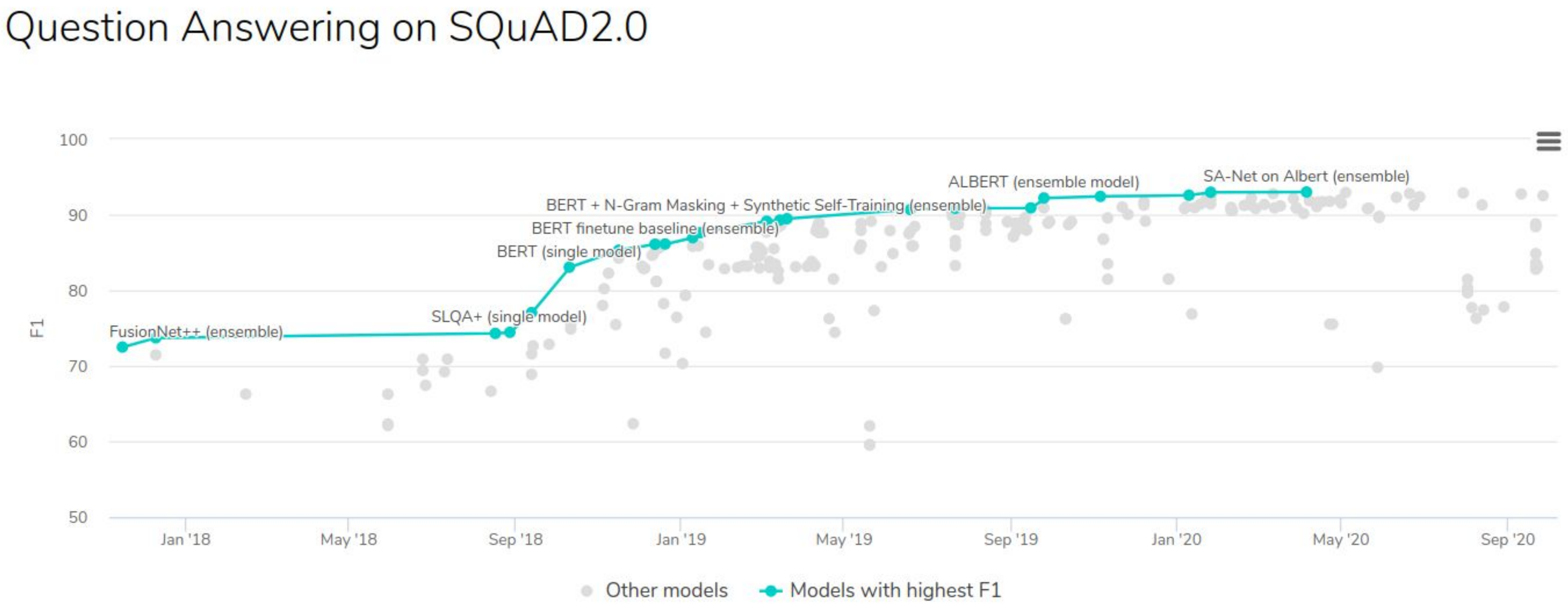
Attention sparsification

Look into the future and the past (behind the context window)

FI:PV212: Readings in Digital ...
Michal Štefánik
stefanik.m@mail.muni.cz

Why talk about it?



https://paperswithcode.com/sota/question-answering-on-squad20

Why talk about it?

Benchmarks

Add a Result

TREND	DATASET	BEST METHOD	PAPER TITLE	PAPER	CODE	COMPARE
	CoNLL 2003 (English)	CNN Large + fine-tune	Cloze-driven Pretraining of Self-attention Networks			See all
	Ontonotes v5 (English)	BERT-MRC+DSC	Dice Loss for Data-imbalanced NLP Tasks			See all
	ACE 2005	BERT-MRC	A Unified MRC Framework for Named Entity Recognition			See all
	GENIA	BERT-MRC	A Unified MRC Framework for Named Entity Recognition			See all
	CoNLL++	CrossWeigh + Pooled Flair	CrossWeigh: Training Named Entity Tagger from Imperfect Annotations			See all
	Long-tail emerging entities	Flair embeddings	Contextual String Embeddings for Sequence Labeling			See all
	BC5CDR	NER+PA+RL (PubMed)	Reinforcement-based denoising of distantly supervised NER with partial annotation			See all
	JNLPBA	BioBERT	BioBERT: a pre-trained biomedical language representation model for biomedical text mining			See all
	SciERC	SpERT	Span-based Joint Entity and Relation Extraction with Transformer Pre-training			See all

https://paperswithcode.com/task/named-entity-recognition-ner

Why talk about it?

Benchmarks

Add a Result

TREND	DATASET	BEST METHOD	PAPER TITLE	PAPER	CODE	COMPARE
	Cityscapes test	HRNet-OCR (Hierarchical Multi-Scale Attention)	Hierarchical Multi-Scale Attention for Semantic Segmentation			See all
	PASCAL VOC 2012 test	EfficientNet-L2+NAS-FPN (single scale test, with self-training)	Rethinking Pre-training and Self-training			See all
	PASCAL Context	ResNeSt-269	ResNeSt: Split-Attention Networks			See all
	Cityscapes val	HRNet-OCR (Hierarchical Multi-Scale Attention)	Hierarchical Multi-Scale Attention for Semantic Segmentation			See all
	ADE20K val	ResNeSt-200	ResNeSt: Split-Attention Networks			See all
	ADE20K	ResNeSt-200	ResNeSt: Split-Attention Networks			See all

https://paperswithcode.com/task/semantic-segmentation

Attention [1]

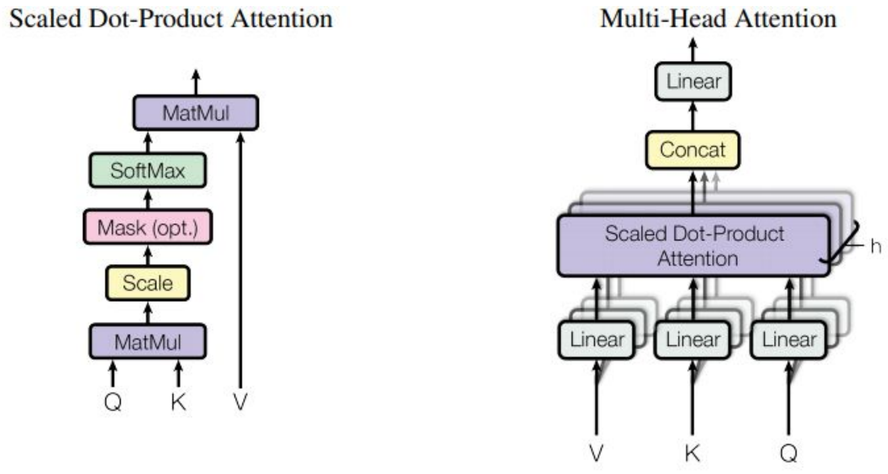
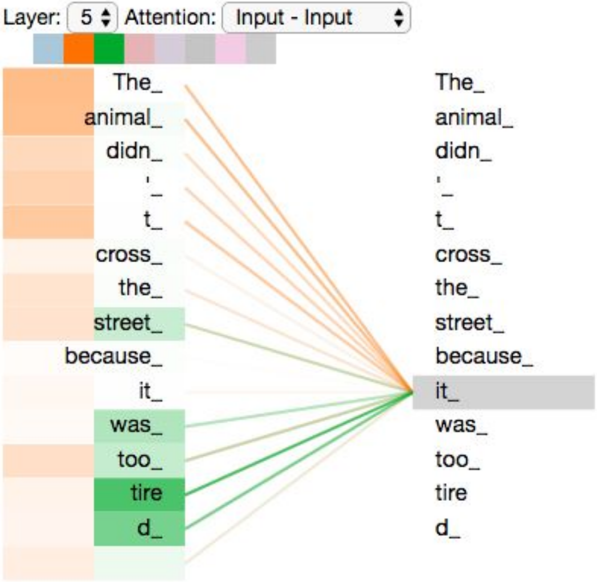


