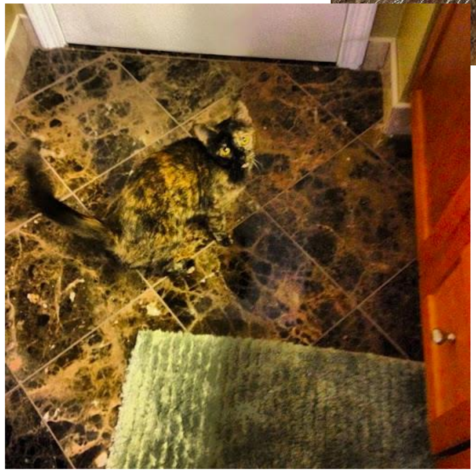
Attention, Transformers and Indirection

Neural Networks

CSCI 4850/5850

High-Dimensional Space – Distractions

Source: thiscatdoesnotexist.com



Source: izzcat.com

Source: reddit.com

Source: cnet.com

Competing Features – Distractions

Source: thiscatdoesnotexist.com

Source: amazon.com

Temporal Relations – Distractions

What is the subject of this poem?

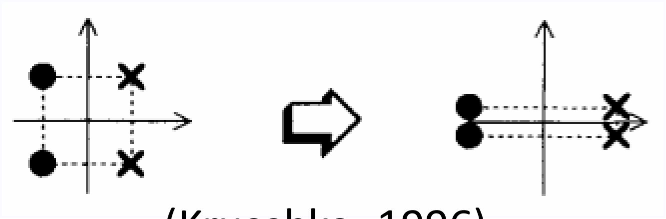
Every morning, suit,
you are waiting on a chair
to be filled
by my vanity, my love,
my hope, my body.

- Excerpt from 'Ode to my suit'
    - Author: Pablo Neruda
    - Translator: Margaret Peden

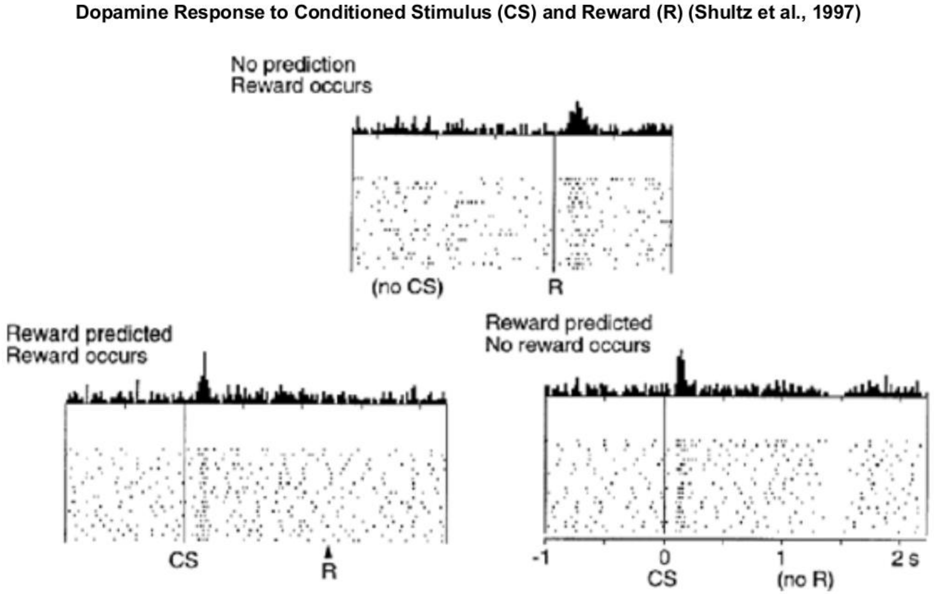
Source: http://www.reedfurnituredesign.com/blog/2015/1/valet-chair

Traditional Attention Mechanisms

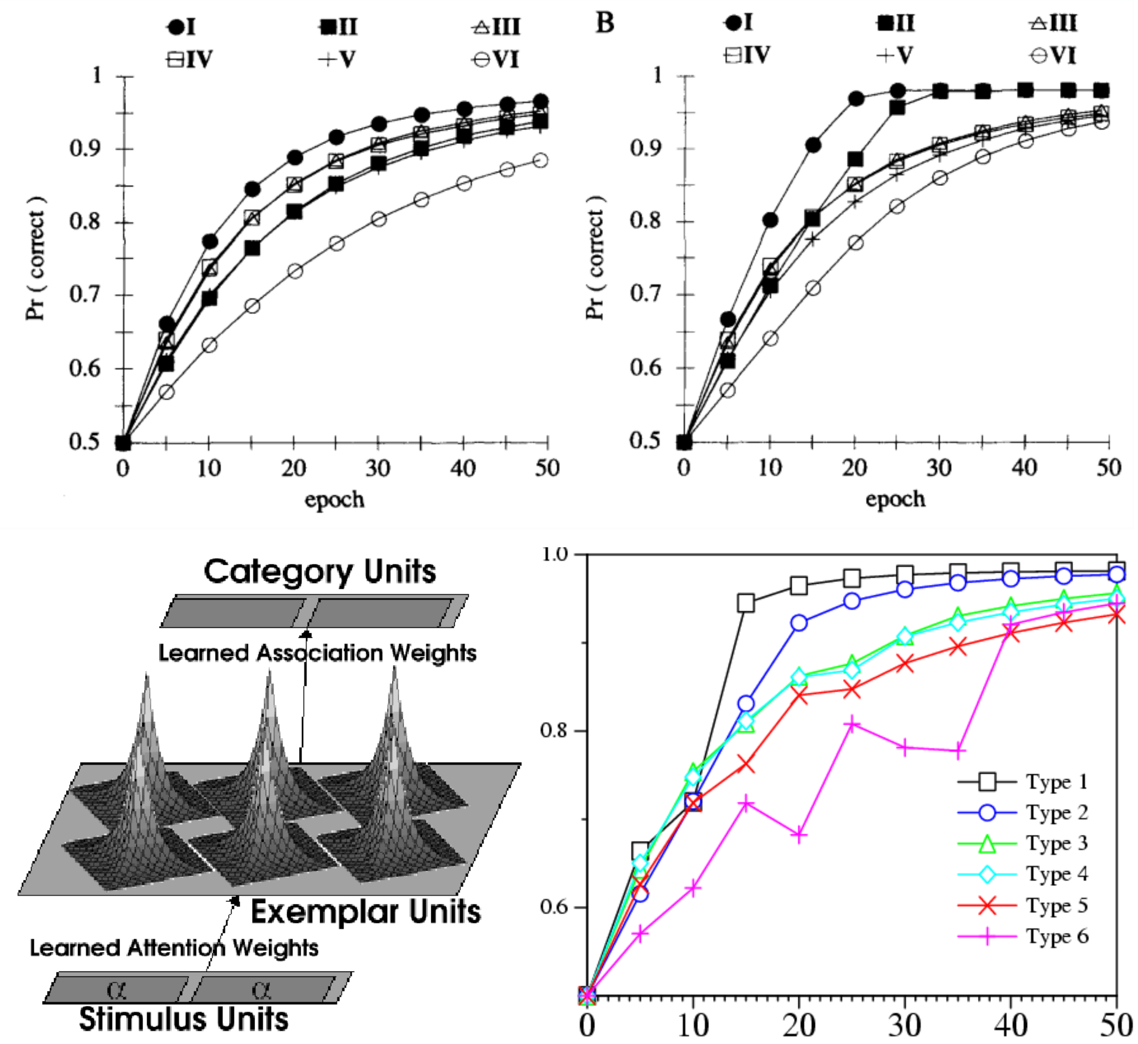
Dimensional Attention



(Kruschke, 1996)



(Schultz, 1999)



(Phillips & Noelle, 2004)

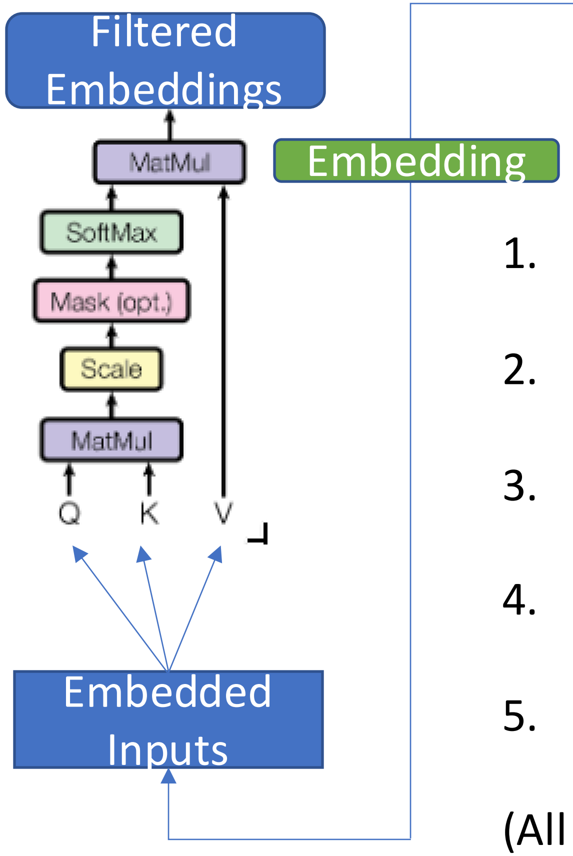
Attention – Filtering Out the Irrelevant

How do we **learn** which features are relevant?

What information in the image **guides our attention**?

