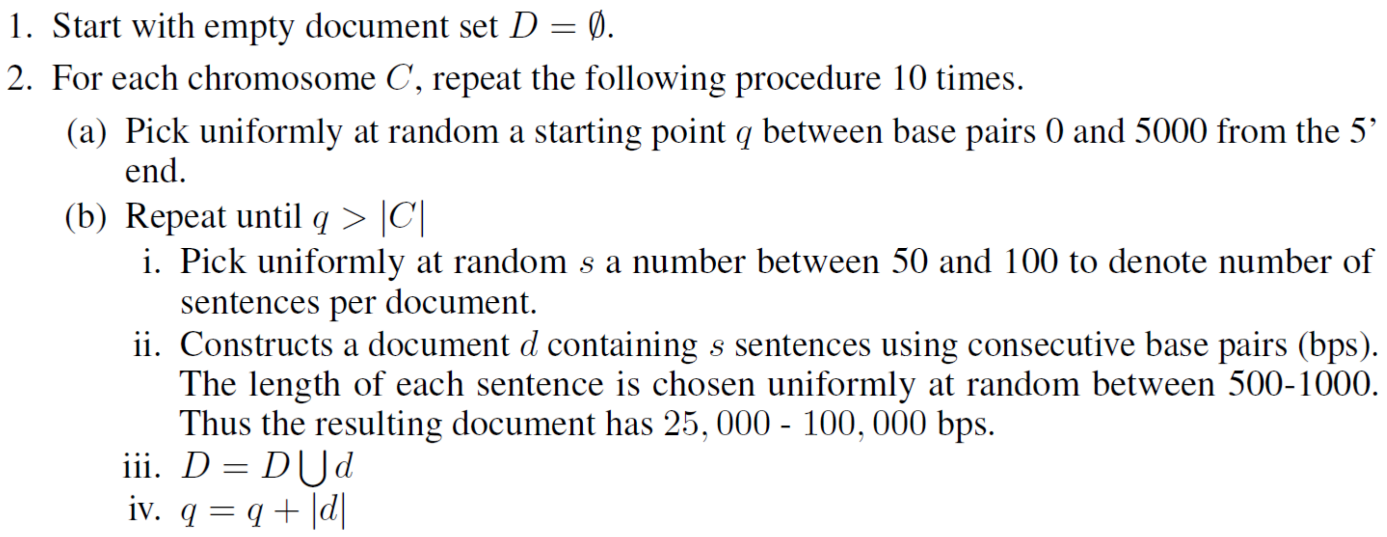
-> *DNA words*: learned via Byte-Pair Encoding (BPE)

Applications: Genomics

- DNA MLM

Self-supervised learning to give each *DNA word* a contextualized embedding according to its *DNA sentence*

-> *DNA sentences*:



```
1. Start with empty document set $D = \emptyset$.
2. For each chromosome $C$, repeat the following procedure 10 times.
   (a) Pick uniformly at random a starting point $q$ between base pairs 0 and 5000 from the 5' end.
   (b) Repeat until $q > |C|$
       i. Pick uniformly at random $s$ a number between 50 and 100 to denote number of sentences per document.
       ii. Constructs a document $d$ containing $s$ sentences using consecutive base pairs (bps). The length of each sentence is chosen uniformly at random between 500-1000. Thus the resulting document has $25,000$ - $100,000$ bps.
       iii. $D = D \bigcup d$
       iv. $q = q + |d|$
```

Applications: Genomics

- DNA MLM

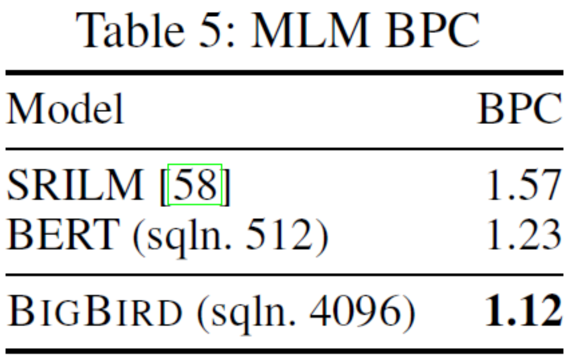


Table 5: MLM BPC

Model	BPC
SRILM [58]	1.57
BERT (sqln. 512)	1.23
BigBird (sqln. 4096)	**1.12**

*SRILM: n-gram (k-mer) models

Applications: Genomics

- Promoter Region Prediction

  Learning to classify a given DNA fragment as a promoter or a non-promoter sequence

Applications: Genomics

DeePromoter: Robust Promoter Predictor Using Deep Learning (Frontiers in genetics 2019)