Figure 2: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.



[1]: https://arxiv.org/abs/1706.03762 (Attention is All You Need)

[3]: https://colab.research.google.com/github/tensorflow/tensor2tensor/blob/master/tensor2tensor/notebooks/hello_t2t.ipynb

Transformer [1]

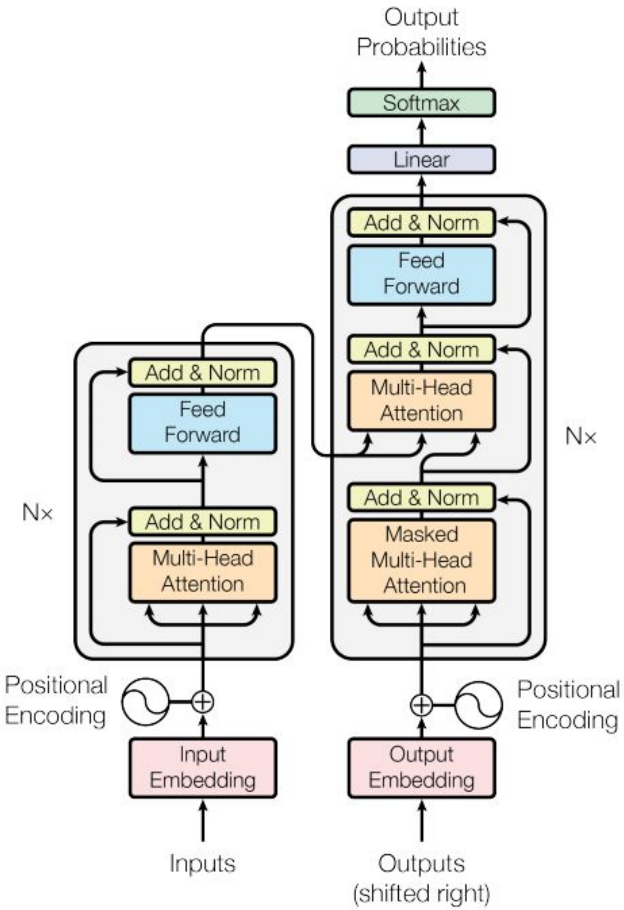


Figure 1: The Transformer - model architecture.

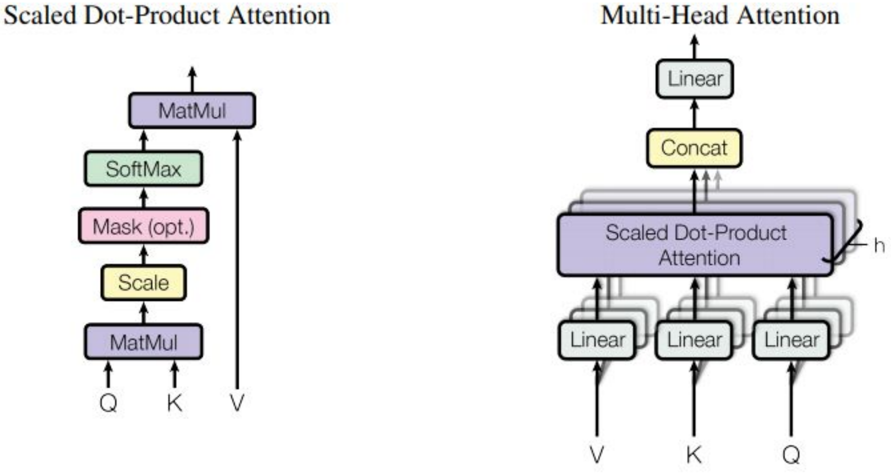
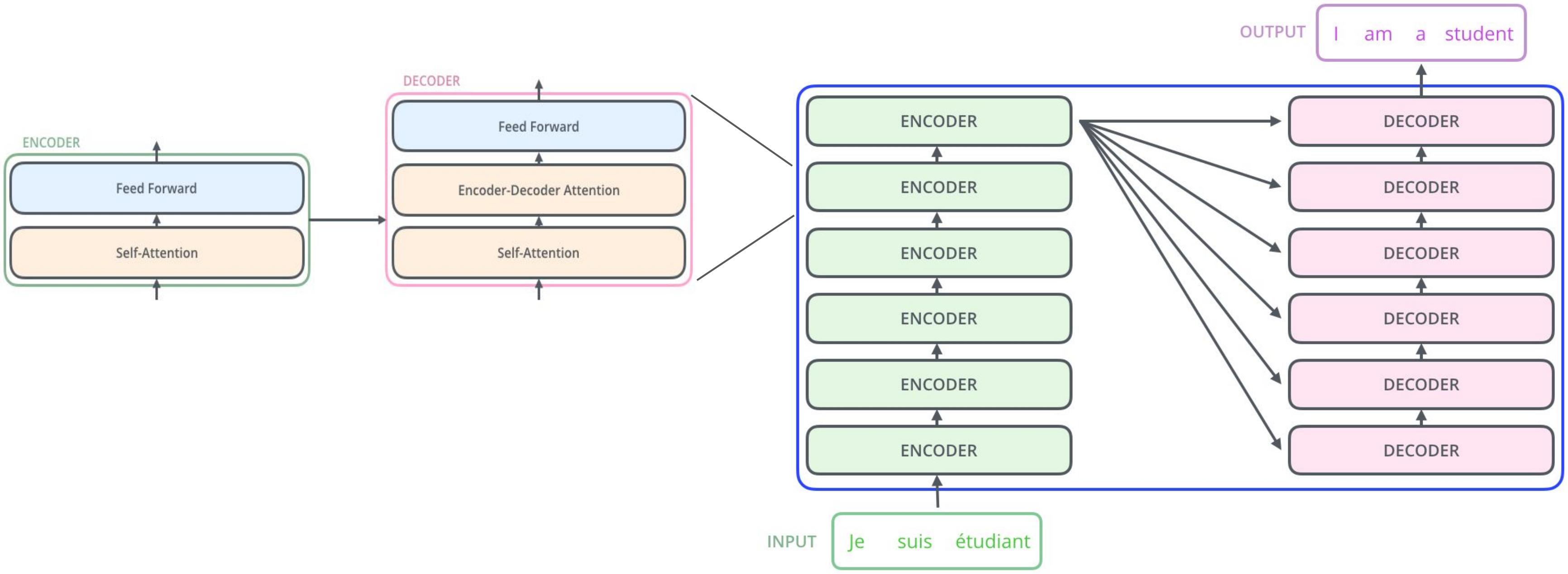


Figure 2: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.

