Attention & Transformers

Advances & Applications

Yingzhen Li (yingzhen.li@imperial.ac.uk)

Attention applications pre-2017

Attention used in RNN generators back in 2013:

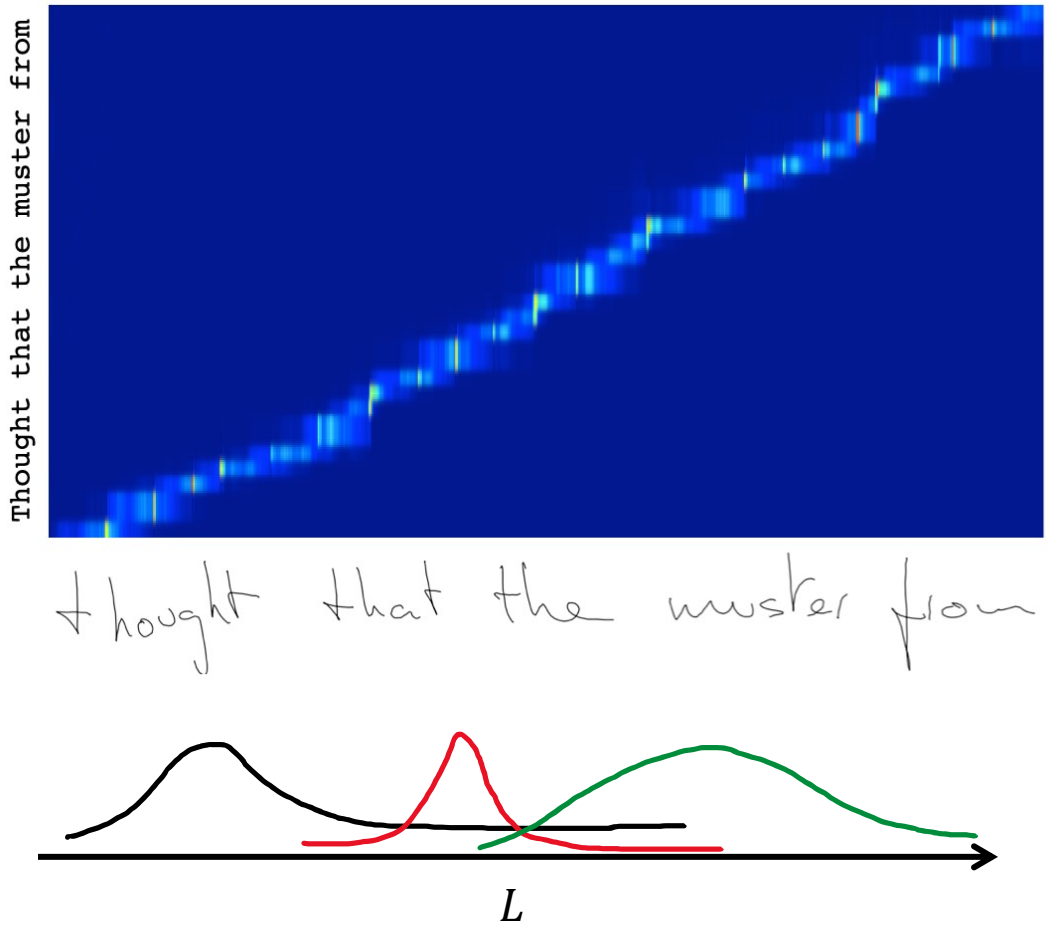
- Idea: align two different sequences with different length
  - $x_{1:T}$: x_t is the pen position at time t
  - $y_{1:L}$: y_l is the l^{th} character in the text