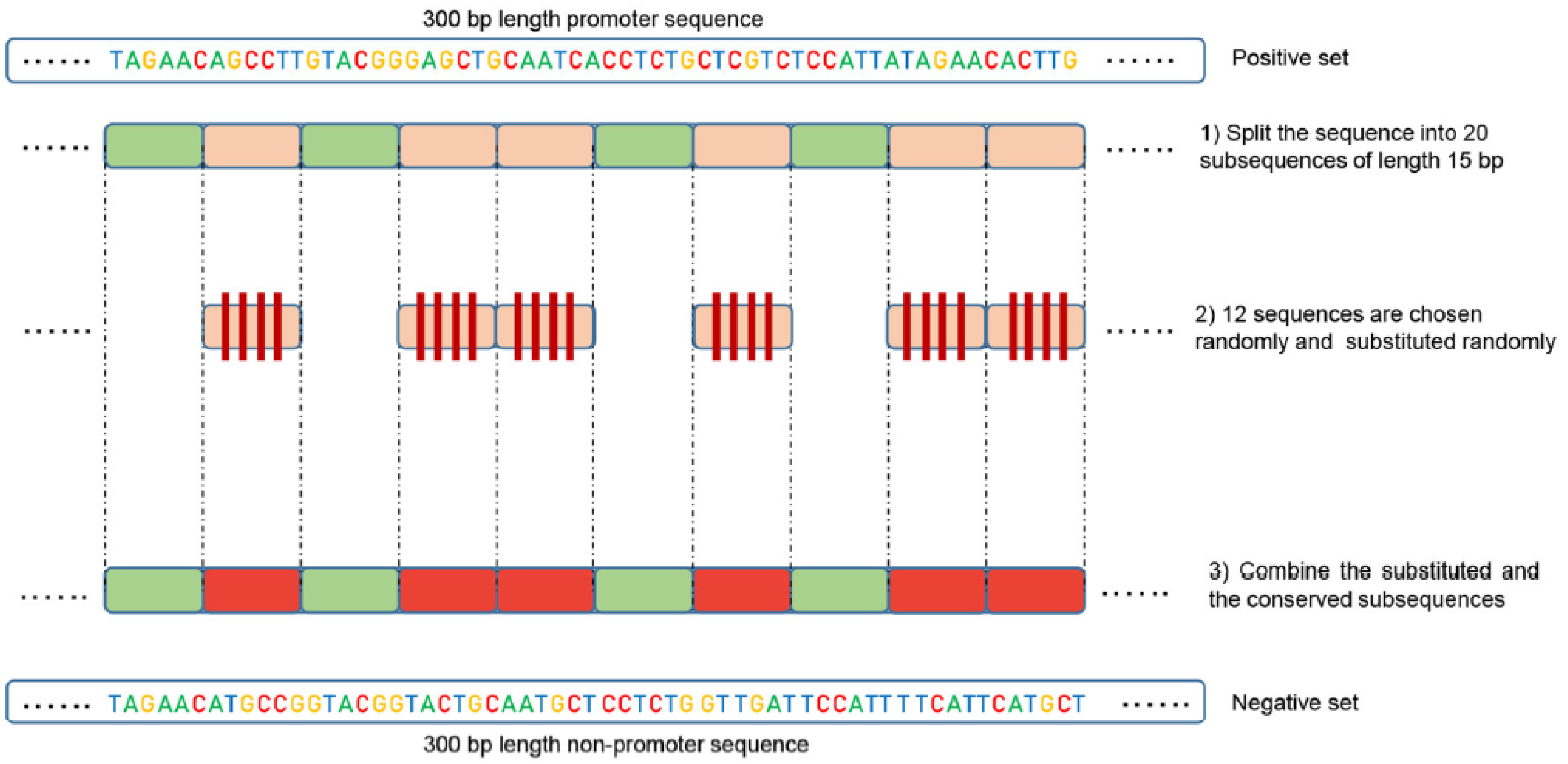

300 bp length promoter sequence
...... TAGAACAGCCTTGTACGGGAGCTGCAATCACCTCTGCTCGTCTCCATTATAGAACACTTG ......
Positive set
1) Split the sequence into 20 subsequences of length 15 bp
2) 12 sequences are chosen randomly and substituted randomly
3) Combine the substituted and the conserved subsequences
...... TAGAACATGCCGGTACGGTACTGCAATGCT CCTCTG GT TGAT TCCATT T TCATTCATGCT ......
Negative set
300 bp length non-promoter sequence

FIGURE 1 | Illustration of the negative set construction method. Green represents the randomly conserved subsequences while red represents the randomly chosen and substituted ones.