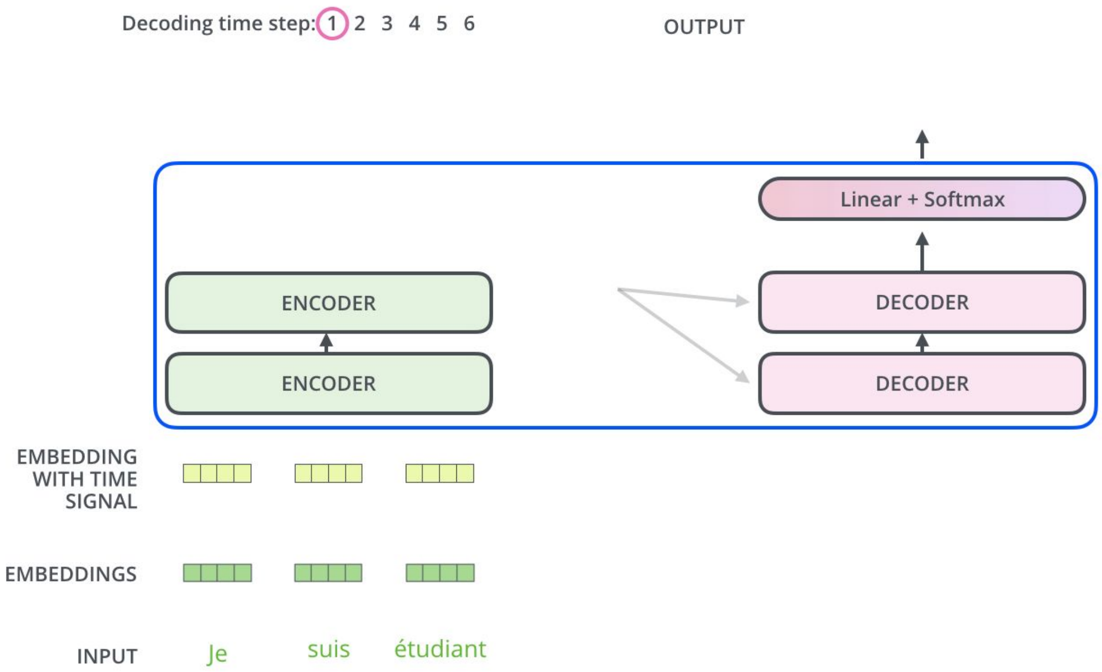
[1]: https://arxiv.org/abs/1706.03762 (Attention is All You Need)

Transformer as autoencoder



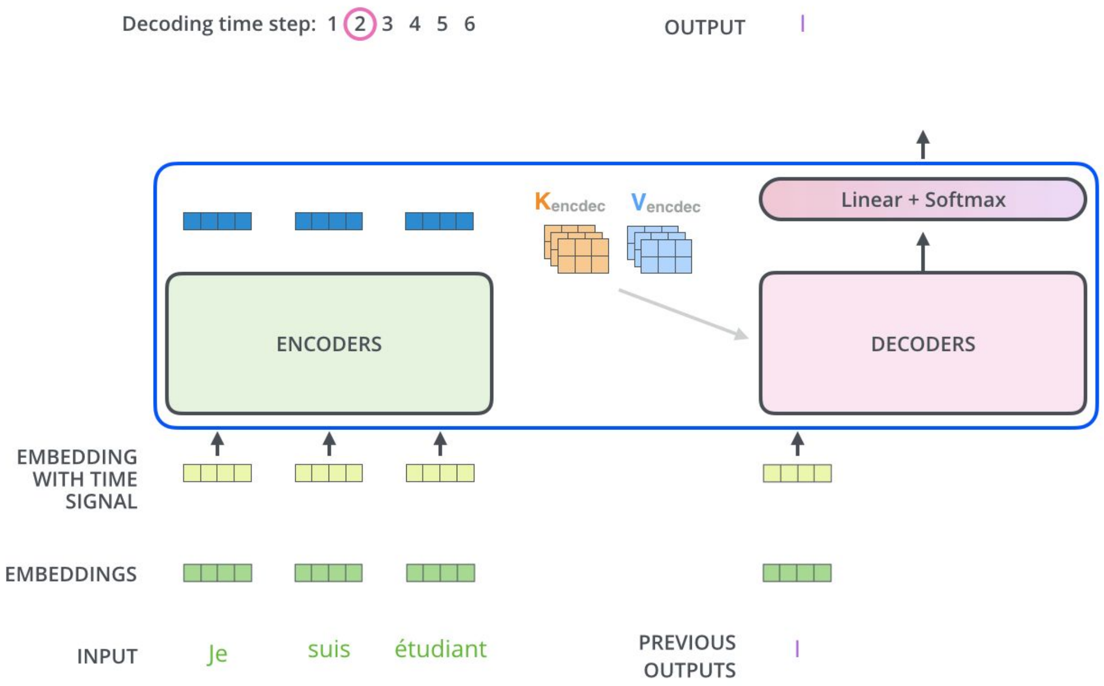
[2]: http://jalammar.github.io/illustrated-transformer/

Transformer as autoencoder



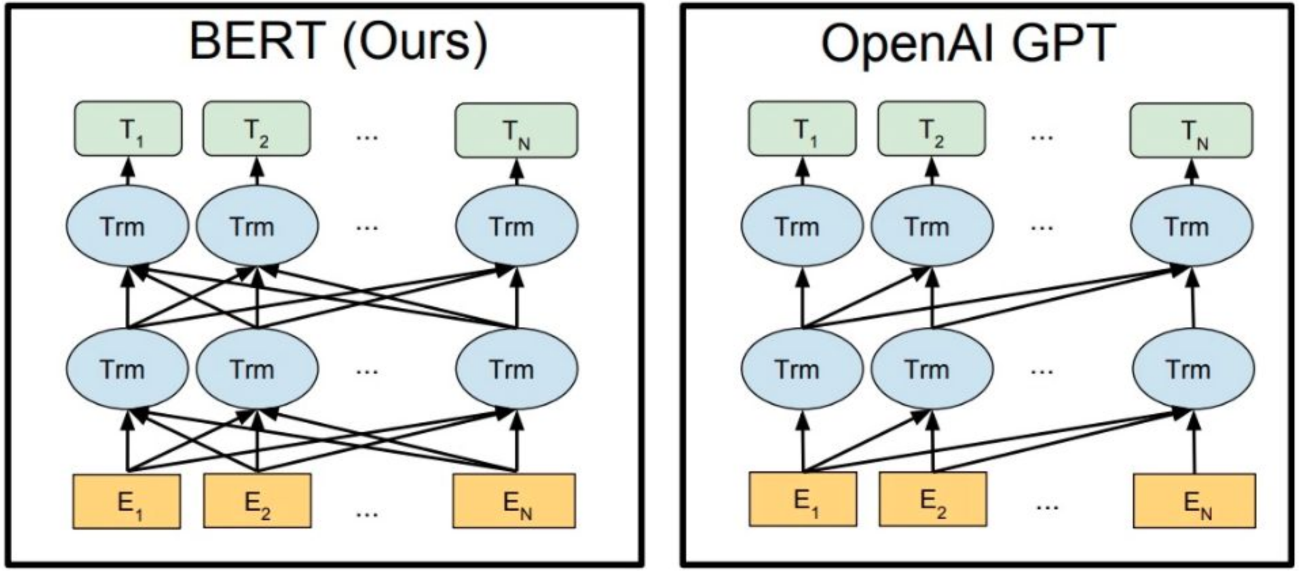
[2]: http://jalammar.github.io/illustrated-transformer/