RNN parameterised mixture of Gaussian attention

$$Attention(x_t, y_{1:L}) = \sum_{l=1}^{L} \phi(t,l) emb(y_l)$$

$$\phi(t,l) = \sum_{k=1}^{K} \alpha_t^k \exp(-\beta_t^k (l - \kappa_t^k)^2)$$

$$\alpha_t^k, \beta_t^k, \kappa_t^k = MLP(h_t), h_{1:T} = RNN(x_{1:T})$$



Graves. Generating Sequences with Recurrent Neural Networks. arXiv:1308.0850

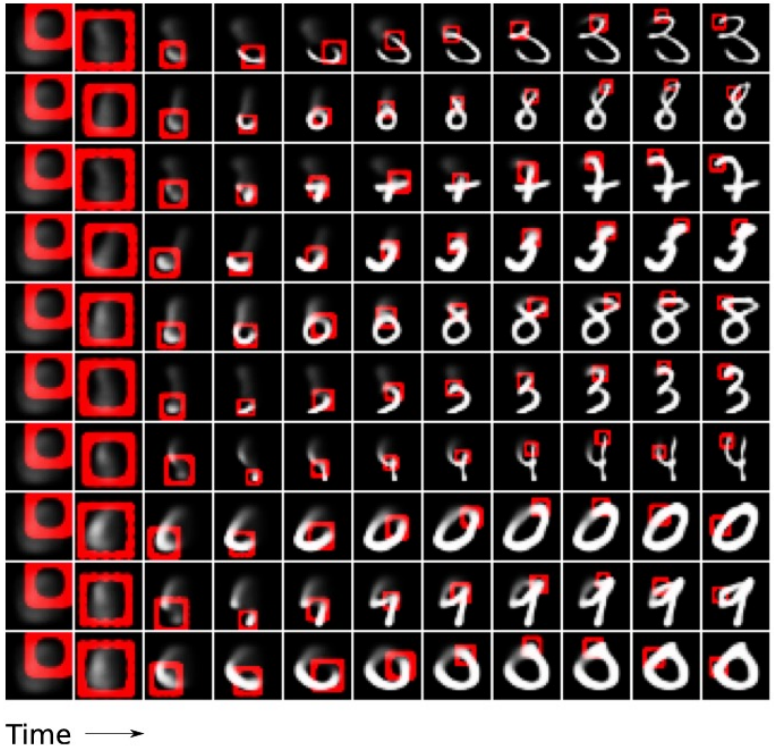
Attention applications pre-2017

DRAW model, extending Graves (2013):

- A VAE generative model for images combining RNNs and Gaussian attention
- Treat the intermediate "drawing state" as latent variable:

$$p(x|z) = p(x|z_{1:T})$$

- Decoder uses "read" and "write" operations guided by attention
- Gaussian attention to select focused patch at time t
- The Gaussian attention filter parameters are obtained from the RNN in the decoder

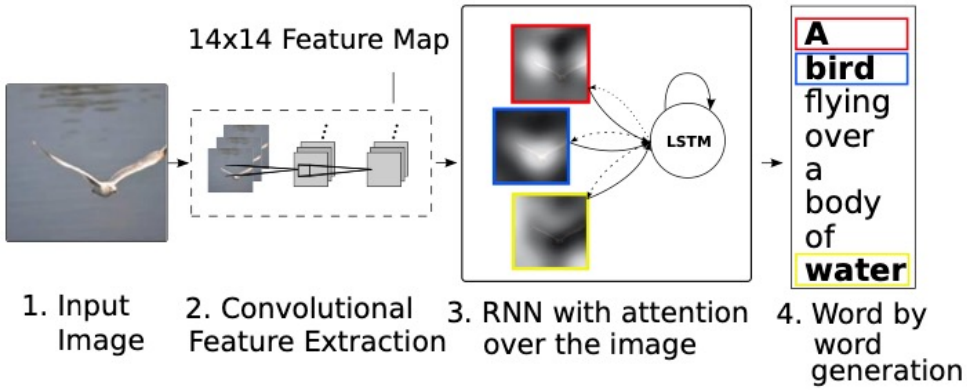


Gregor et al. DRAW: A Recurrent Neural Network For Image Generation. ICML 2015

Attention applications pre-2017

Attention applied in image captioning:

- CNN low-level features are indexed by spatial location
- Sort CNN features and then process with RNNs
- Apply similar RNN attention methods in Bahdanau et al. (2014)



A woman is throwing a frisbee in a park.

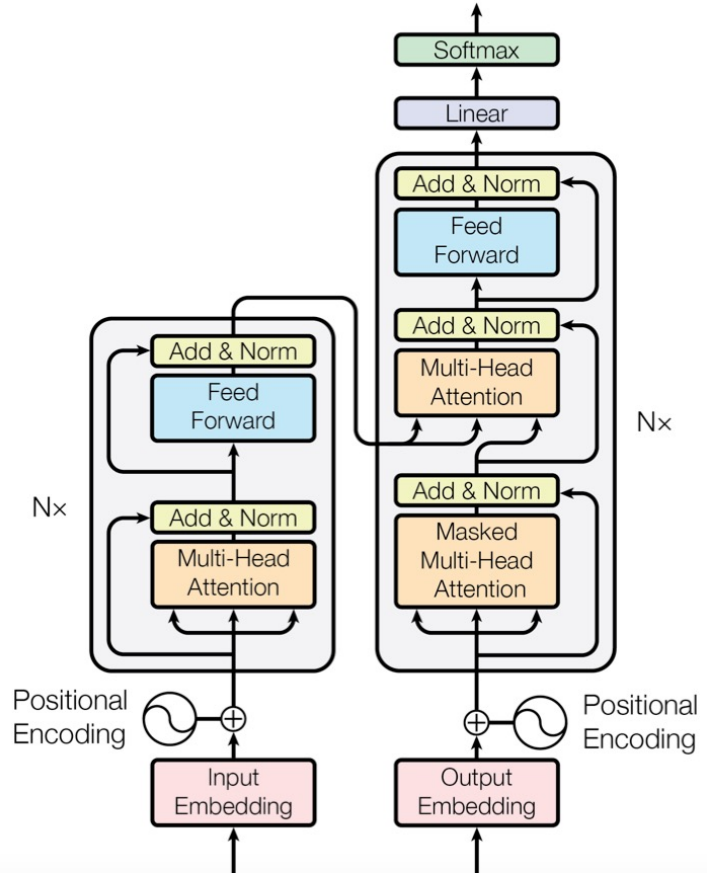
A dog is standing on a hardwood floor.