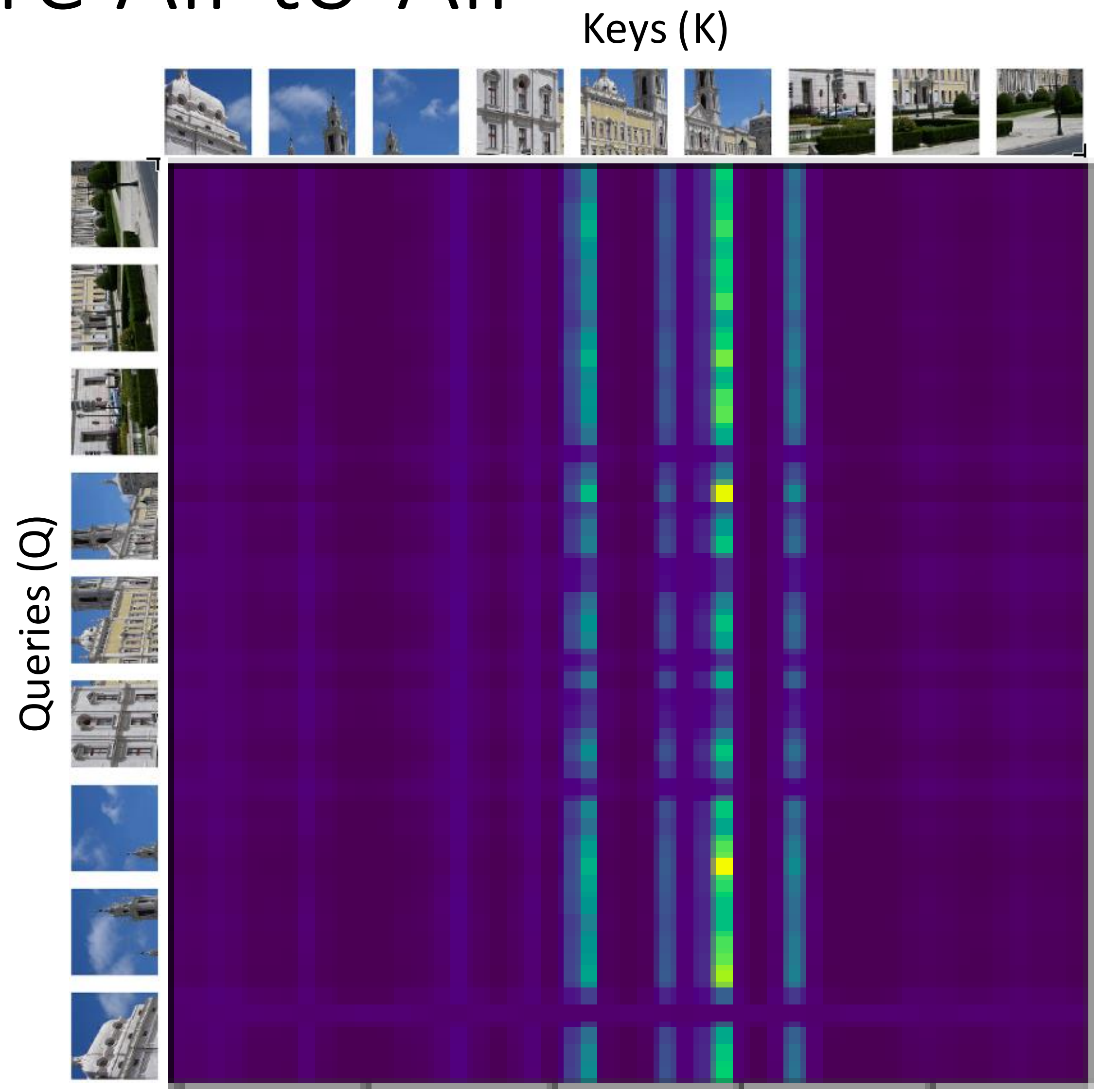
Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

Self-Attention – Compare All-to-All



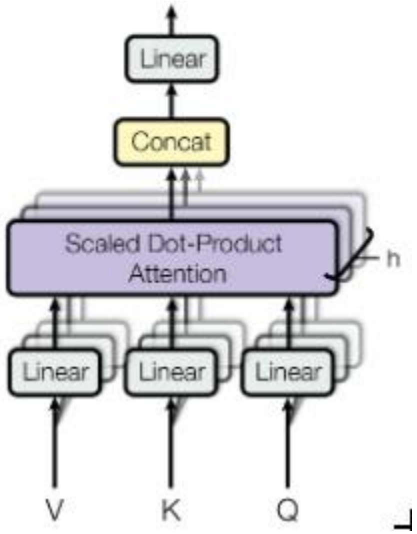
1. Embeddings can be linear, CCNs, etc.
2. Split into Queries, Keys, Values
3. All Queries are compared to All Keys
4. Matches allow Values to pass forward
5. Non-matching Values are filtered out

(All choices on what to **self-attend** to using Q and K are learned based on **feedback**)



Vaswani, Ashish, et al. "Attention is all you need." arXiv preprint arXiv:1706.03762 (2017).

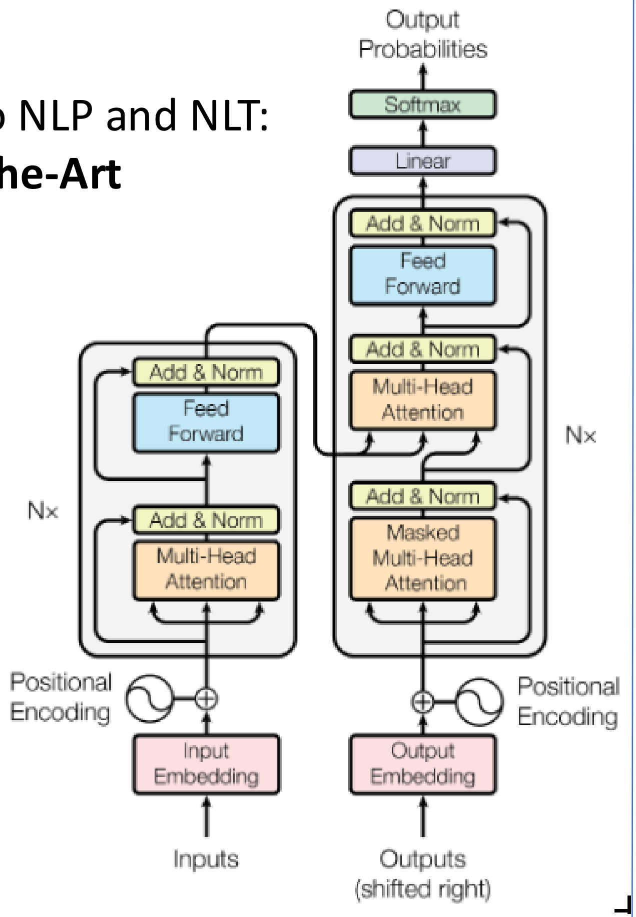
Multiple Filters? – Multihead Attention



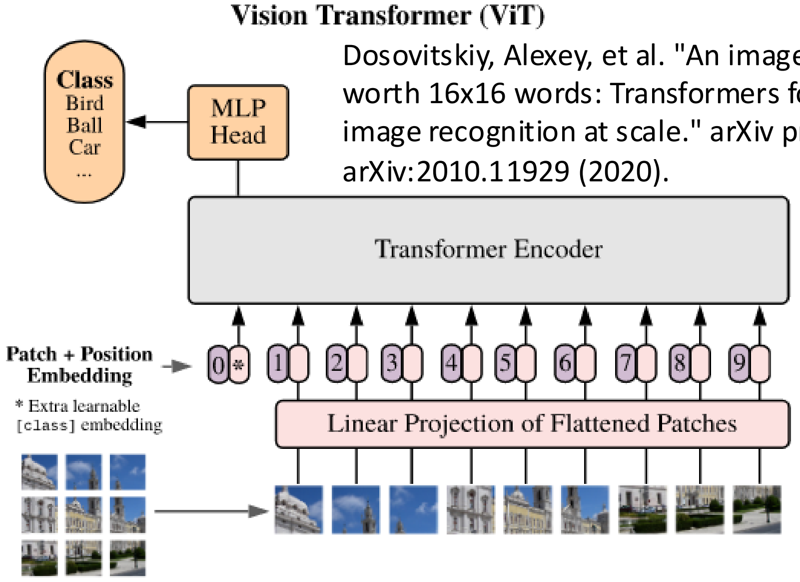
Transform the input several different ways (Linear) and then apply self-attention: results are combined together (Concat + Linear) for a **more robust embedding** capable of **focusing attention on several types of features**.

NOTE: No recurrent layers…

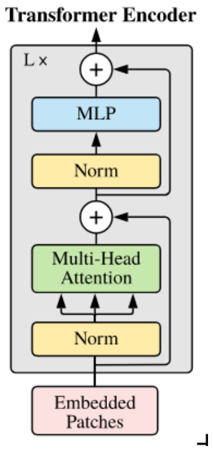
Applied to NLP and NLT:
State-of-the-Art



Vaswani, Ashish, et al. "Attention is all you need." arXiv preprint arXiv:1706.03762 (2017).



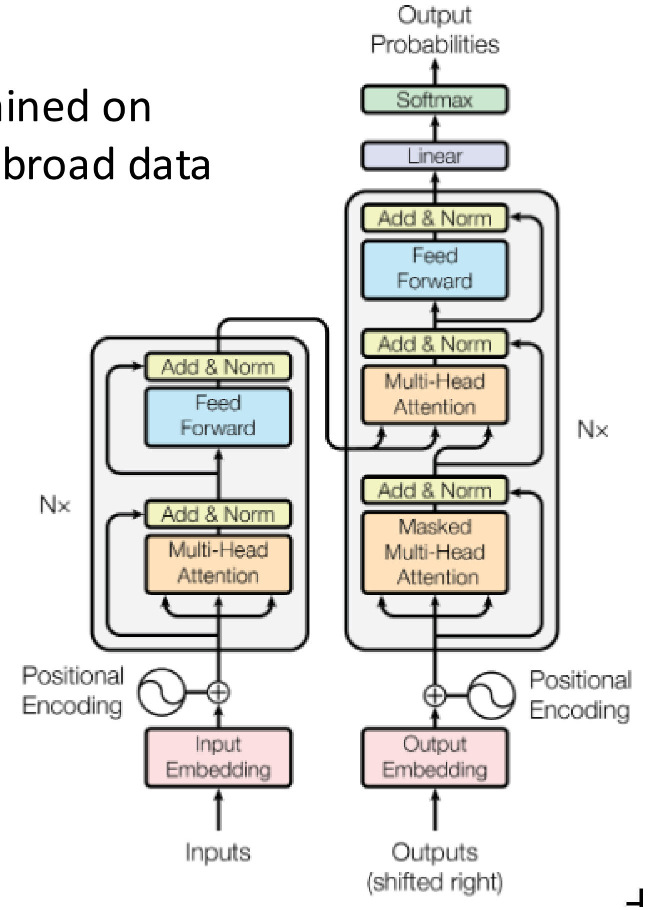
Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).