Transformer as autoencoder

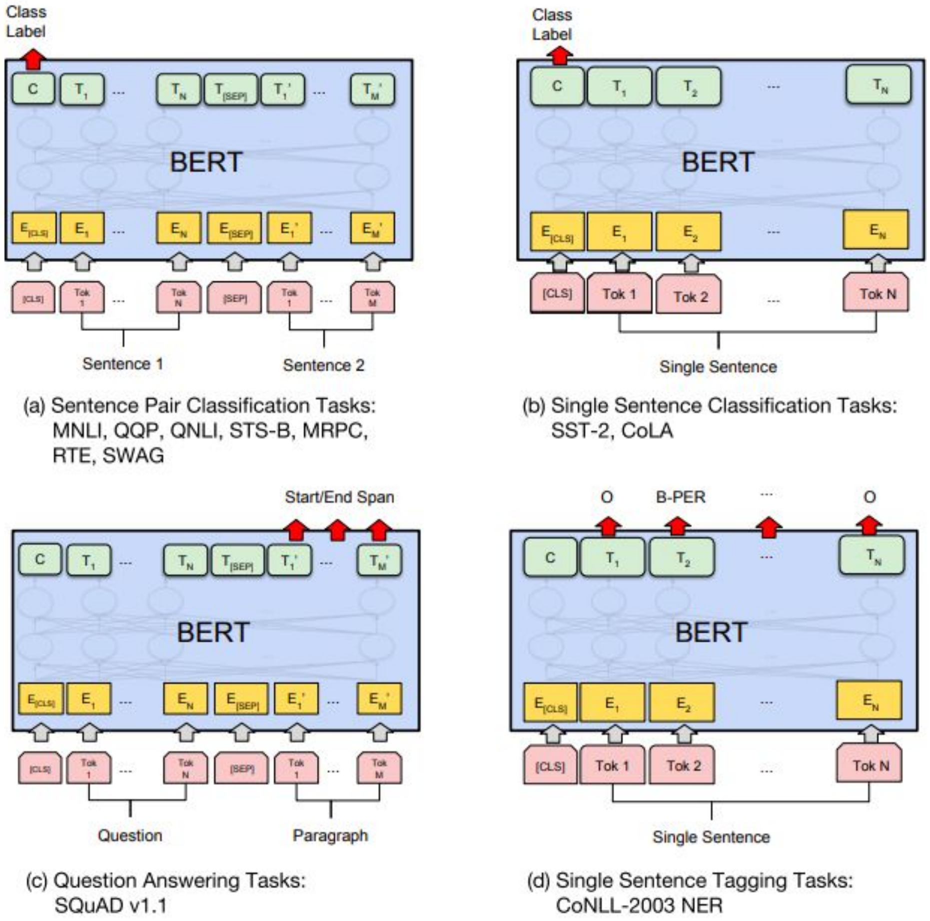


[2]: http://jalammar.github.io/illustrated-transformer/

Transformer families



https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html



[4]: https://arxiv.org/pdf/1810.04805.pdf (BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding)