Xu et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. ICML 2015

Transformer in NLP applications

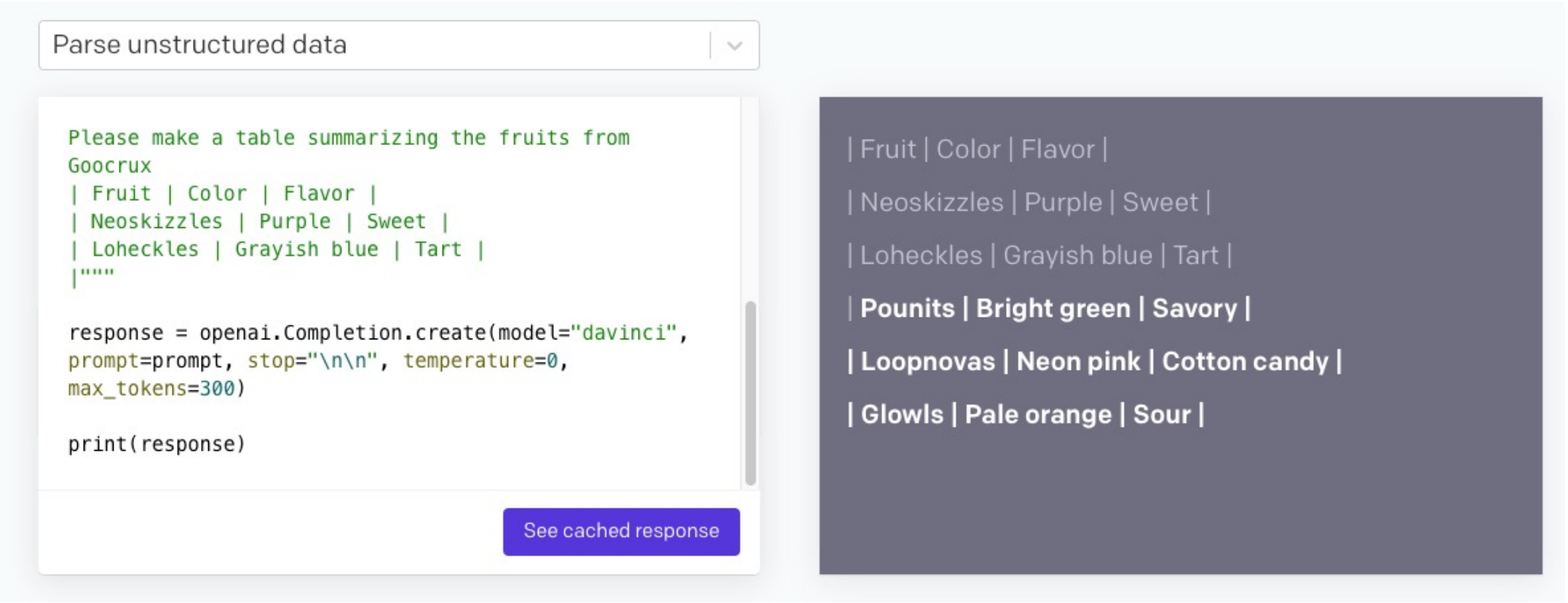
- Pretraining Transformer models on massive data

- Deeper & bigger Transformer architecture with modifications
- Pretrained on very big corpus by e.g. randomly masking out and predicting words in a sequence
- Fine-tune on specific tasks that the user cares



Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL 2019
Brown et al. Language Models are Few-shot Learners. NeurIPS 2020