Applied to Image Processing and Classification:
State-of-the-Art

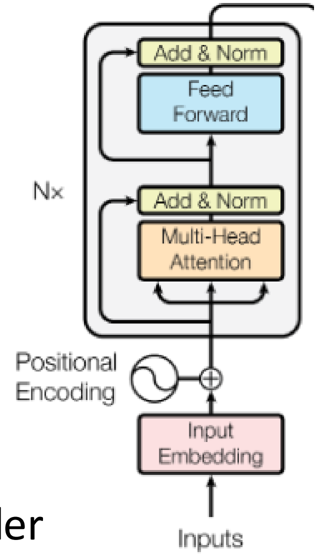
Best Case Scenario – Transfer Learning

Pretrained on huge/broad data sets



Remove decoder

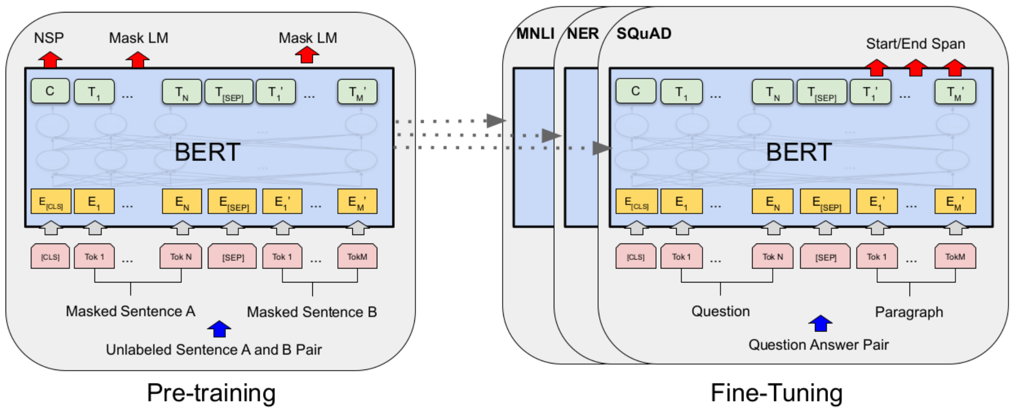
Pretrained embedding model

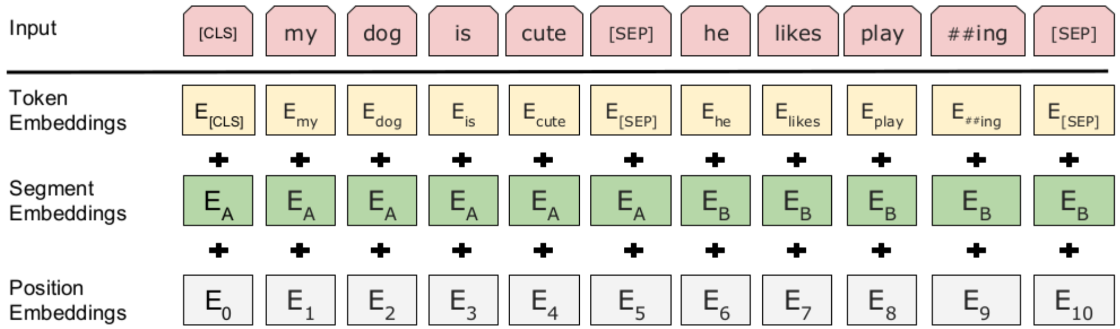


NLP Examples:

BERT, mBERT, RoBERTa

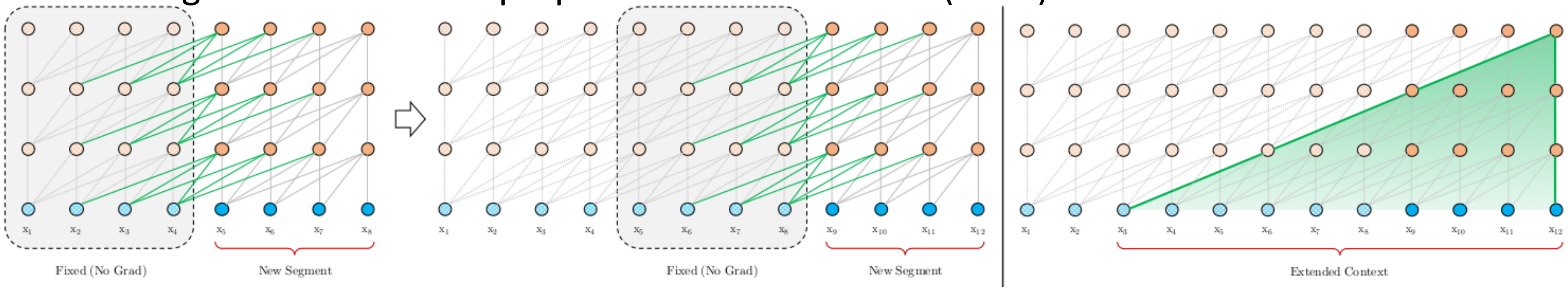
Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).





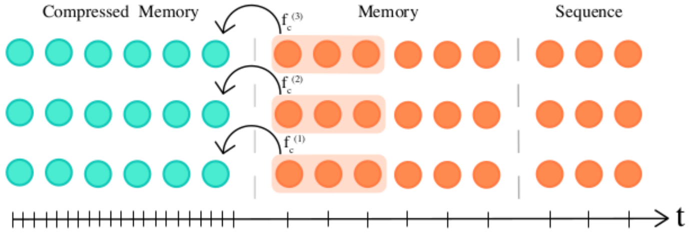
Arbitrary Sequence Lengths?

Dai, Zihang, et al. "Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context." arXiv preprint arXiv:1901.02860 (2019).

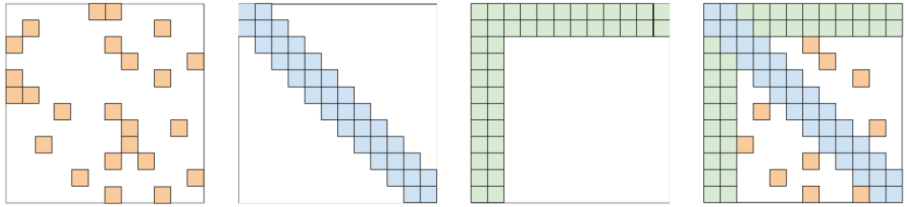


(a) Training phase. (b) Evaluation phase.

Utilizes **relative position encodings** to allow arbitrary time-scales for position tokens

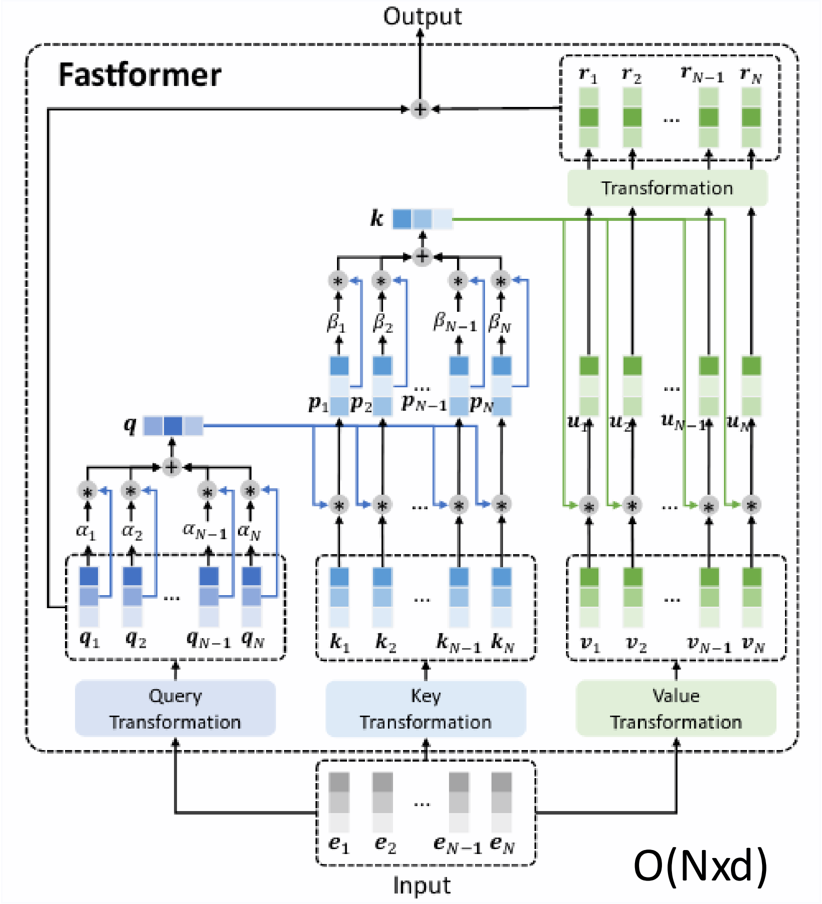


Rae, Jack, et al. "Compressive Transformers for Long-Range Sequence Modelling." arXiv preprint arXiv:1911.05507 (2019).


(a) Random attention (b) Window attention (c) Global Attention (d) BIGBIRD

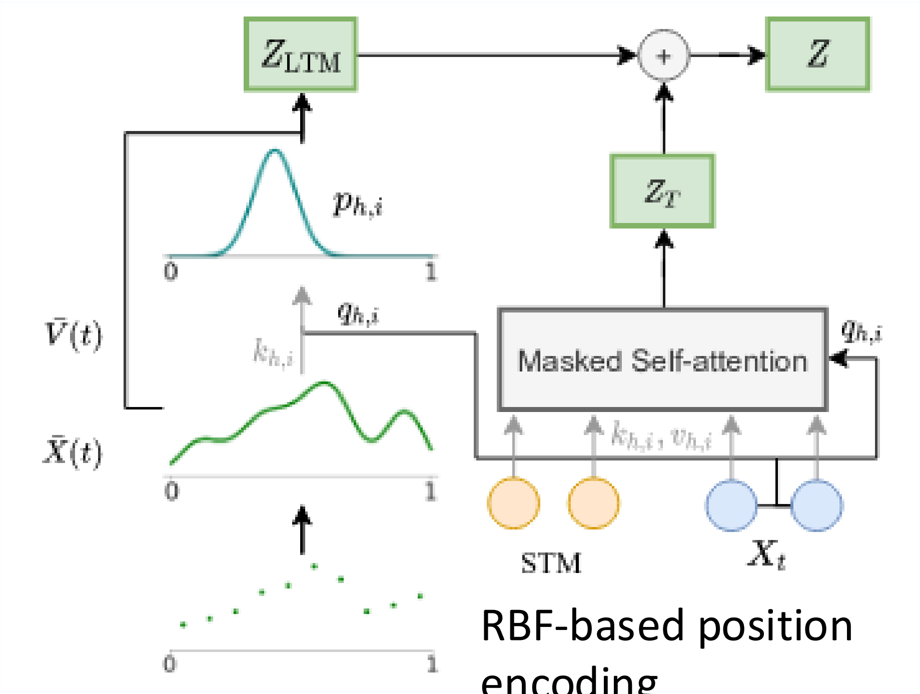
Zaheer, Manzil, et al. "Big Bird: Transformers for Longer Sequences." arXiv preprint arXiv:2007.14062 (2021).

Speed/Memory Issues?