Transformers: scaling

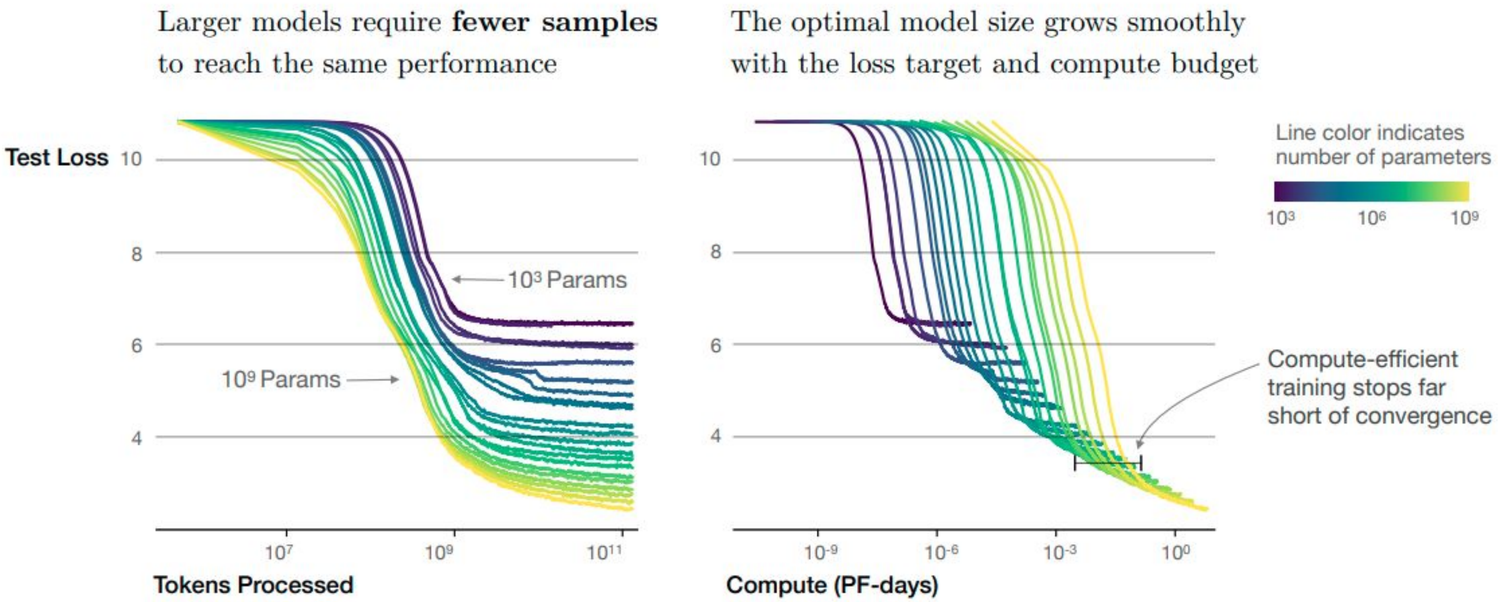
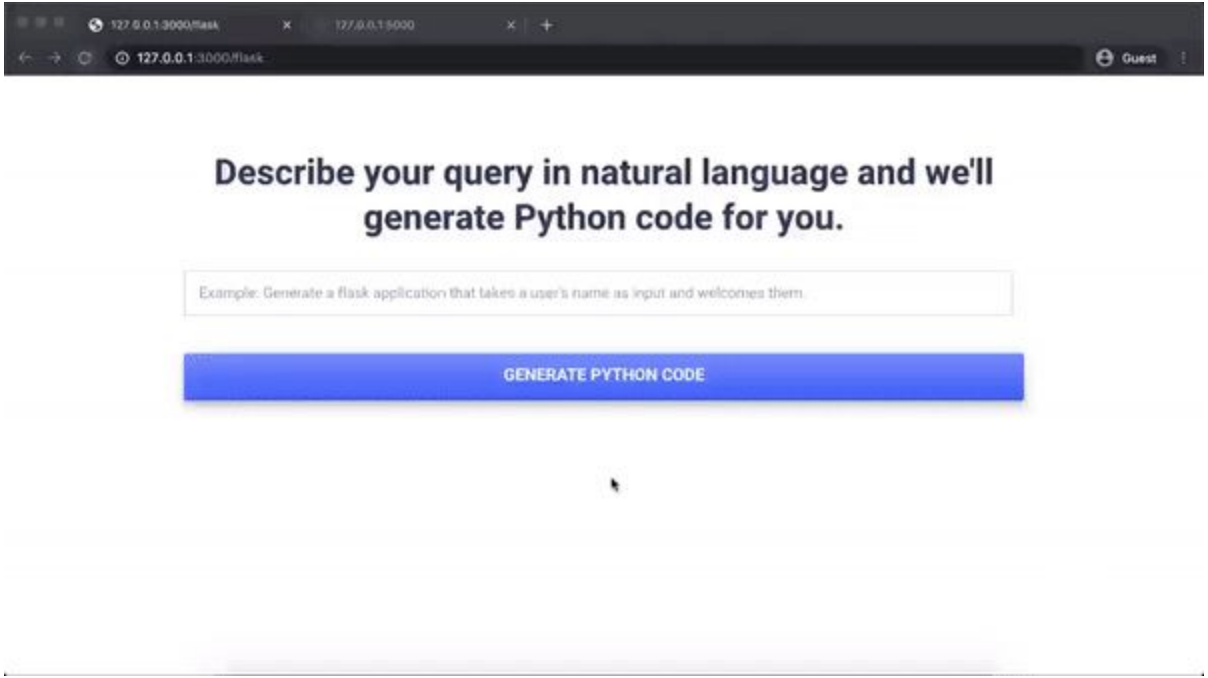


Figure 2 We show a series of language model training runs, with models ranging in size from 10^3 to 10^9 parameters (excluding embeddings).

[5]: https://arxiv.org/pdf/2001.08361.pdf (Scaling Laws for Neural Language Models, 2020)

Transformers: scaling

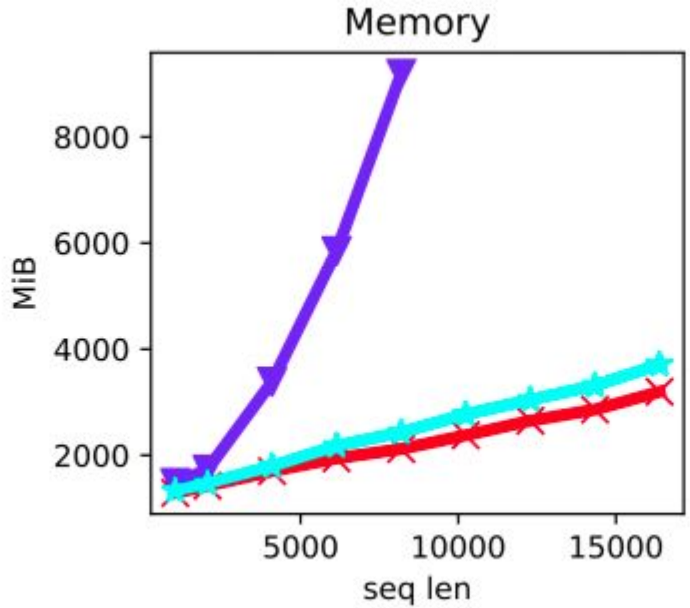


https://twitter.com/pavtalk/status/1285410751092416513

Transformers: (down)scaling

```
(base) xstefan3@michal-ideapad-520S:~$ curl -X POST "localhost:4321/translate/" --data '{"source_lang":"cs", "target_lang":"en"
, "text":"Žila jednou jedna hodná a milá dívenka. Všichni ji měli velice rádi a ze všech nejvíce maminka s babičkou. Babička jí
 ušila červený čepeček a podle něj jí začali říkat Červená Karkulka. Babička bydlela na samotě u lesa, kde široko daleko nebyla
 žádná jiná chaloupka. Babička se tam starala o lesní zvířátka. Jednou v létě maminka napekla bábovku, do košíku přidala láhev
vína a řekla Karkulce: „Babička má dneska svátek. Vezmi košík a zanes ho k babičce do chaloupky. Ale jdi rovnou, ať se v lese n
ezatouláš!"}' | jq -C
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  1145  100   530  100   615    246    286  0:00:02  0:00:02 --:--:--   533
{
  "source_lang": "cs",
  "target_lang": "en",
  "translation": "There was a nice girl who lived once. Everyone loved her very much and most of all her mother and grandmother
. Grandma made her a red hat and he said they started calling her Red Riding Hood. Grandma lived alone in the woods, where ther
e was no other cottage. Grandma took care of the forest animals. One summer, my mom baked a cake, added a bottle of wine to the
 basket and told Riding Hood: \"Grandma's holiday is tonight. Take the basket and take it to Grandma's cottage."
}
(base) xstefan3@michal-ideapad-520S:~$
```

Transformers: scaling



[6]: https://arxiv.org/pdf/2004.05150.pdf (Longformer: The Long-Document Transformer)