Transformer in NLP applications



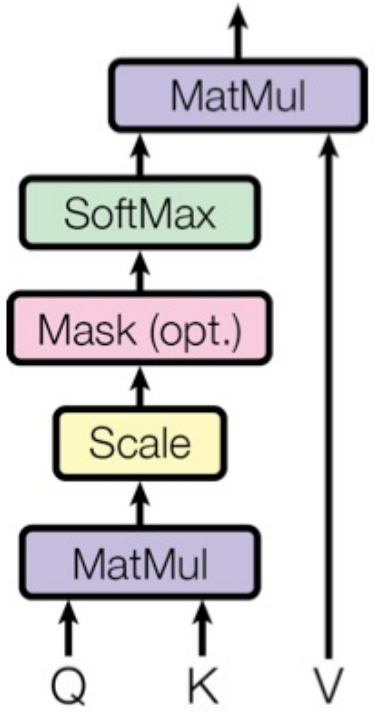
beta.openai.com

Brown et al. Language Models are Few-shot Learners. NeurIPS 2020

Multi-head Attention in Other Applications

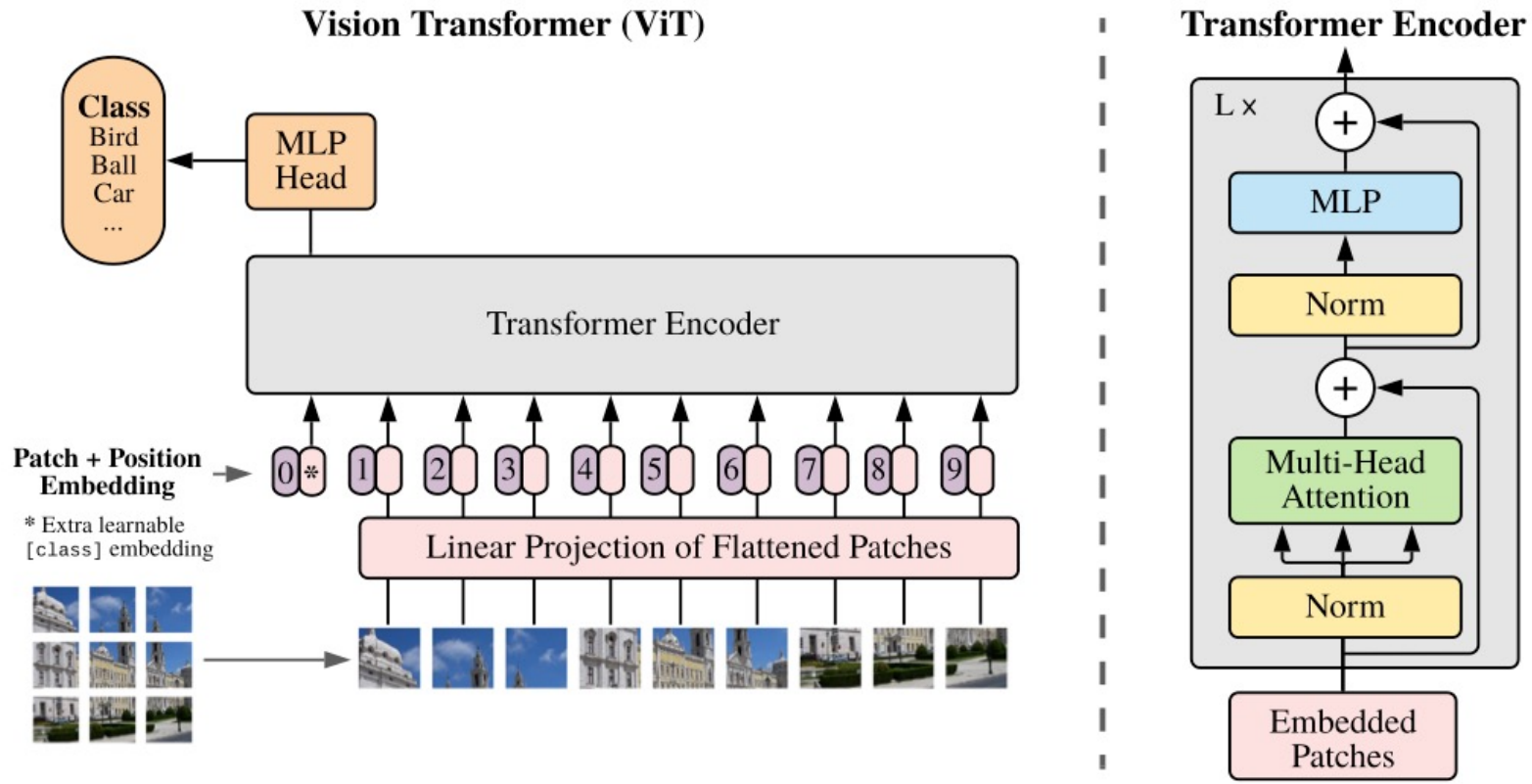
⇒ Can be applied to any Set Data with/out indexing!

- Text = set of words, Image = set of pixels (or set of patches), Graph = set of nodes and edges, point cloud = set of points, …
- Cross-modality application: embed points from diff. modality to the same space then apply attention



Vaswani et al. Attention is all you need. NeurIPS 2017

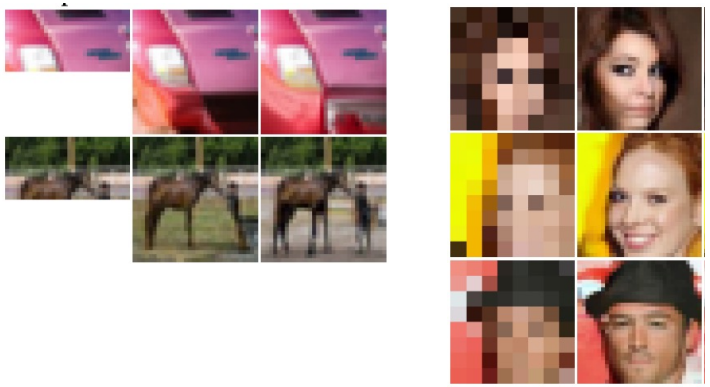
Multi-head Attention in Other Applications