O(Nxd)

Wu et al. “Fastformer: Additive Attention Can Be All You Need” (2021). https://arxiv.org/abs/2108.09084



RBF-based position encoding

RBFs determines complexity and is independent of sequence length

Martins, Marinho, and Martins “∞-former: Infinite Memory Transformer” (2021). https://arxiv.org/abs/2109.00301

Overcoming Existing Context Limits

Model	#Param	PPL
Grave et al. (2016b) - LSTM	-	48.7
Bai et al. (2018) - TCN	-	45.2
Dauphin et al. (2016) - GCNN-8	-	44.9
Grave et al. (2016b) - LSTM + Neural cache	-	40.8
Dauphin et al. (2016) - GCNN-14	-	37.2
Merity et al. (2018) - QRNN	151M	33.0
Rae et al. (2018) - Hebbian + Cache	-	29.9
Ours - Transformer-XL Standard	151M	**24.0**
Baevski and Auli (2018) - Adaptive Input[◇]	247M	20.5
Ours - Transformer-XL Large	257M	**18.3**

Table 1: Comparison with state-of-the-art results on WikiText-103. ◇ indicates contemporary work.

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.

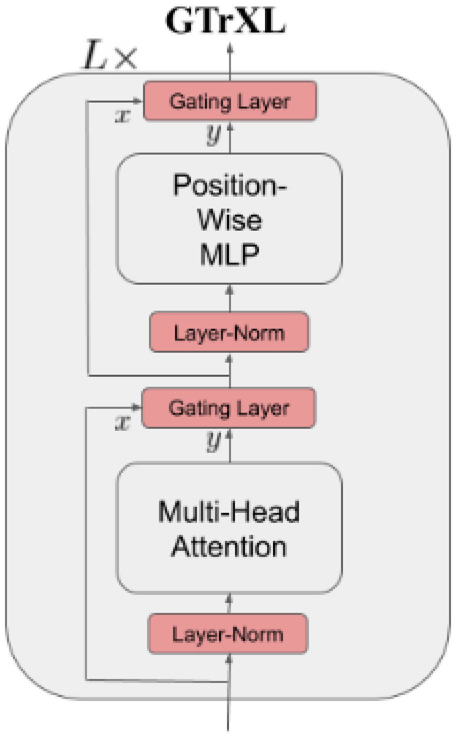
Model	#Param	bpc
Cooijmans et al. (2016) - BN-LSTM	-	1.36
Chung et al. (2016) - LN HM-LSTM	35M	1.29
Zilly et al. (2016) - RHN	45M	1.27
Krause et al. (2016) - Large mLSTM	45M	1.27
Al-Rfou et al. (2018) - 12L Transformer	44M	1.18
Al-Rfou et al. (2018) - 64L Transformer	235M	1.13
Ours - 24L Transformer-XL	277M	**1.08**

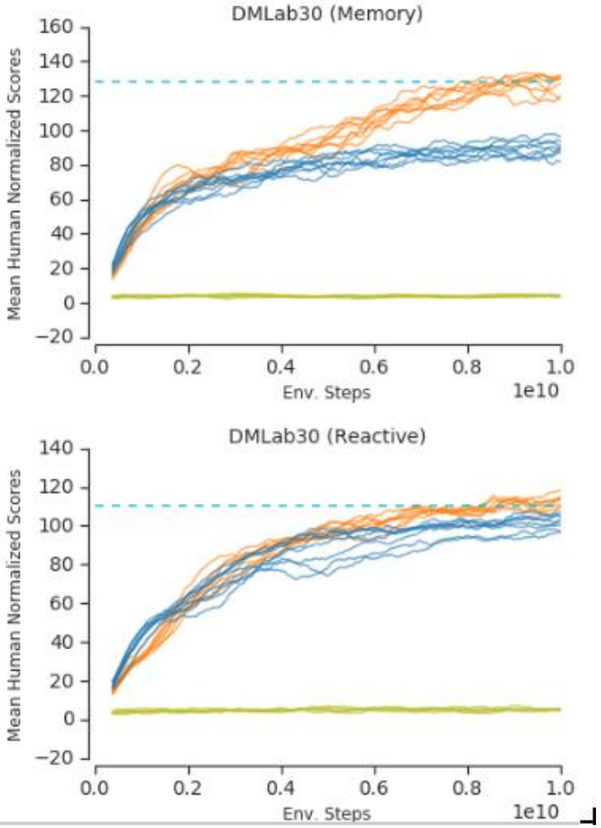
Table 3: Comparison with state-of-the-art results on text8.

Model	#Param	PPL
Shazeer et al. (2014) - Sparse Non-Negative	33B	52.9
Chelba et al. (2013) - RNN-1024 + 9 Gram	20B	51.3
Kuchaiev and Ginsburg (2017) - G-LSTM-2	-	36.0
Dauphin et al. (2016) - GCNN-14 bottleneck	-	31.9
Jozefowicz et al. (2016) - LSTM	1.8B	30.6
Jozefowicz et al. (2016) - LSTM + CNN Input	1.04B	30.0
Shazeer et al. (2017) - Low-Budget MoE	~5B	34.1
Shazeer et al. (2017) - High-Budget MoE	~5B	28.0
Shazeer et al. (2018) - Mesh Tensorflow	4.9B	24.0
Baevski and Auli (2018) - Adaptive Input[◇]	0.46B	24.1
Baevski and Auli (2018) - Adaptive Input[◇]	1.0B	23.7
Ours - Transformer-XL Base	0.46B	23.5
Ours - Transformer-XL Large	0.8B	**21.8**

Table 4: Comparison with state-of-the-art results on One Billion Word. ◇ indicates contemporary work.

Dai, Zihang, et al. "Transformer-xl: Attentive language models beyond a fixed-length context." arXiv preprint arXiv:1901.02860 (2019).

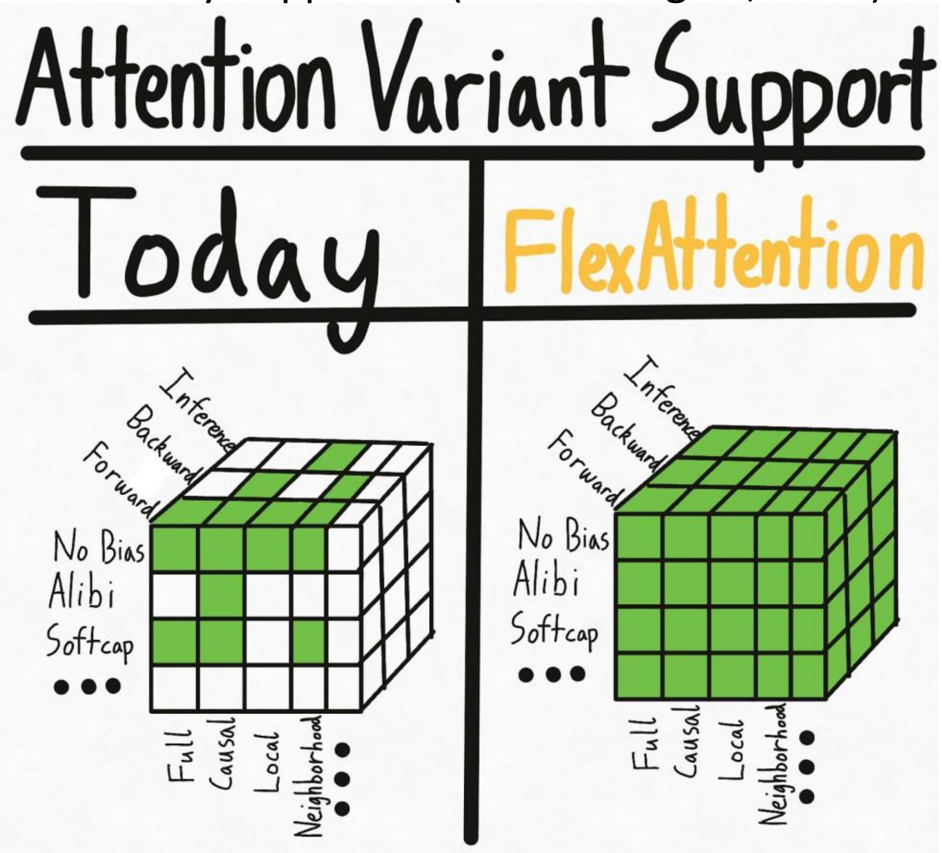




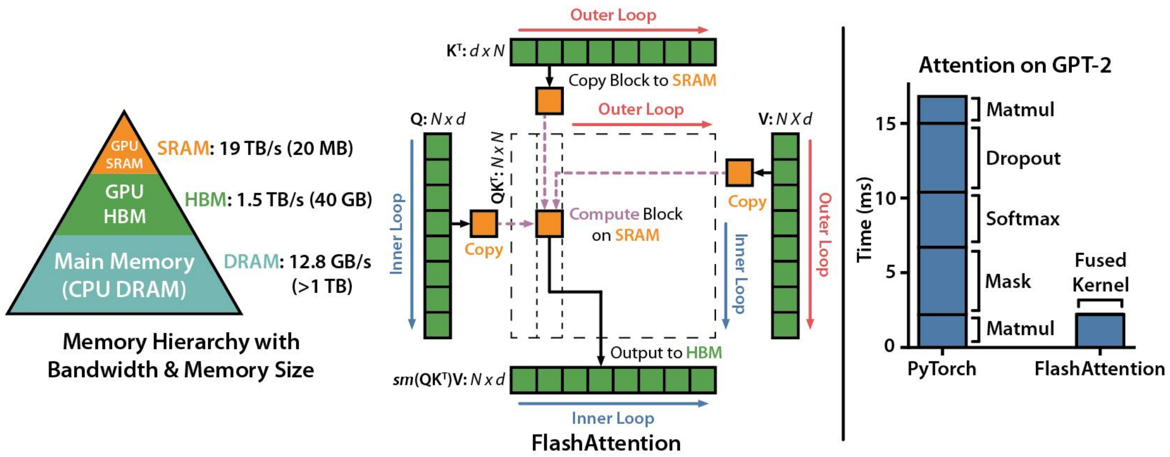
Parisotto, Emilio, et al. "Stabilizing transformers for reinforcement learning." International Conference on Machine Learning. PMLR, 2020.

