Learnable Fourier Features for Multi-Dimensional Spatial Positional Encoding

Yang Li, Si Si, Gang Li, Cho-Jui Hsieh*, Samy Bengio

Google Research UCLA*

Neural Attentional Mechanisms

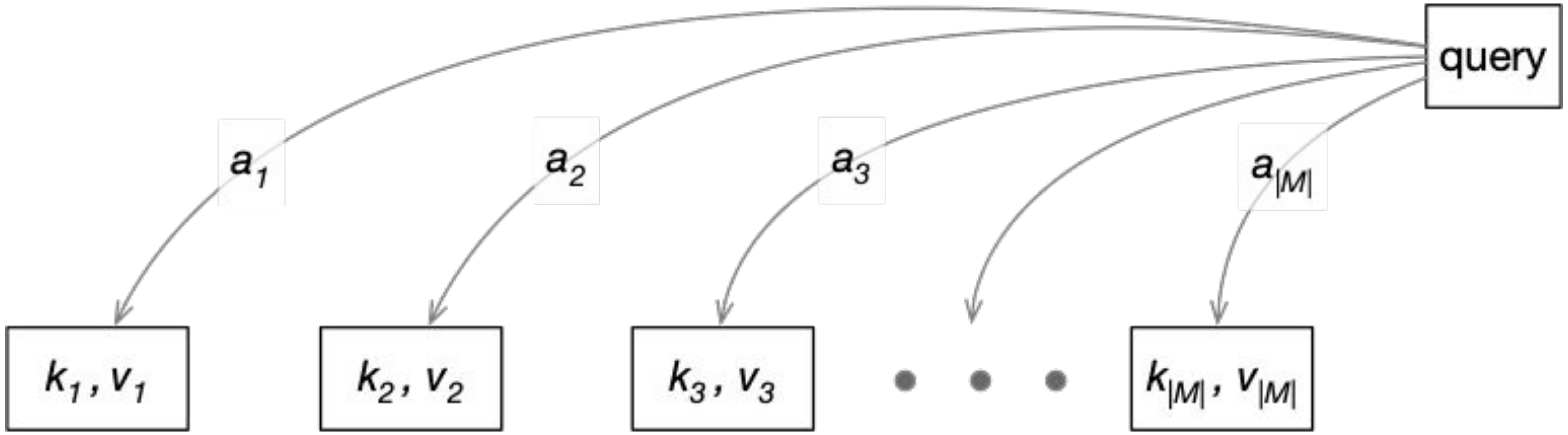
Attention weights