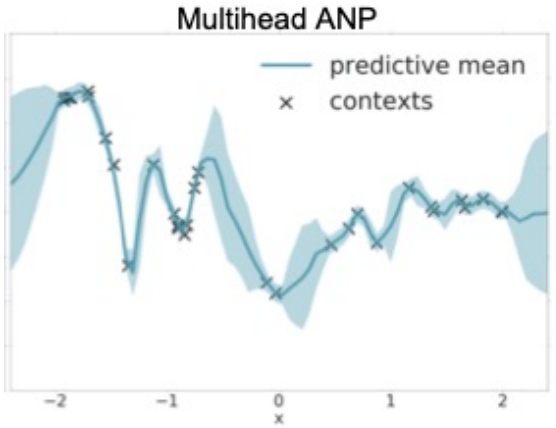
Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021

Multi-head Attention in Other Applications

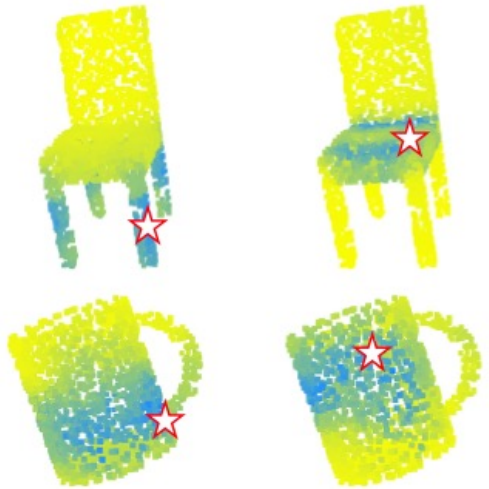
Image completion & super-resolution



Regression & supervised learning



Point cloud applications



Parmar et al. Image Transformer. ICML 2018
Kim et al. Attentive Neural Processes. ICLR 2019
Guo et al. PCT: Point Cloud Transformer. arXiv:2012.09688