Current SOA

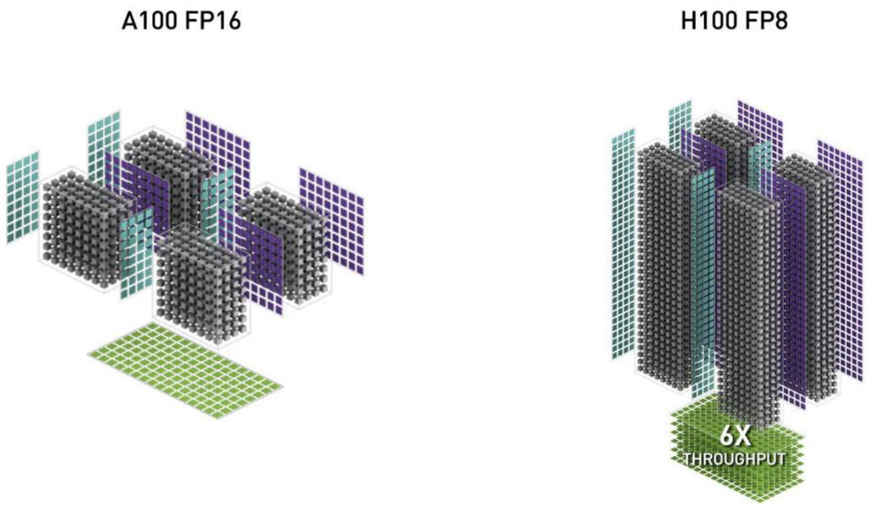
Currently Supported (started August, 2024)



https://pytorch.org/blog/flexattention/



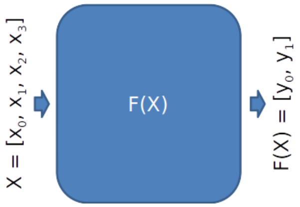
Dao et al., 2022 - https://arxiv.org/abs/2205.14135
https://github.com/Dao-AILab/flash-attention
Already up to version 3 – as of July 2024
(only supports specific hardware)



Unsupervised Learning: Generative Pretrained Transformer (GPT)

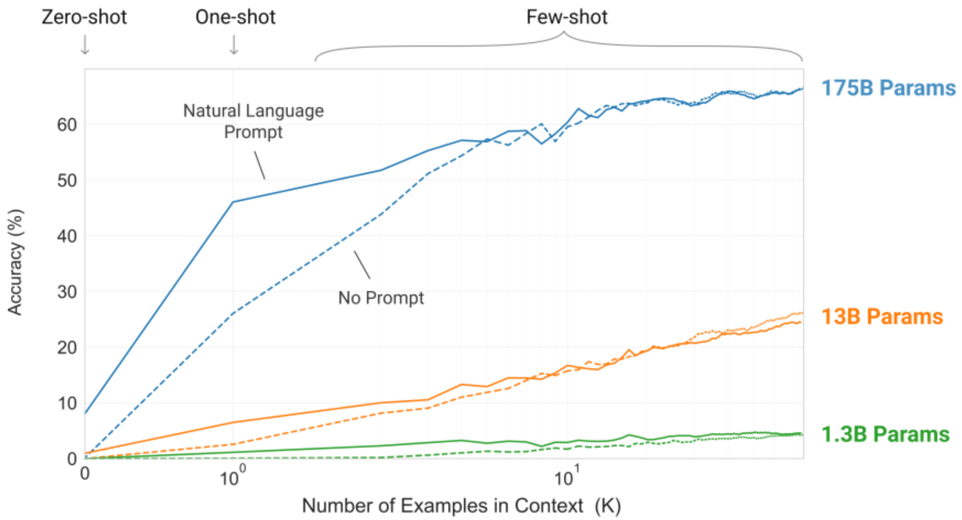
- Vaswani et al., 2017 - **Transformer** architecture
- Radford et al., 2018 and Brown et al., 2020
- Simple *generative training* and *testing* procedure, perfectly suited for the *transformer* architecture.
- Very large model, very large data set

1. The
2. Cat
3. Ran
4. Fast

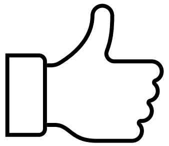


1. Cat
2. Ran
3. Fast
4. <STOP>

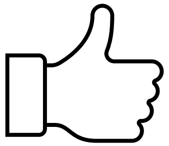
The [P(duck), P(cat), P(fast), P(no), ...]
The cat [P(duck), P(cat), P(ran), ...]
The cat ran [P(fast), P(quickly), P(slowly), P(no) ...]



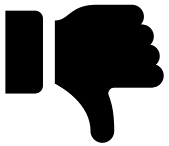
GPT-3 (Brown et al. 2020)



[To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:]
One day when I was playing tag with my little sister, she got really excited and she started doing these crazy farduddles.

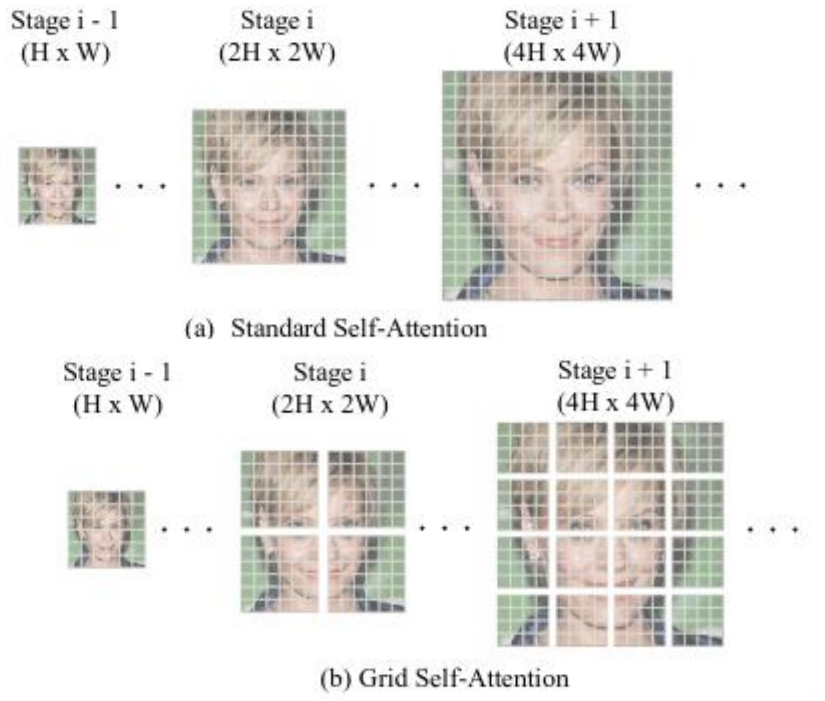


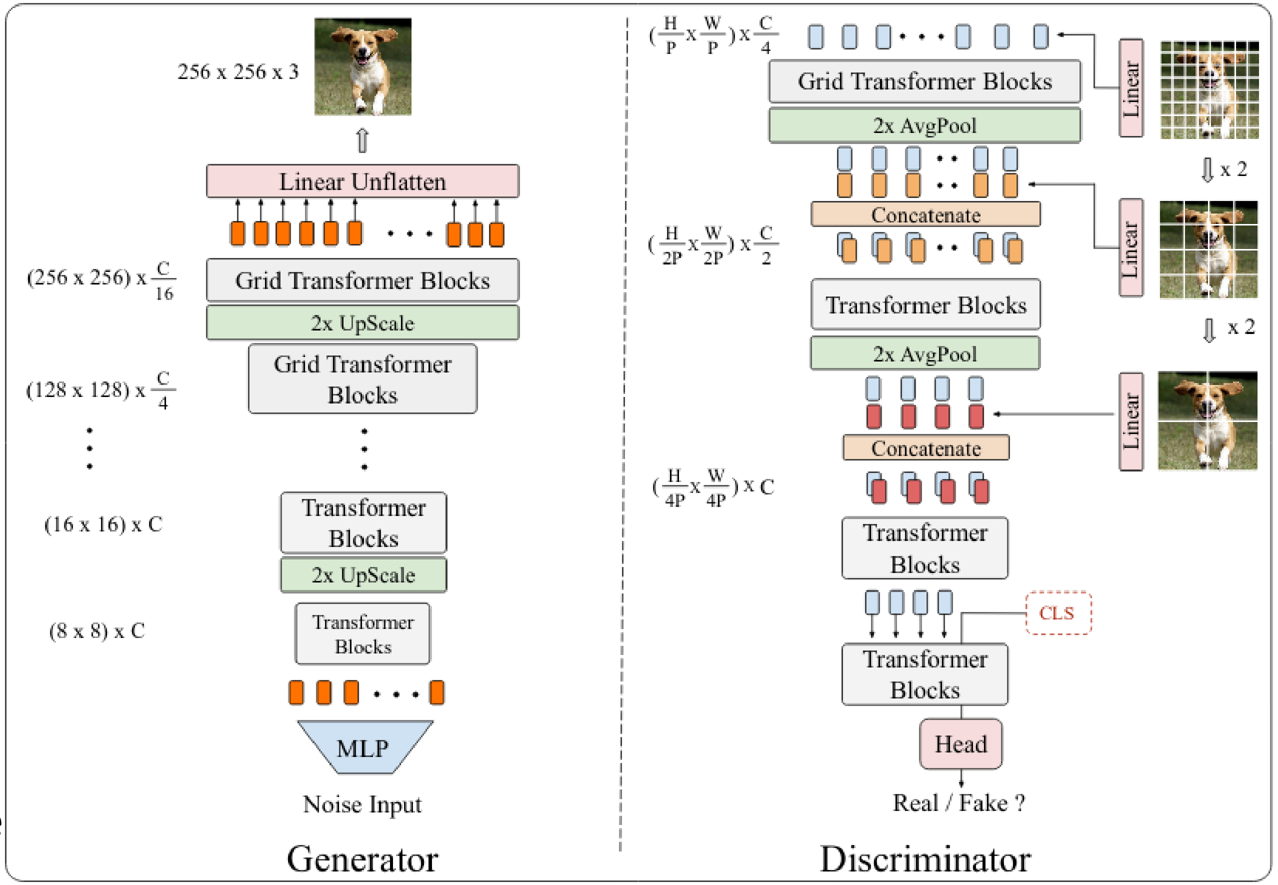
[A "yalubalu" is a type of vegetable that looks like a big pumpkin. An example of a sentence that uses the word yalubalu is:]
I was on a trip to Africa and I tried this yalubalu vegetable that was grown in a garden there. It was delicious.



[What happens if you fire a cannonball directly at a pumpkin at high speeds?]
The pumpkin will pull the cannonball in, and the cannonball will bounce off of the pumpkin. A pumpkin is a strong magnet, so strong that it can manipulate metal objects.