$$a_i = \frac{\exp(f_{att}(q, k_i))}{\sum_{j=1}^{|M|} \exp(f_{att}(q, k_j))}$$

Attention output

$$O_q^M = \sum_{i=1}^{|M|} a_i v_i$$

Items to attend



Bahdanau et al., ICLR 2015
Luong et al., ACL 2015

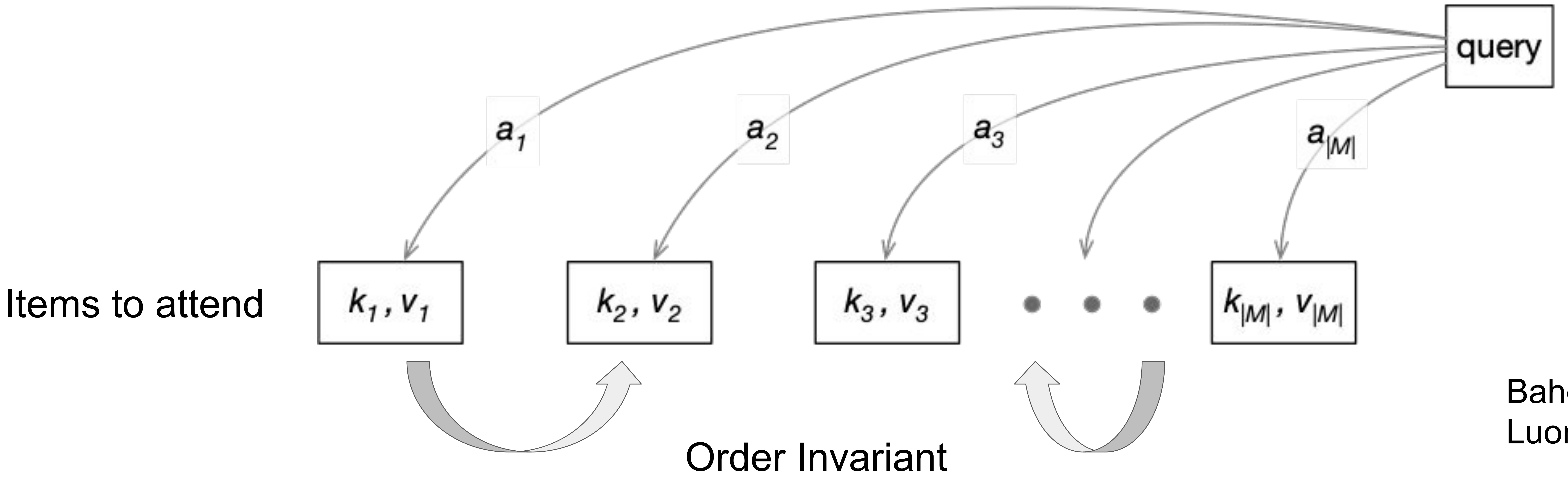
Neural Attentional Mechanisms

Attention weights