Applications: Genomics

- Promoter Region Prediction

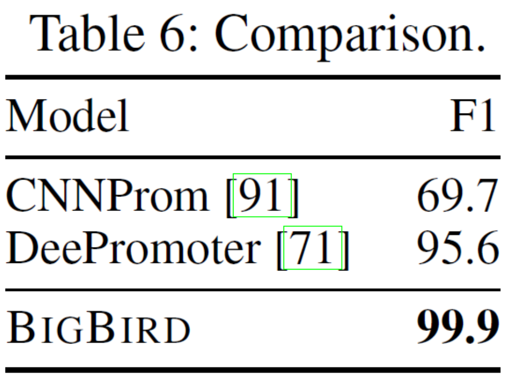


Table 6: Comparison.

Model	F1
CNNProm [91]	69.7
DeePromoter [71]	95.6
BigBird	**99.9**

*DeePromoter: CNN + LSTM

Applications: Genomics

- Chromatin-Profile Prediction

    Learning to predict chromatin-profiling from non-coding genomic sequence

Applications: Genomics

Predicting Effects of Noncoding Variants with Deep Learning-based Sequence Model (Nat Methods 2015)

- 2.4M noncoding variants
- 919 chromatin-profile
  - 690 transcription factors (TF) binding profiles for 160 different TFs
  - 125 DNase I sensitivity (DHS) profiles
  - 104 histone-mark (HM) profiles

Applications: Genomics

- Chromatin-Profile Prediction

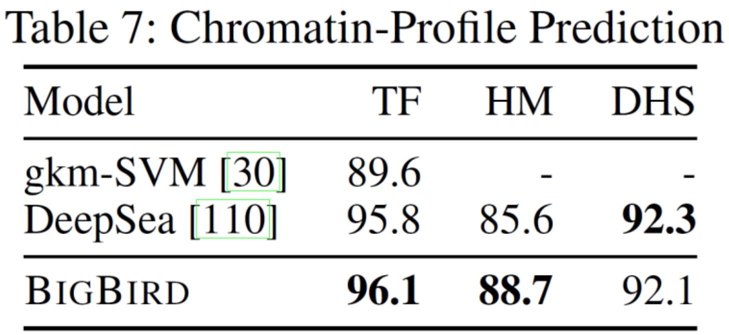
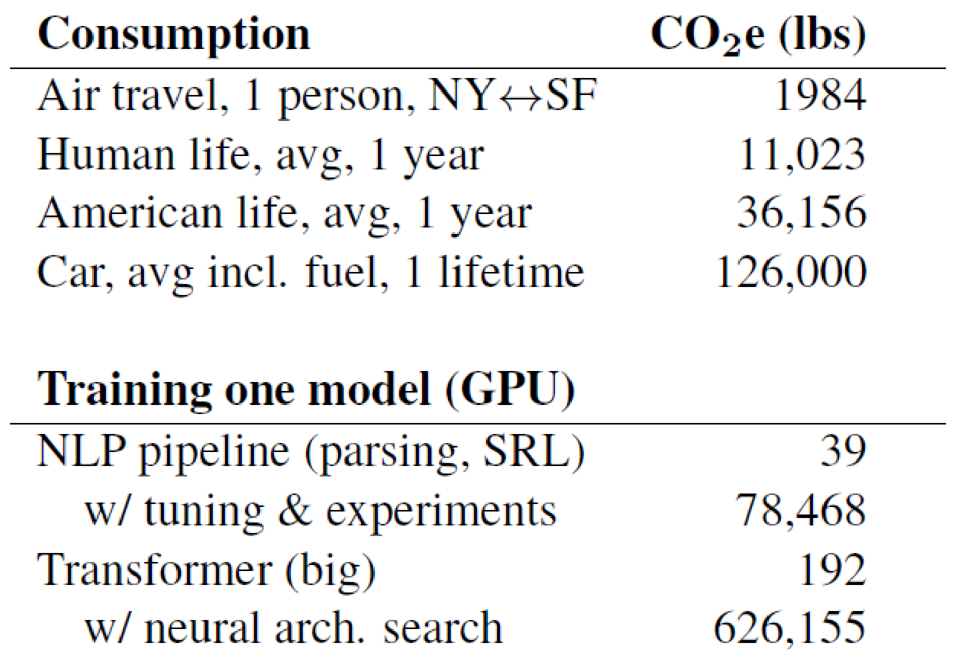


Table 7: Chromatin-Profile Prediction

Model	TF	HM	DHS
gkm-SVM [30]	89.6	-	-
DeepSea [110]	95.8	85.6	**92.3**
BIGBIRD	**96.1**	**88.7**	92.1

Earth Day 2022

Energy and Policy Considerations for Deep Learning in NLP (ACL 2019)



Consumption	**CO_2e (lbs)**
Air travel, 1 person, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experiments	78,468
Transformer (big)	192
w/ neural arch. search	626,155

Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]