Generative Models

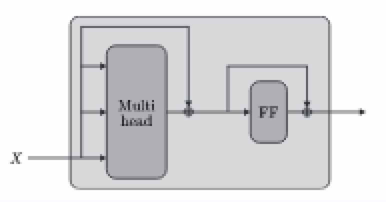




- Expanded the idea of patch-based structure to **multiscale patches**
- GANs (both **generator and discriminatory**) constructed purely from residual transformer blocks

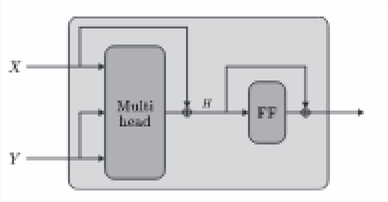
Jiang, Chang, and Wang. "TransGAN: Two Pure Transformers Can Make One Strong GAN, and That Can Scale Up" (under review) **State-of-theArt** https://arxiv.org/abs/2102.07074

Unstructured Data (Sets)

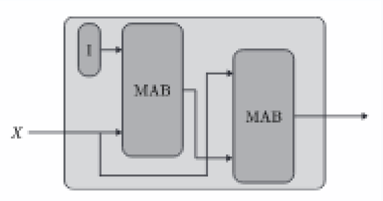


NxN

Set Attention Block



Multihead Attention Block (MAB)

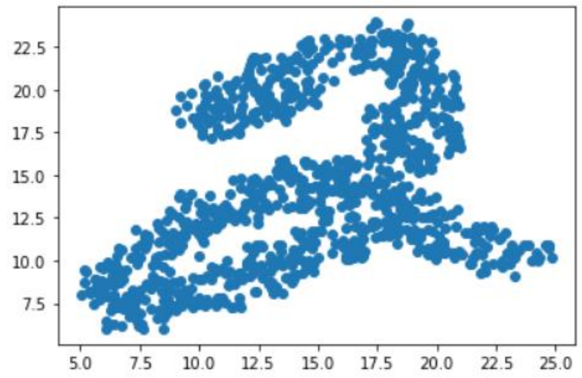


LxN

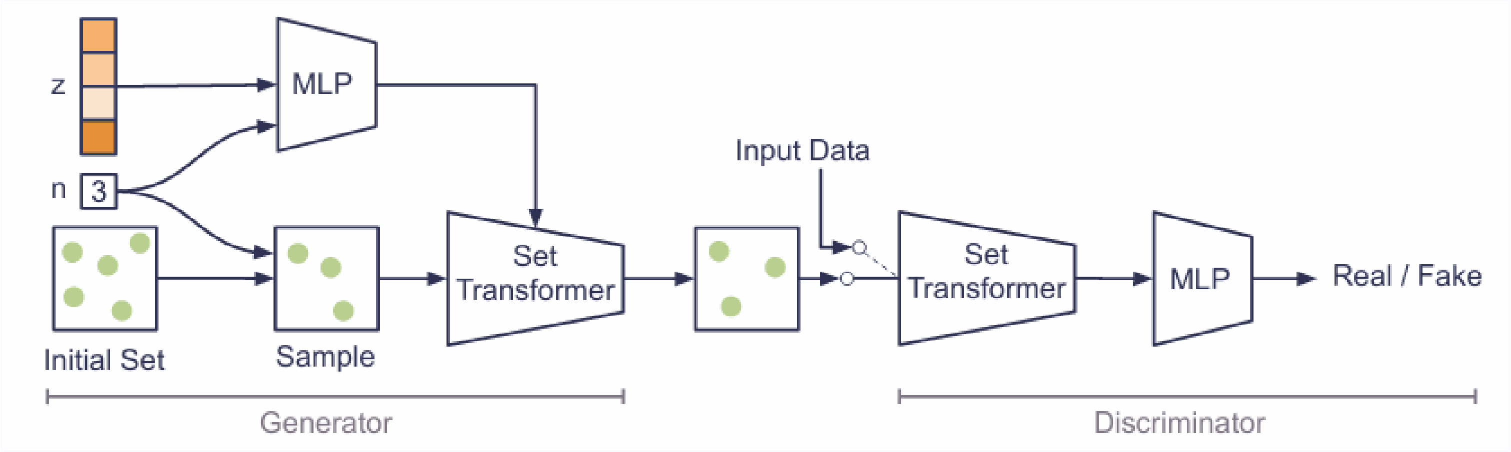
Induced Set Attention Block

Lee et al. “Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks” *ICML* (2019). https://arxiv.org/abs/1810.00825

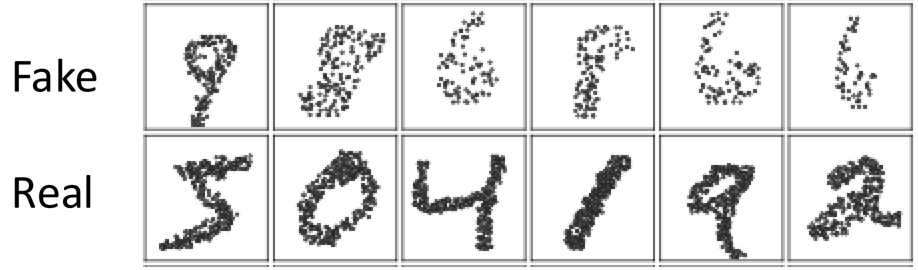
- Set-based Data?
  - Groups of points
    - (general embeddings?)
  - Groups of images
  - Groups of sentences (captioning)
- Problems
  - Traditional approaches cannot produce *permutation equivariant* representations
    - Changing the order of the inputs impacts which weights are used and therefore how the data is encoded
  - Traditional loss functions are not *permutation invariant*
    - Changing the order of the output results in a matching problem (which permutation is it?)
- Solutions
  - Transformers
    - Remove position encodings and they naturally produce permutation equivariant transforms
    - No matter that order the input data is presented in, it’s encoded in the same manner
  - Hungarian Loss and Chamfer Loss
    - More computationally expensive, but expressive loss functions
    - H is O(n^3) and C is O(n^2) – approximation trade-off



Generative Models (for Sets)

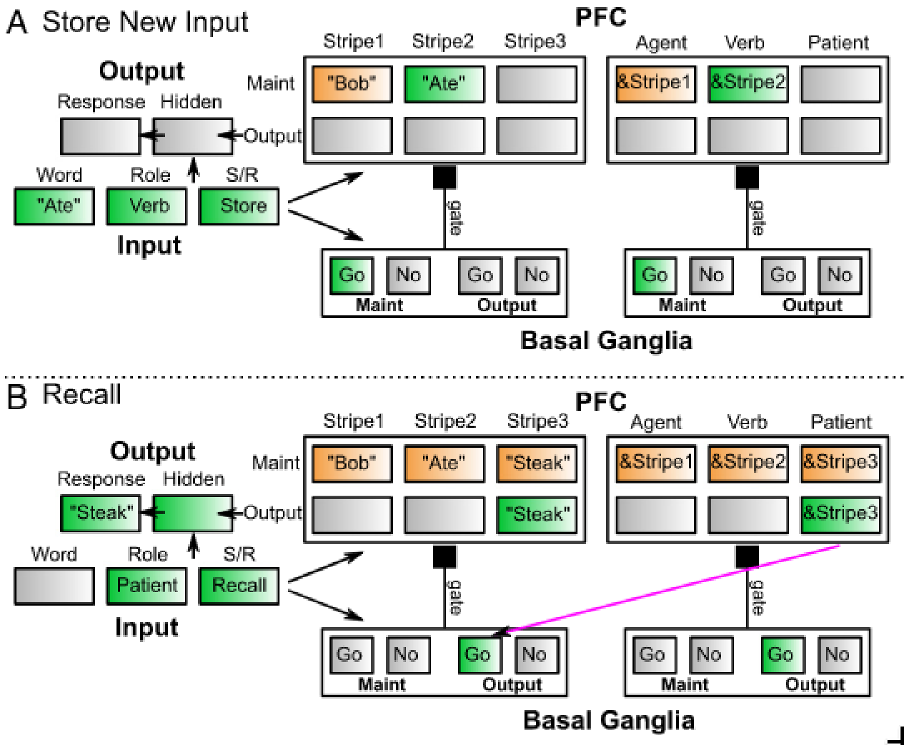


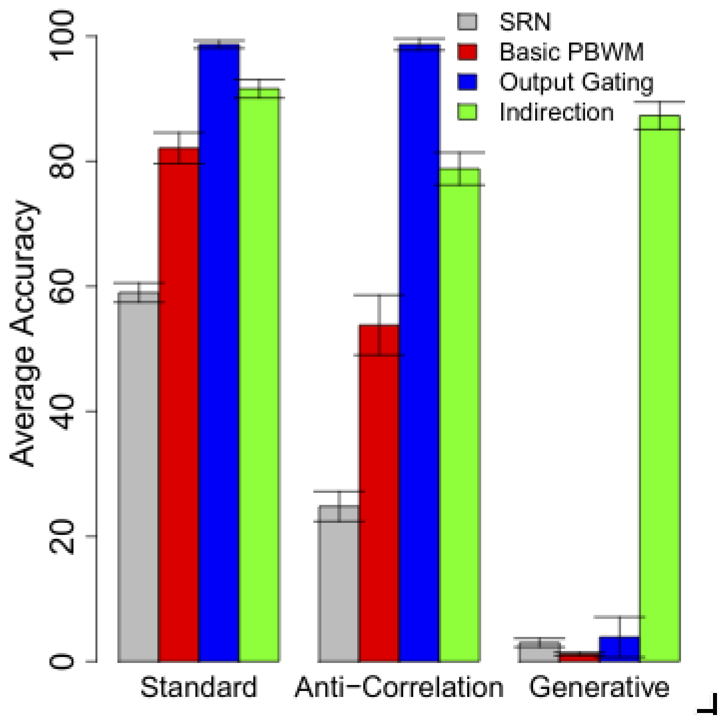
- Generate sets of arbitrary cardinatily
- No need to expensive loss function
- General embeddings for set-structured data



Stelzner, Kersting, and Kosiorek. “Generative Adversarial Set Transformers” ICML (2020). https://www.ml.informatik.tu-darmstadt.de/papers/stelzner2020ood_gast.pdf

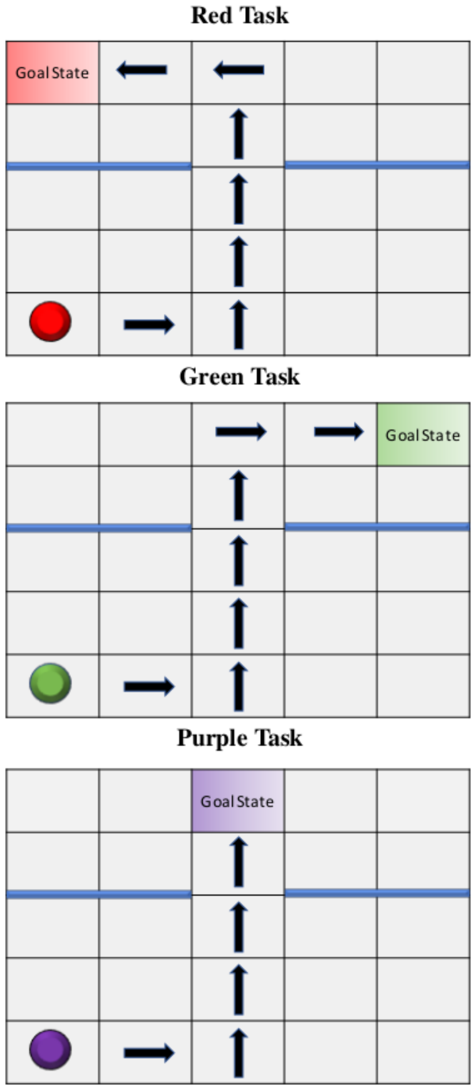
Another Type/Use of Attention – Indirection





Kriete, Trenton, et al. "Indirection and symbol-like processing in the prefrontal cortex and basal ganglia." Proceedings of the National Academy of Sciences 110.41 (2013): 16390-16395.