$$a_i = \frac{\exp(f_{att}(q, k_i))}{\sum_{j=1}^{|M|} \exp(f_{att}(q, k_j))}$$

Attention output

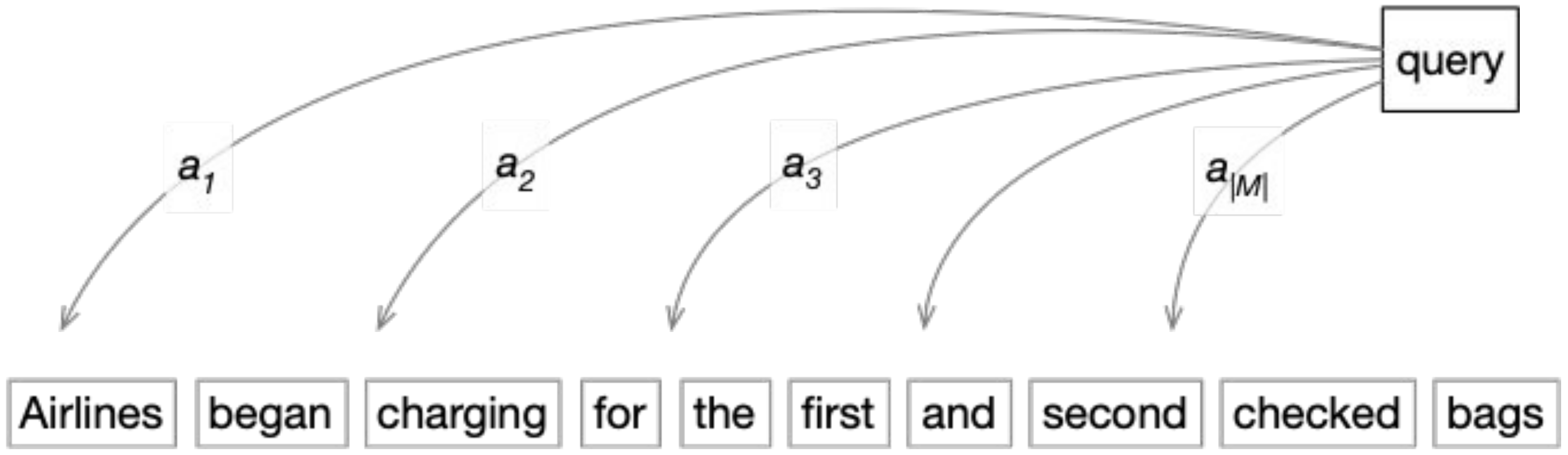
$$O_q^M = \sum_{i=1}^{|M|} a_i v_i$$



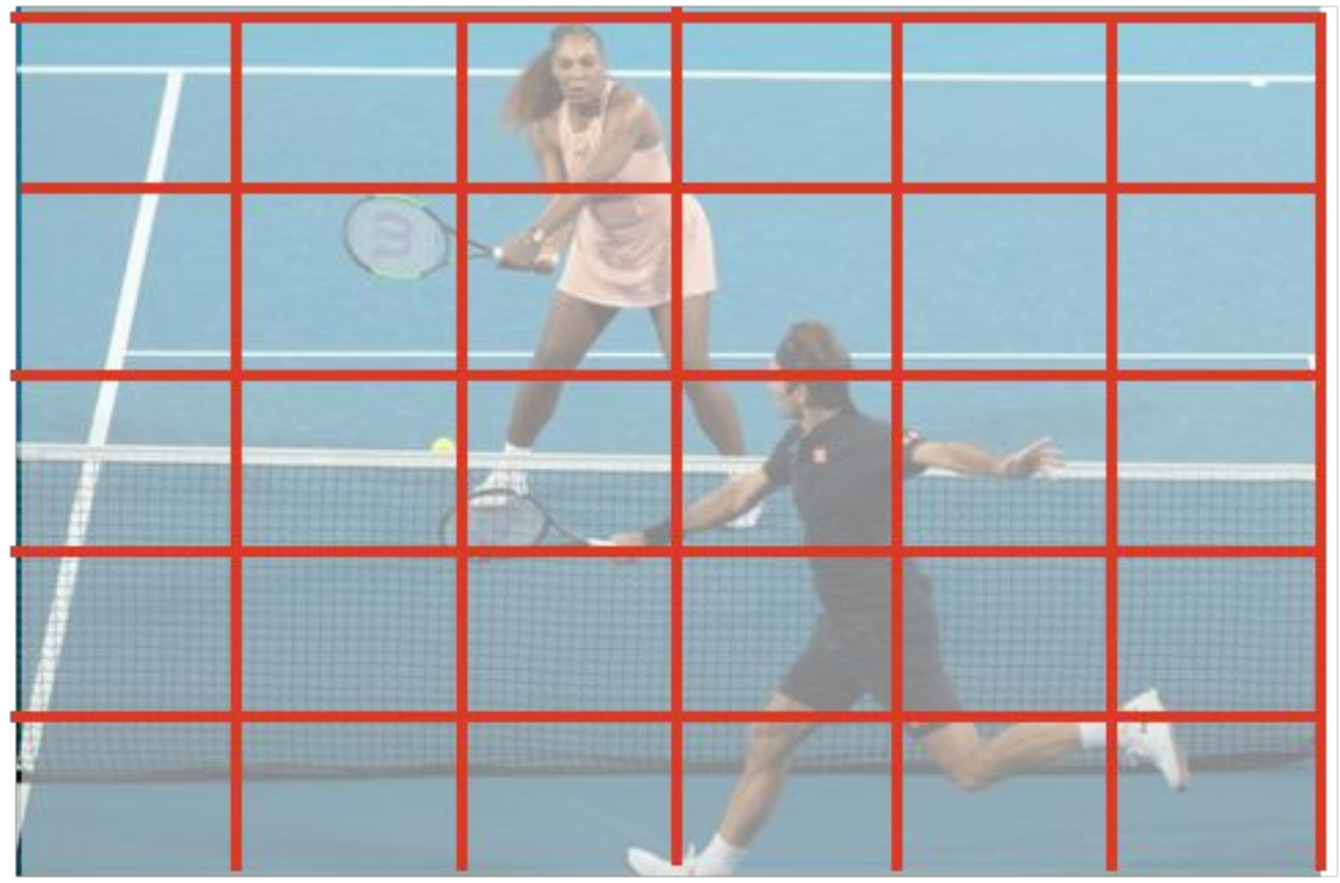
Bahdanau et al., ICLR 2015
Luong et al., ACL 2015

Why Positional Encoding?