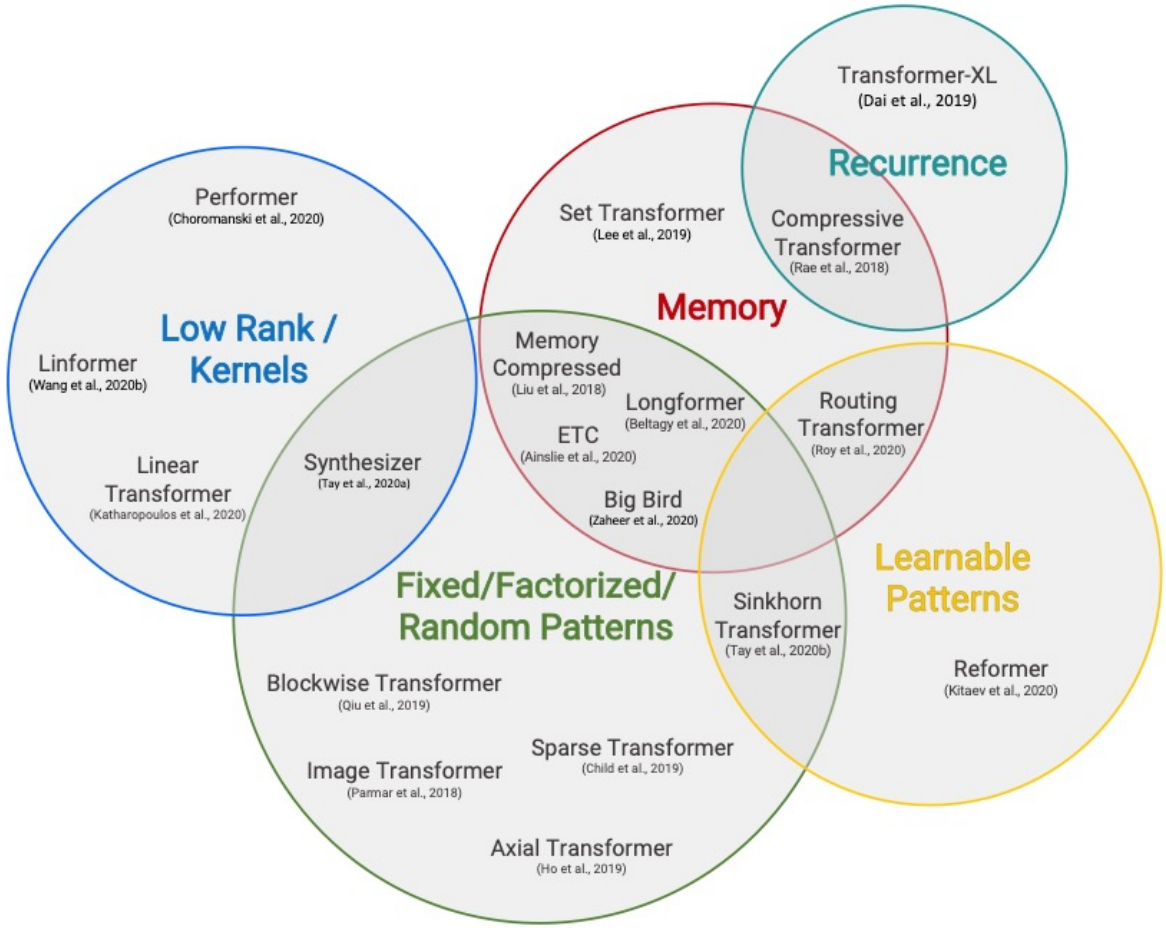
Efficient Transformers

Self attention complexities:

Assume the query $Q \in R^{N \times d_q}$

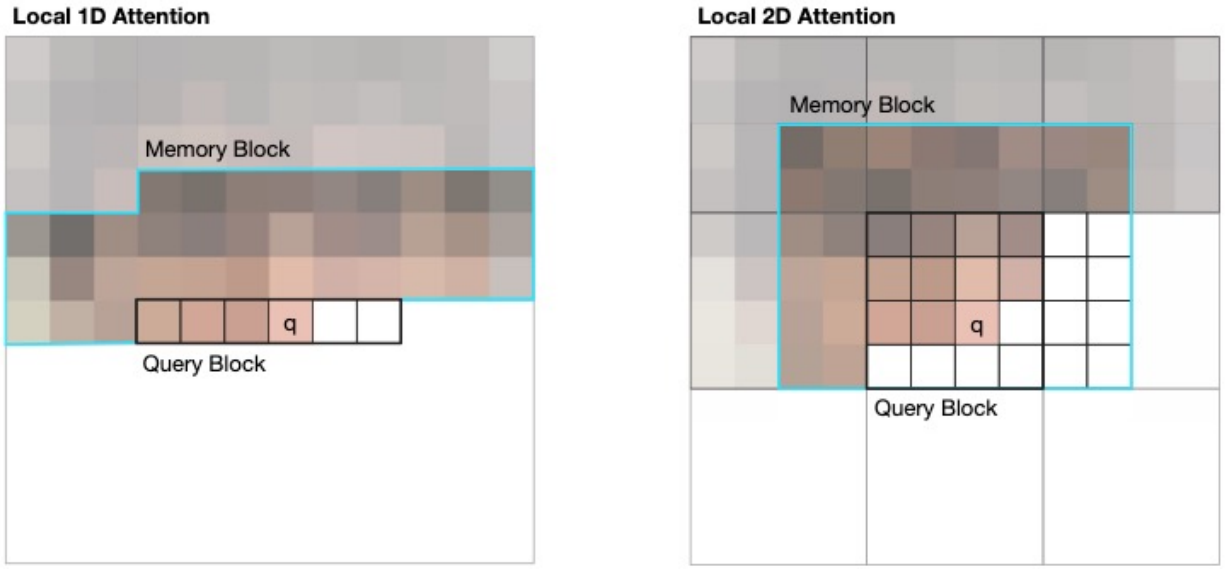
- Time complexity: $O(N^2)$
- Space complexity: $O(N^2)$
- Both complexities also scales linearly with d_q

Many existing approaches for improving efficiency!



Tay et al. Efficient Transformers: A Survey. arXiv:2009.06732

Efficient Transformers

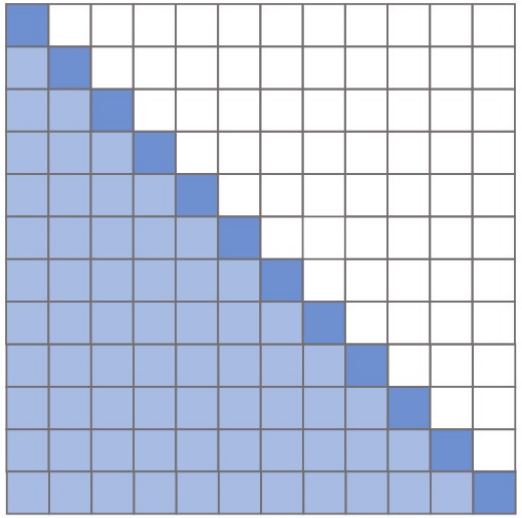


Local attention by defining “neighbourhood”:

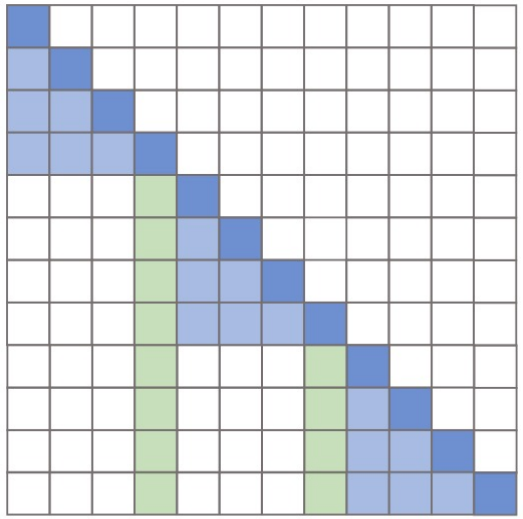
- Treat the input as a sequence with position ordering and build an auto-regressive model
- At current position, both the query and the key (or memory) inputs are “local”

Parmar et al. Image Transformer. ICML 2018

Efficient Transformers



(a) Transformer



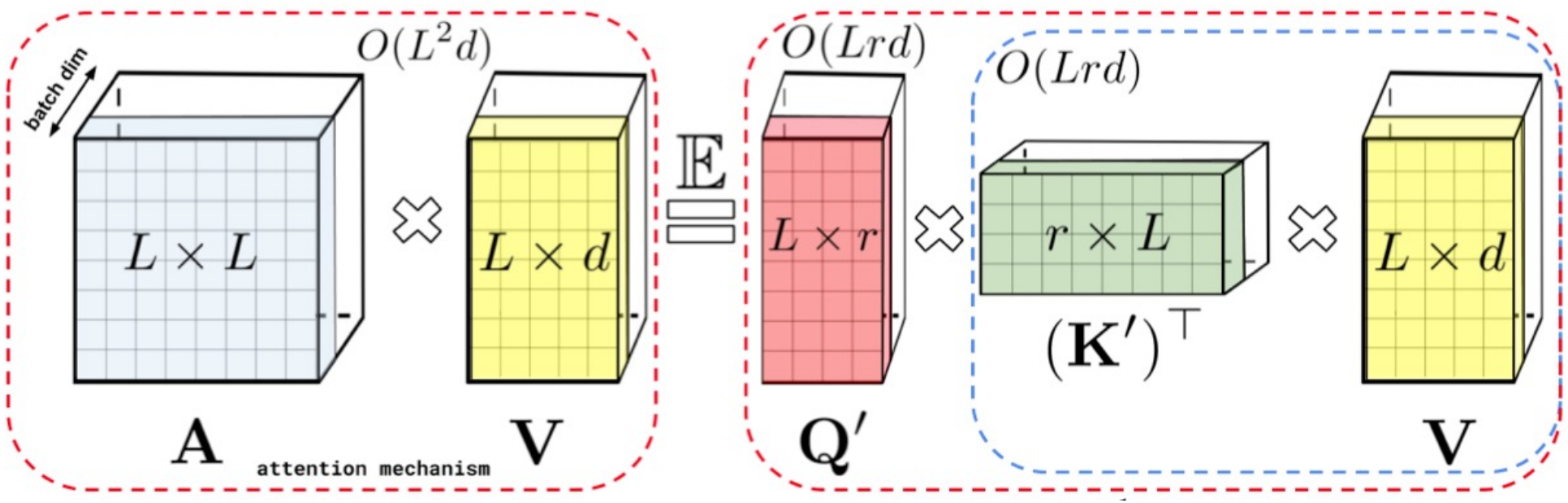
(b) Sparse Transformer

Sparse attention by defining attention weight matrix patterns:

- Full attention matrix before non-linearity: $(QK^T)_{ij} = \langle q_i, k_j \rangle$
- Sparse attention matrix: fix rules such that only for certain pairs of (i, j) the attention entry is computed
- Different attention heads can have different sparsity patterns

Child et al. Generating Long Sequences with Sparse Transformers. arXiv:1904.10509

Efficient Transformers



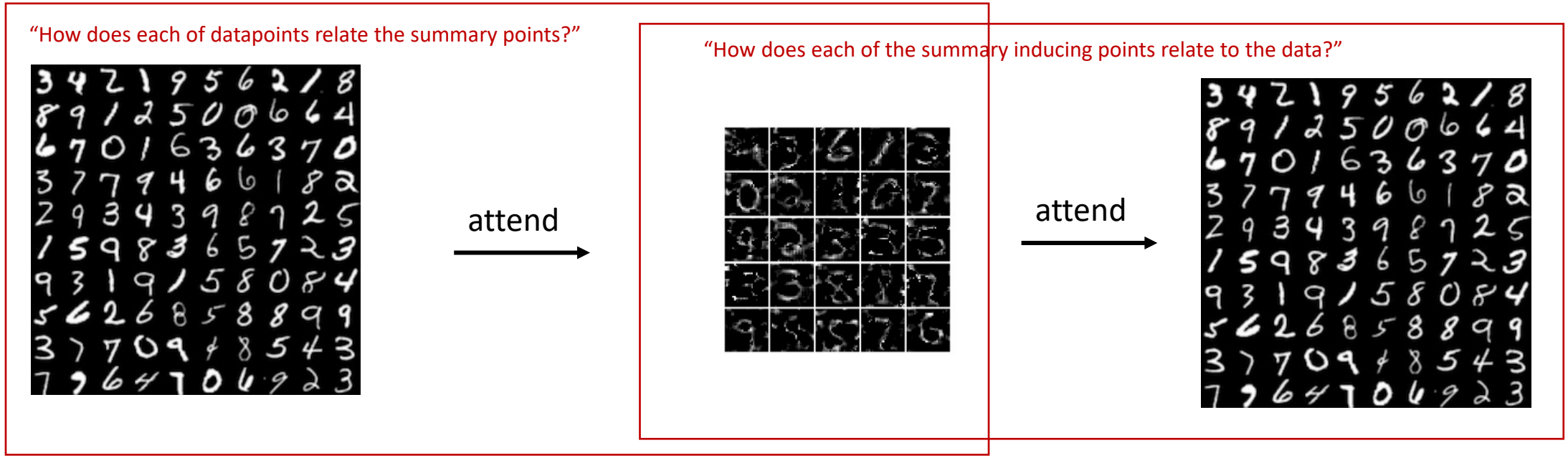
Low-rank approximations:

Can be much faster to compute!

- If the activation $a(\cdot)$ in $Attention(Q, K, V, a)$ is linear: $(QK^T)V = Q(K^TV)$
- In general the attention matrix has entries $A_{ij} = K(q_i, k_j)$ with $K(\cdot,\cdot)$ a kernel function
- Random feature approximation can be done to (approximately) compute A fastly

Choromanski et al. Rethinking attention with Performers. ICLR 2021

Efficient Transformers



Learnable inducing points for set summary:

- $MAB(X, X)$ as building block, requiring $MultiHeadAttention(X, X)$
- Idea: Using learnable inducing points I_M as a "bridge":

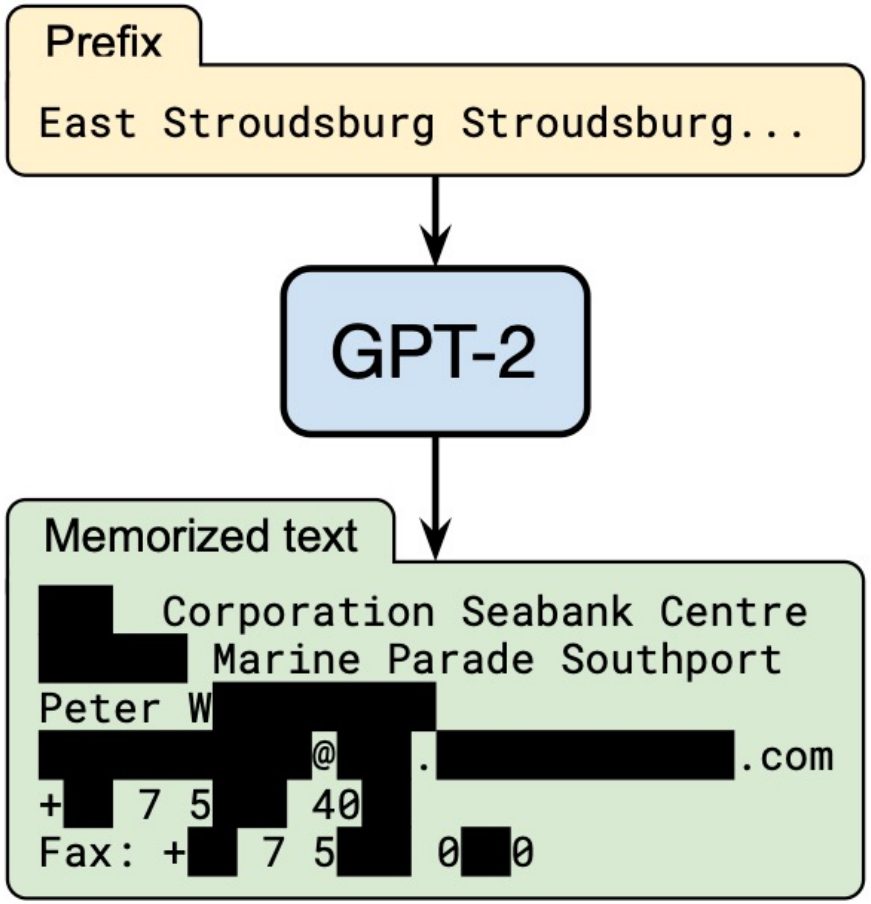
$$ISAB = MAB(X, MAB(I_M, X))$$

Lee et al. Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks. ICML 2019
Figure adapted from: Tomczak and Welling. VAE with a VampPrior. AISTATS 2018.

Memorisation in Transformers

Potential privacy issues:

- Transformer-based Language models has billions of parameters
- Capable for memorising the input data
- Currently they are pre-trained on open-web text without any privacy protection
- Attacks can steal sensitive input data



Carlini et al. Extracting Training Data from Large Language Models. arXiv:2012.07805