Indirection – Transfer Learning



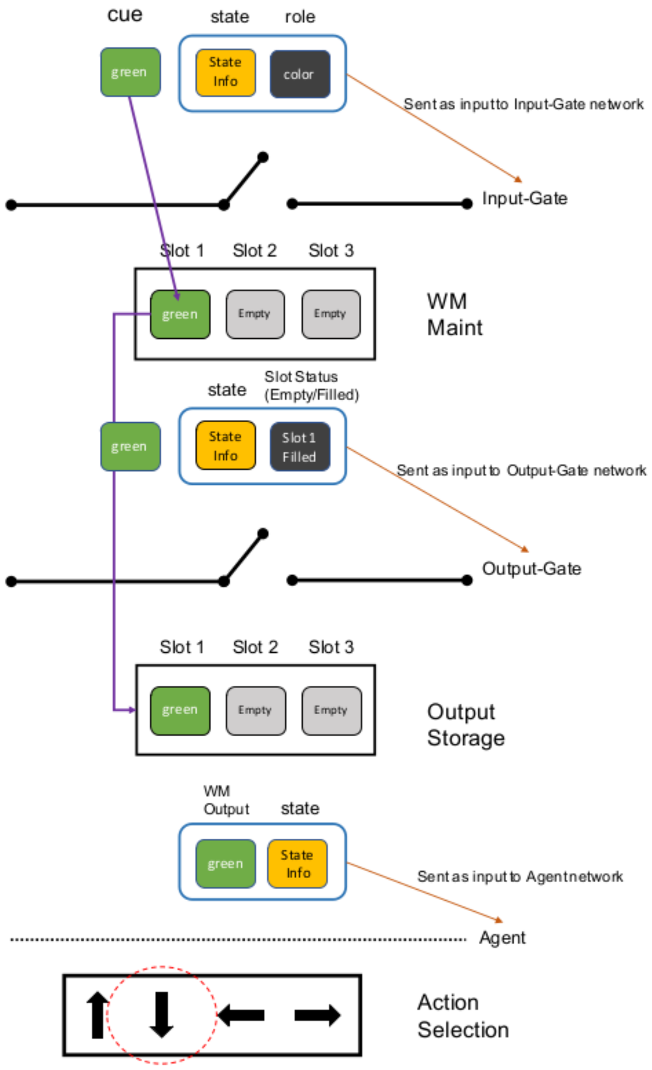
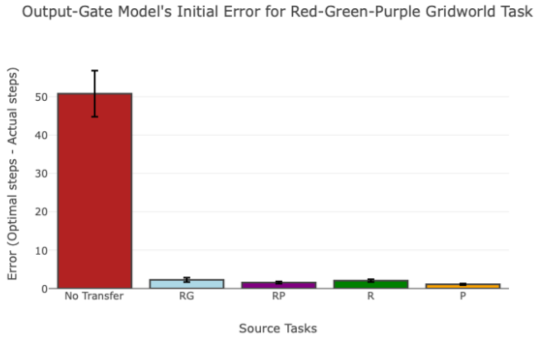
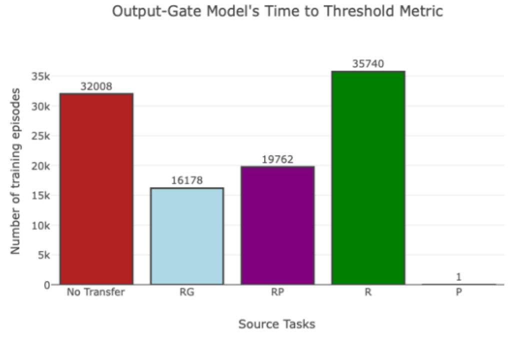


Table 1: Parameter Descriptions and Values

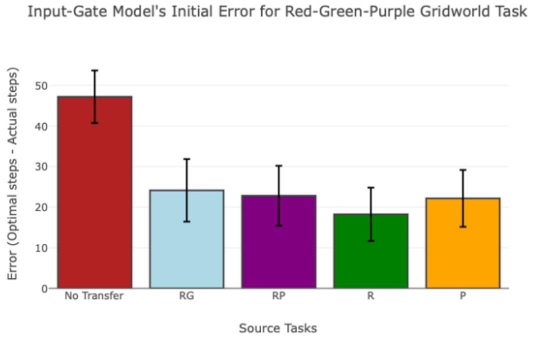
Name	Value	Description
n	1024	Size of HRR vectors
ε	0.1	Probability of non-greedy action choice
γ	0.9	Discount factor
α	0.1	Learning rate
λ	0.9	Trace decay
b	1	Network bias



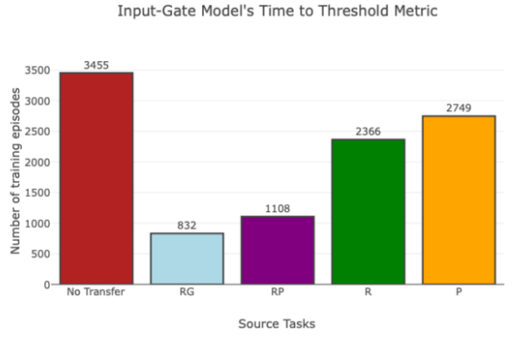
(a) Jumpstart metrics for the Output-Gate model



(a) Time to threshold metrics for the Output-Gate model



(b) Jumpstart metrics for the Input-Gate model

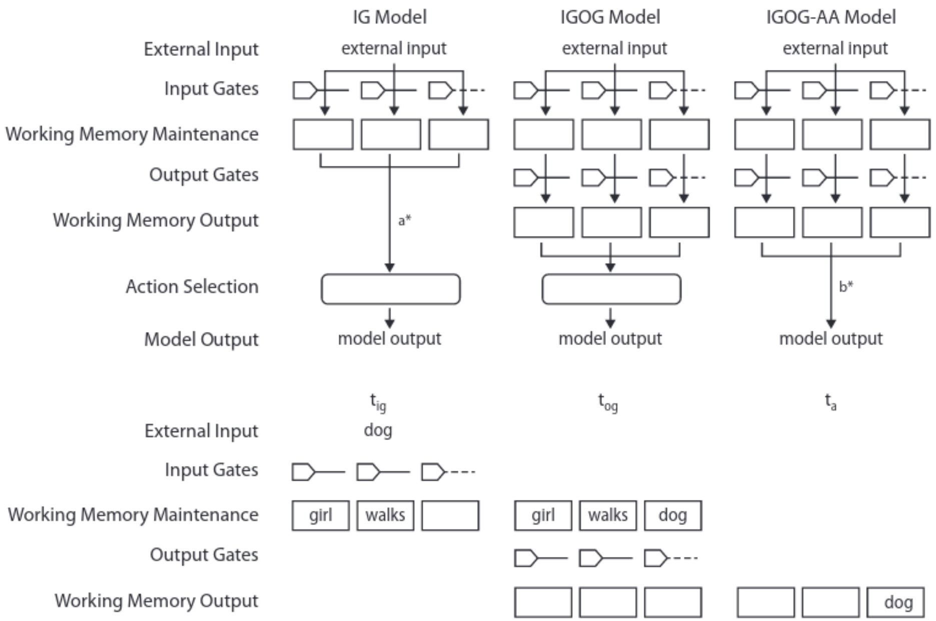


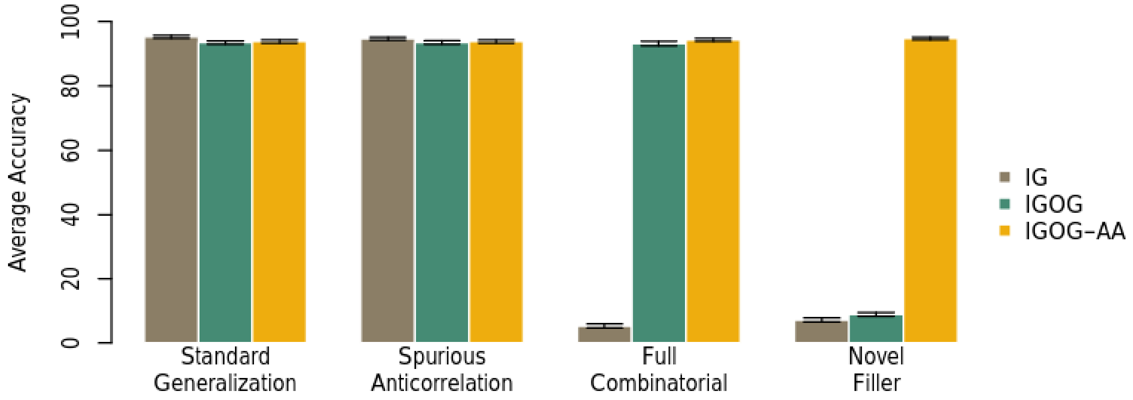
(b) Time to threshold metrics for the Input-Gate model

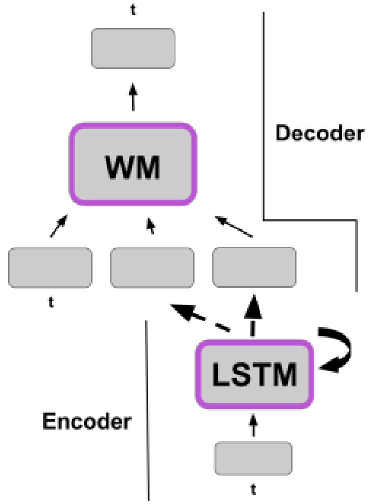
Williams and Phillips (2020) *34th AAAI Conference on Artificial Intelligence*