Attention customizations

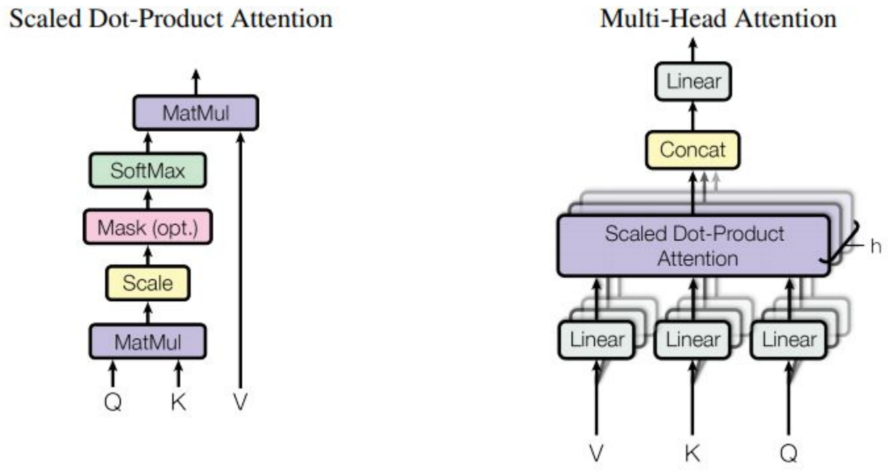
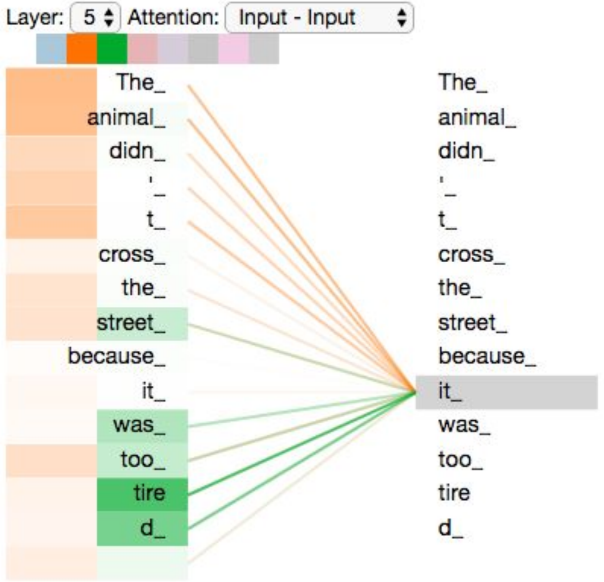


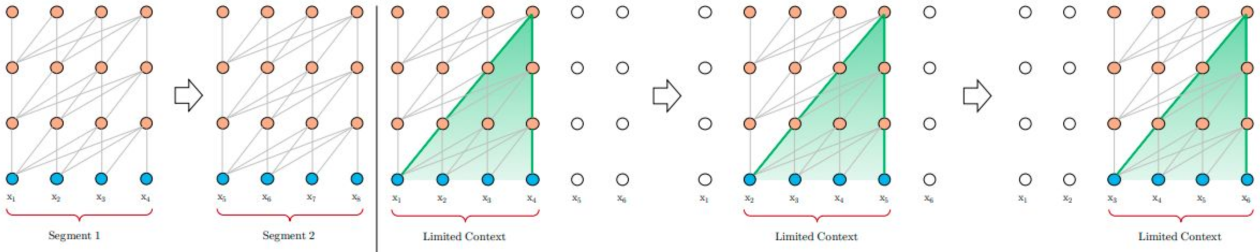
Figure 2: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.



[1]: https://arxiv.org/abs/1706.03762 (Attention is All You Need)

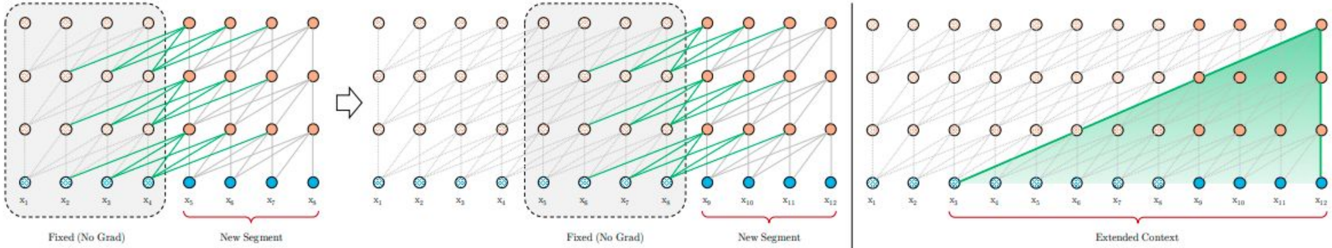
[3]: https://colab.research.google.com/github/tensorflow/tensor2tensor/blob/master/tensor2tensor/notebooks/hello_t2t.ipynb

Transformer-XL



(a) Train phase. (b) Evaluation phase.

Figure 1: Illustration of the vanilla model with a segment length 4.



(a) Training phase. (b) Evaluation phase.

Figure 2: Illustration of the Transformer-XL model with a segment length 4.

[7]: https://arxiv.org/abs/1706.03762 (Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context)

Transformer-XL [7]

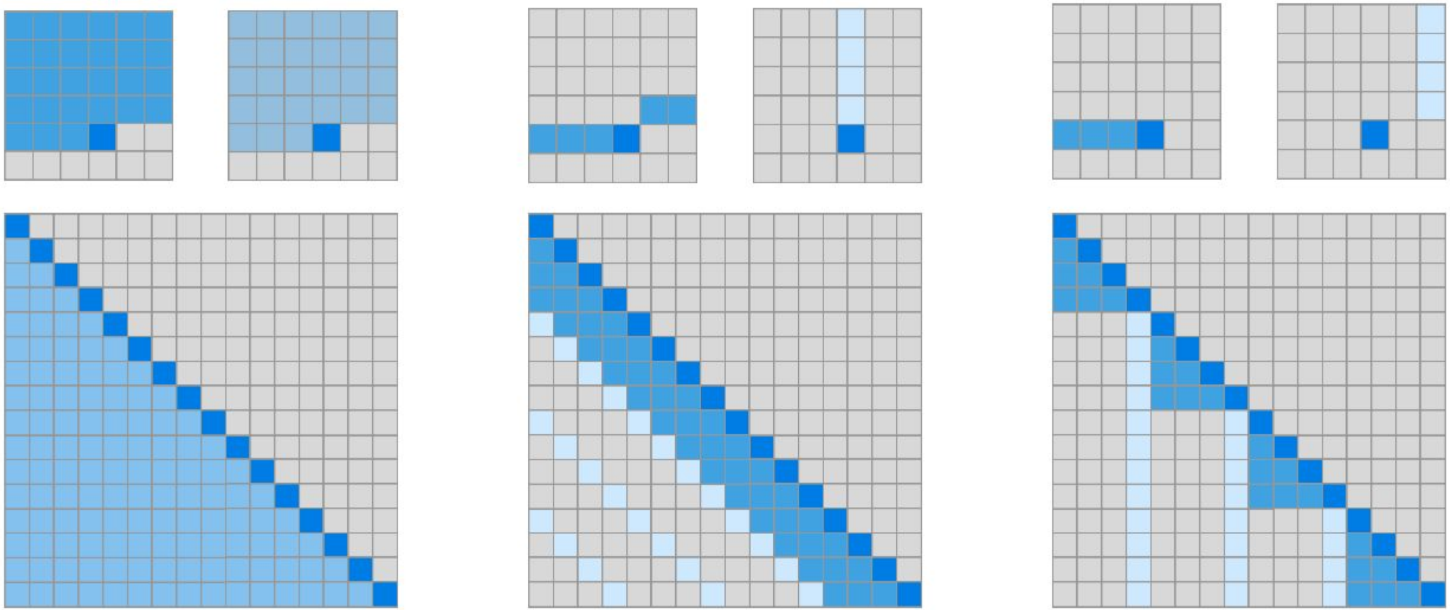
- Window length **784 on training**, **3800 on evaluation**
- Novel Relative positional encodings
- Not very relevant evaluation (bpc: Bytes-per-character)
- Not any smaller