Indirection – Standard Neural Networks

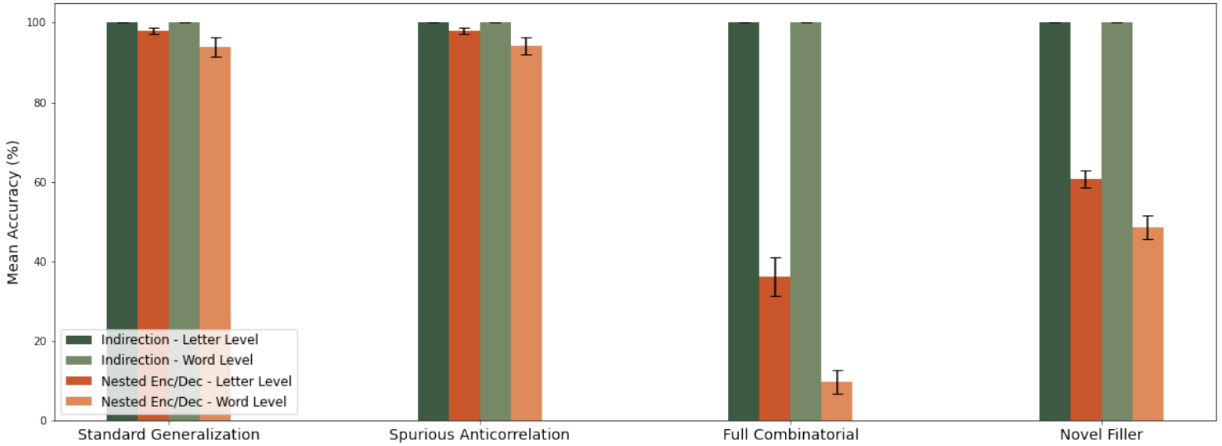
Jovanovich, 2017 http://jewlscholar.mtsu.edu/xmlui/handle/mtsu/5561



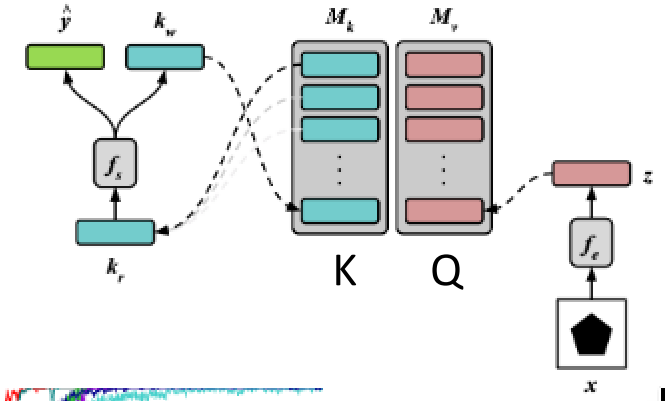


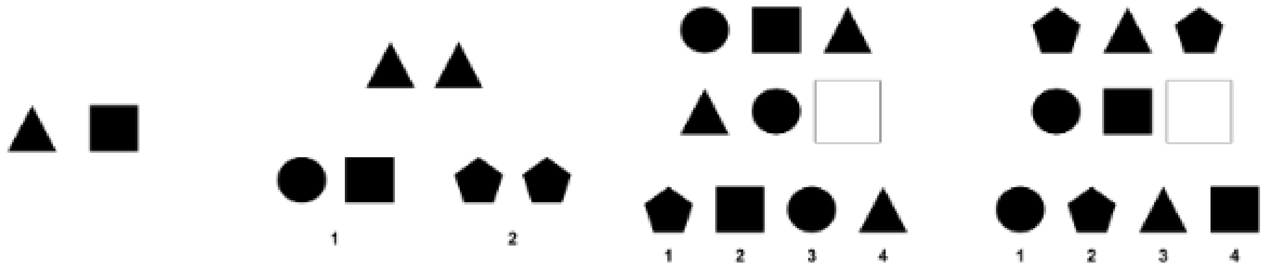


Mullinax, 2020 https://jewlscholar.mtsu.edu/handle/mtsu/6360

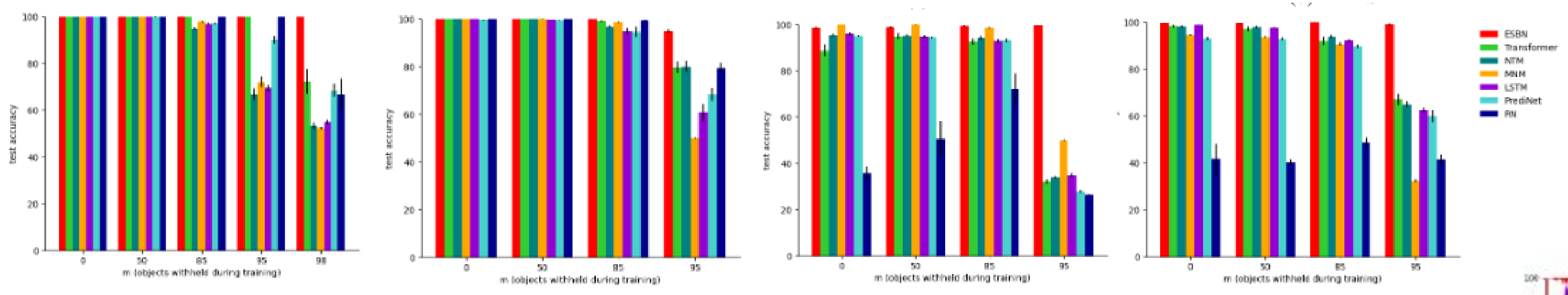


Indirection – Emergent Symbols

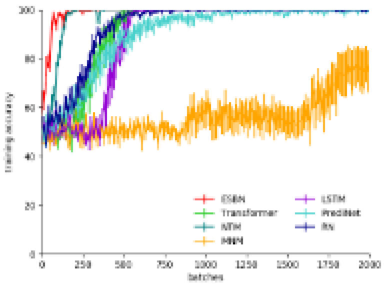




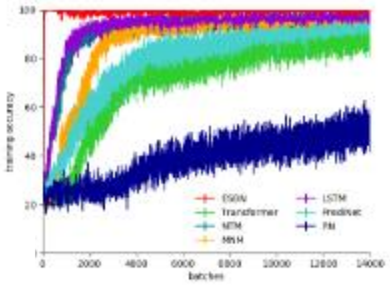
(a) Same/different (b) RMTS (c) Distribution-of-three (d) Identity rules



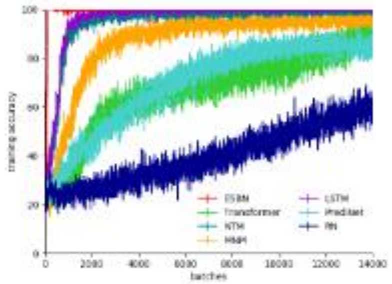
(a) Same/different (b) RMTS (c) Distribution-of-three (d) Identity rules



(a) RMTS



(b) Distribution-of-three



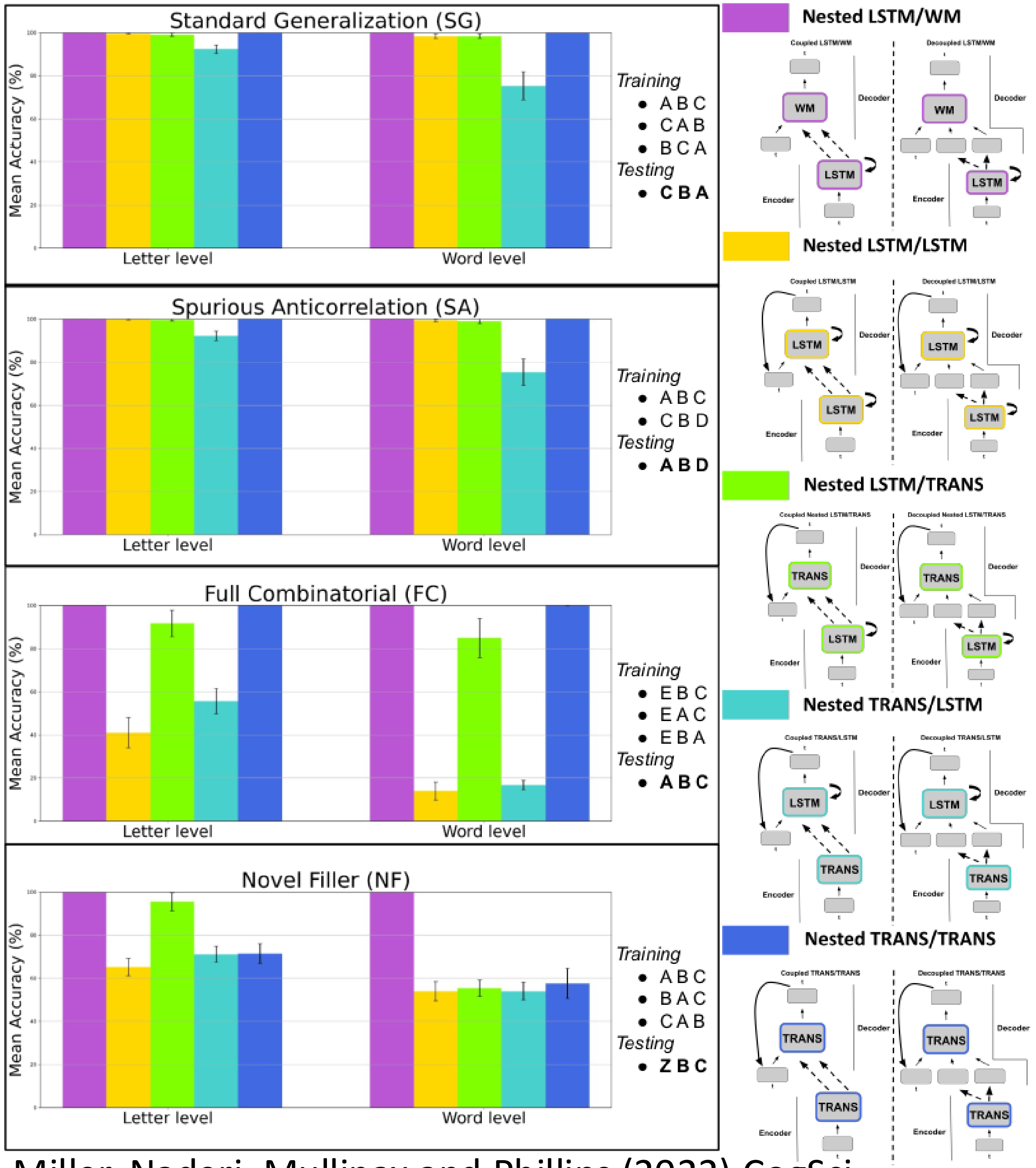
(c) Identity rules

Webb, Taylor W., Ishan Sinha, and Jonathan D. Cohen. "Emergent Symbols through Binding in External Memory." arXiv preprint arXiv:2012.14601 (2020).

Webb, Taylor, et al. "Learning representations that support extrapolation." International Conference on Machine Learning. PMLR, 2020.

Transformer Limitations

- The **transformer** is also a **clear improvement in extrapolation** compared to existing recurrent networks
- Note: *quantifiable* difference of a *qualitatively* different behavior
- Extrapolation abilities do not extend to true **indirection**
- However, **working memory** can currently still overcome this limitation by providing the correct **inductive bias**
- Perhaps, this property emerges spontaneously when training on large data sets: so-called **induction heads**? Olsson et al., 2022: https://arxiv.org/abs/2209.11895



Miller, Naderi, Mullinax and Phillips (2022) CogSci