Model	#Param	bpc
Ha et al. (2016) - LN HyperNetworks	27M	1.34
Chung et al. (2016) - LN HM-LSTM	35M	1.32
Zilly et al. (2016) - RHN	46M	1.27
Mujika et al. (2017) - FS-LSTM-4	47M	1.25
Krause et al. (2016) - Large mLSTM	46M	1.24
Knol (2017) - cmix v13	-	1.23
Al-Rfou et al. (2018) - 12L Transformer	44M	1.11
Ours - 12L Transformer-XL	41M	**1.06**
Al-Rfou et al. (2018) - 64L Transformer	235M	1.06
Ours - 18L Transformer-XL	88M	1.03
Ours - 24L Transformer-XL	277M	**0.99**

Table 2: Comparison with state-of-the-art results on enwik8.

[7]: https://arxiv.org/abs/1706.03762 (Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context)

Sparse Transformers [8]



(a) Transformer (b) Sparse Transformer (strided) (c) Sparse Transformer (fixed)

[8]: https://arxiv.org/pdf/1904.10509.pdf (Generating Long Sequences with Sparse Transformers)

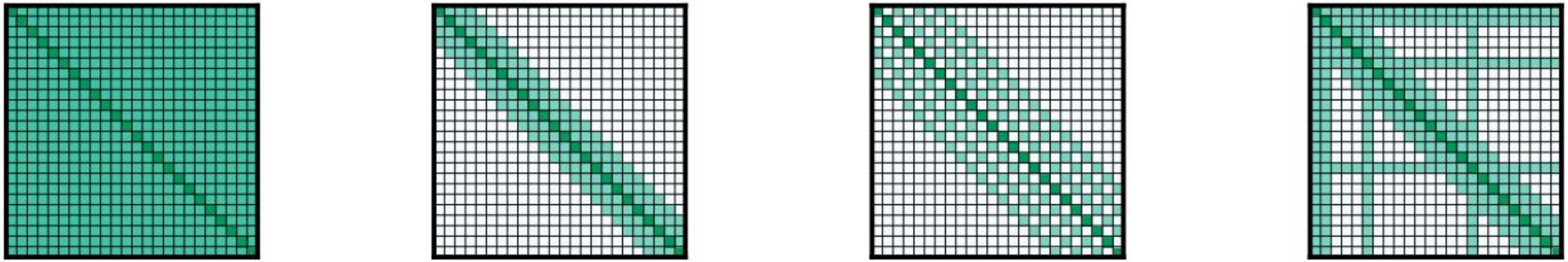
Sparse Transformers [8]

- **Head factorization** - decomposition of functionality to two heads
- **Global token positions** - first attempt to propagate information over attention
- Shows well-performing **replacement of convolution** with attention
- Evaluation on other sequences (Classical music)
- Missing evaluation on actual NLP end-tasks

Model	Bits per byte
CIFAR-10	
PixelCNN (Oord et al., 2016)	3.03
PixelCNN++ (Salimans et al., 2017)	2.92
Image Transformer (Parmar et al., 2018)	2.90
PixelSNAIL (Chen et al., 2017)	2.85
Sparse Transformer 59M (strided)	**2.80**
Enwik8	
Deeper Self-Attention (Al-Rfou et al., 2018)	1.06
Transformer-XL 88M (Dai et al., 2018)	1.03
Transformer-XL 277M (Dai et al., 2018)	**0.99**
Sparse Transformer 95M (fixed)	**0.99**
ImageNet 64x64	
PixelCNN (Oord et al., 2016)	3.57
Parallel Multiscale (Reed et al., 2017)	3.7
Glow (Kingma & Dhariwal, 2018)	3.81
SPN 150M (Menick & Kalchbrenner, 2018)	3.52
Sparse Transformer 152M (strided)	**3.44**
Classical music, 5 seconds at 12 kHz	
Sparse Transformer 152M (strided)	**1.97**

[8]: https://arxiv.org/pdf/1904.10509.pdf (Generating Long Sequences with Sparse Transformers)

Longformer [6]



(a) Full n^2 attention (b) Sliding window attention (c) Dilated sliding window (d) Global+sliding window

Figure 2: Comparing the full self-attention pattern and the configuration of attention patterns in our Longformer.

[6]: https://arxiv.org/pdf/2004.05150.pdf (Longformer: The Long-Document Transformer)

Longformer [6]

- **Linear scaling** of attention weights' size
- **Different context windows** by layers (1->12: 32->512)
- New idea of **Dilation**: attend to every second position, on 2 bottom layers
- Transfer of existing RoBERTa weights
- Evaluation on **end tasks** (requiring long context window)
- Humble **ablation study**

Model	WikiHop	TriviaQA
Current SOTA	78.3	73.3
Longformer-large	**81.9**	**77.3**

Table 9: Leaderboard results of Longformer-large

Model	Accuracy / Δ
Longformer (seqlen: 4,096)	73.8
RoBERTa-base (seqlen: 512)	72.4 / -1.4
Longformer (seqlen: 4,096, 15 epochs)	75.0 / +1.2
Longformer (seqlen: 512, attention: n^2)	71.7 / -2.1
Longformer (seqlen: 512, attention: window)	68.8 / -5.0
Longformer (seqlen: 2,048)	73.1 / -0.7
Longformer (no MLM pretraining)	73.2 / -0.6
Longformer (no linear proj.)	72.2 / -1.6
Longformer (no linear proj. no global atten.)	65.5 / -8.3

Table 11: WikiHop development set ablations

Model	#Param	Test BPC
Transformer-XL (18 layers)	88M	1.03
Sparse (Child et al., 2019)	≈100M	0.99
Transformer-XL (24 layers)	277M	0.99
Adaptive (Sukhbaatar et al., 2019)	209M	0.98
Compressive (Rae et al., 2020)	277M	0.97
Our Longformer	102M	0.99

Table 3: Performance of *large* models on `enwik8`

[6]: https://arxiv.org/pdf/2004.05150.pdf (Longformer: The Long-Document Transformer)

Big Bird [8]

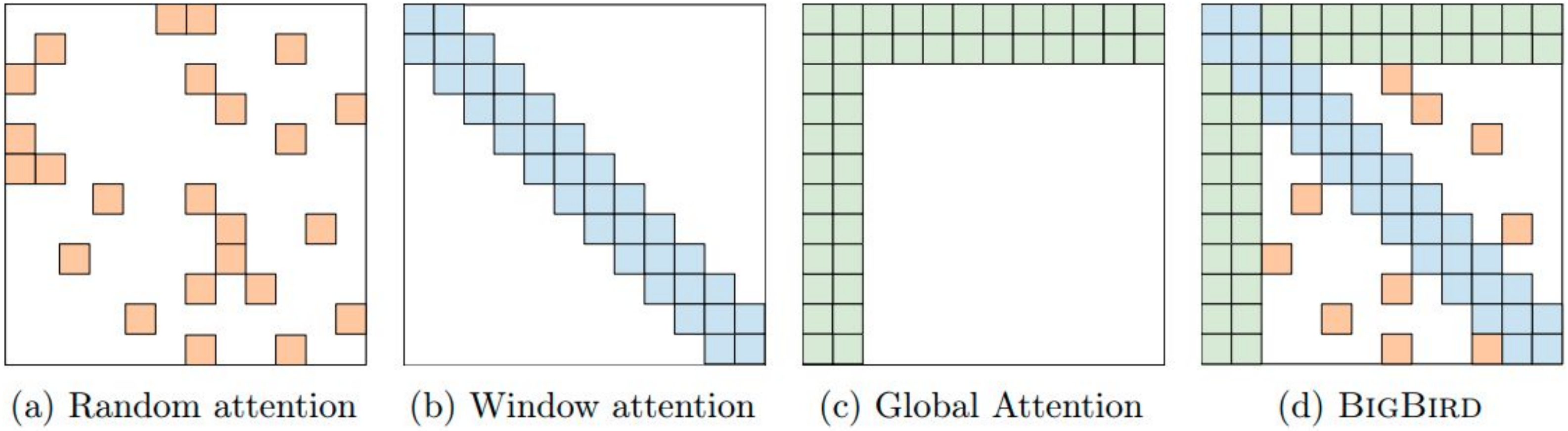
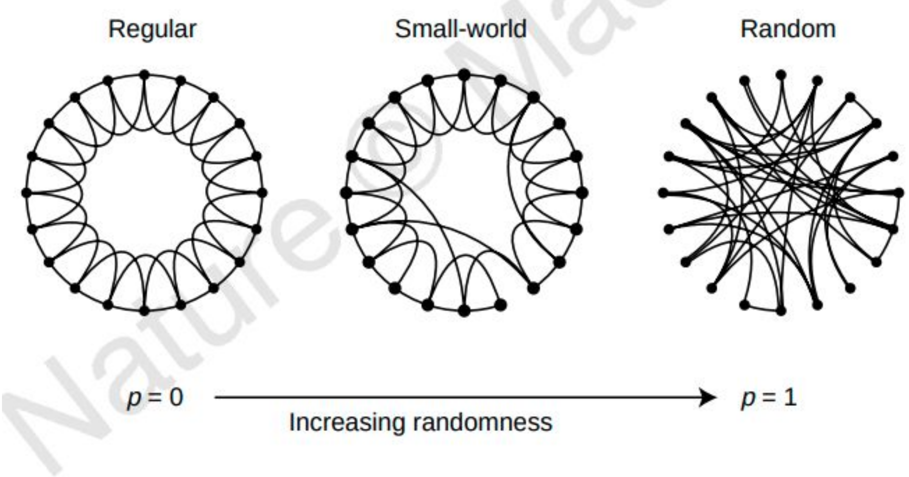


Figure 1: Building blocks of the attention mechanism used in BIGBIRD. White color indicates absence of attention. (a) random attention with $r = 2$, (b) sliding window attention with $w = 3$ (c) global attention with $g = 2$. (d) the combined BIGBIRD model.

[8]: https://arxiv.org/pdf/2007.14062v1.pdf (Big Bird: Transformers for Longer Sequences)

Big Bird [8]

- **Random graph in attention**: model of random positions selection is rationalized by information propagation, reasoned in [9]
- Randomness actually seem to work: ablation study shows +3-5% acc superiority to Longformer (that is a lot)



[8]: https://arxiv.org/pdf/2007.14062v1.pdf (Big Bird: Transformers for Longer Sequences)

[9]: collective-dynamics-of-small-world-networks.pdf (Collective dynamics of ‘small-world’ networks)

Big Bird [8]

- Compared to Longformer, it only **adds random connections**
- It reasons it by **minimizing the distance** of between each pair of nodes = tokens
- Some nice theoretical properties: with random attention heads, Big Bird is ***Universal Approximator*** of any seq2seq function on its context window (like full Transformer)
- Turing complete
- Serious **evaluation** on “long” end tasks (not just bpc) and also some “short” tasks
- Possible cheating by pre-training with Pegasus objective

Model	IMDb [65]	Yelp-5 [108]	Arxiv [36]	Patents [54]	Hyperpartisan [48]
# 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	[70] 87.96	[70] 69.01	[41] 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	$\mathbf{92.2} \pm \mathbf{1.7}$

Table 6: 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.

[8]: https://arxiv.org/pdf/2007.14062v1.pdf (Big Bird: Transformers for Longer Sequences)

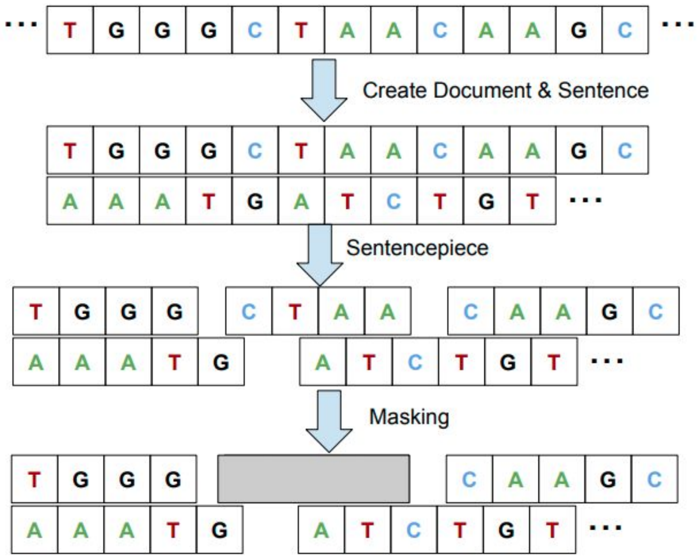


Figure 2: Visual description of how the masked language modeling data was generated from raw DNA dataset. The raw DNA sequences of GRCh37, where split at random positions to create documents with 50-100 sentences where each sentence was 500-1000 base pairs (bps). Thus each document had a continuous strand of 25000-100,000 bps of DNA. This process was repeated 10 times to create 10 sets of document for each chromosome of GRCH37. The resulting set of documents was then passed through Sentencepiece that created tokens of average 8bp. For pretraining we used masked language model and masked 10% of the tokens and trained on predicting the masked tokens.

Thanks!

Feel free to check out our theses:

https://is.muni.cz/auth/rozpis/tema tag MIR

or contact us later!

Michal Štefánik

stefanik.m@mail.muni.cz