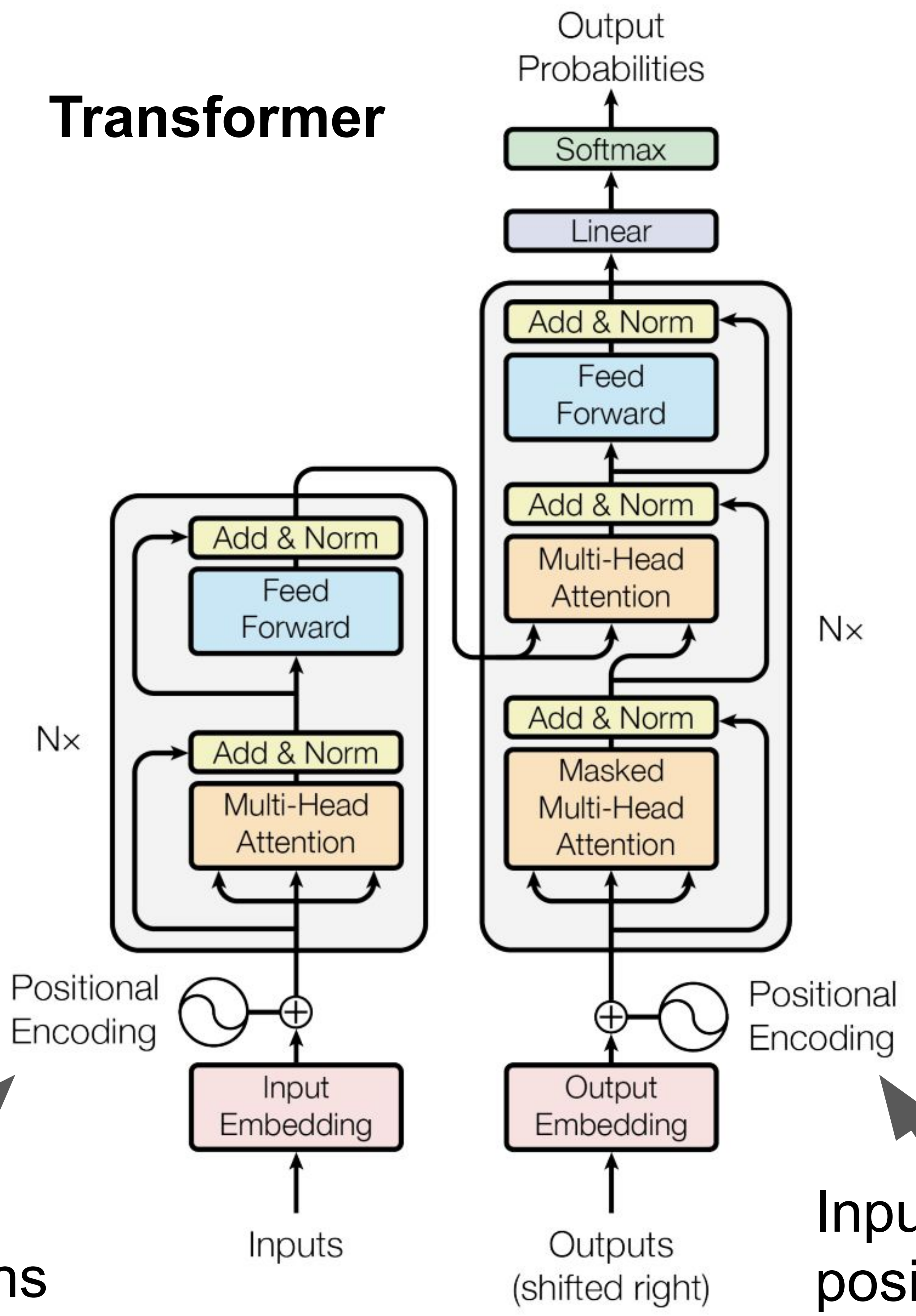
Language



Image



Transformer



Input positions

Input positions

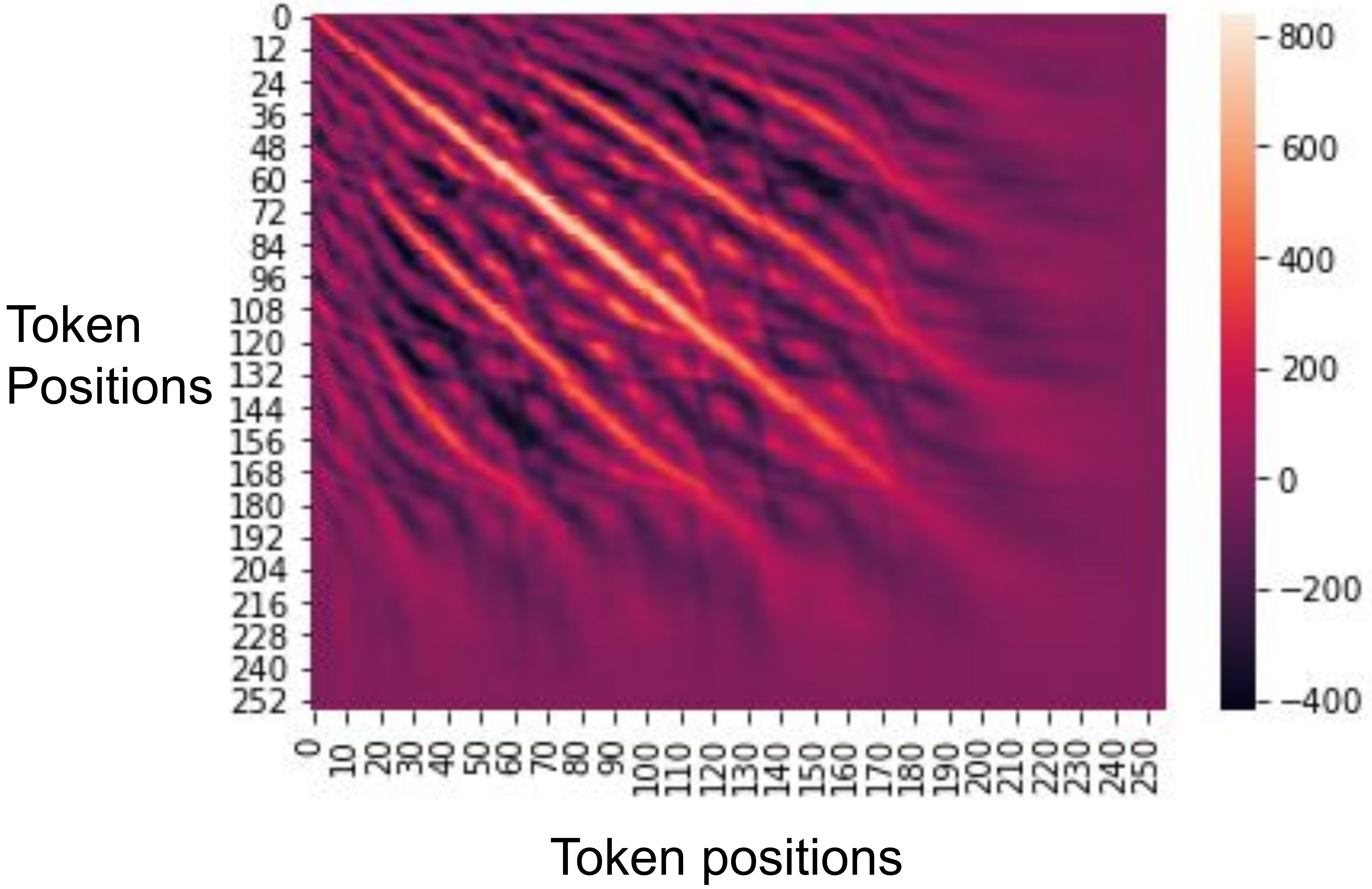
Attention is all you need, Vaswani et al. NIPS 2017

Existing Positional Encoding Methods

- Learnable embedding for discrete positions

Existing Positional Encoding Methods

- Learnable embedding for discrete positions



Learned by Transformer on
En-De WMT32k
Machine Translation Task

Existing Positional Encoding Methods

- Learnable embedding for discrete positions

- Sinusoidal positional encoding

$$PE(p, 2d) = \sin \frac{p}{10000^{2d/D}} \qquad PE(p, 2d+1) = \cos \frac{p}{10000^{2d/D}}$$

Existing Positional Encoding Methods

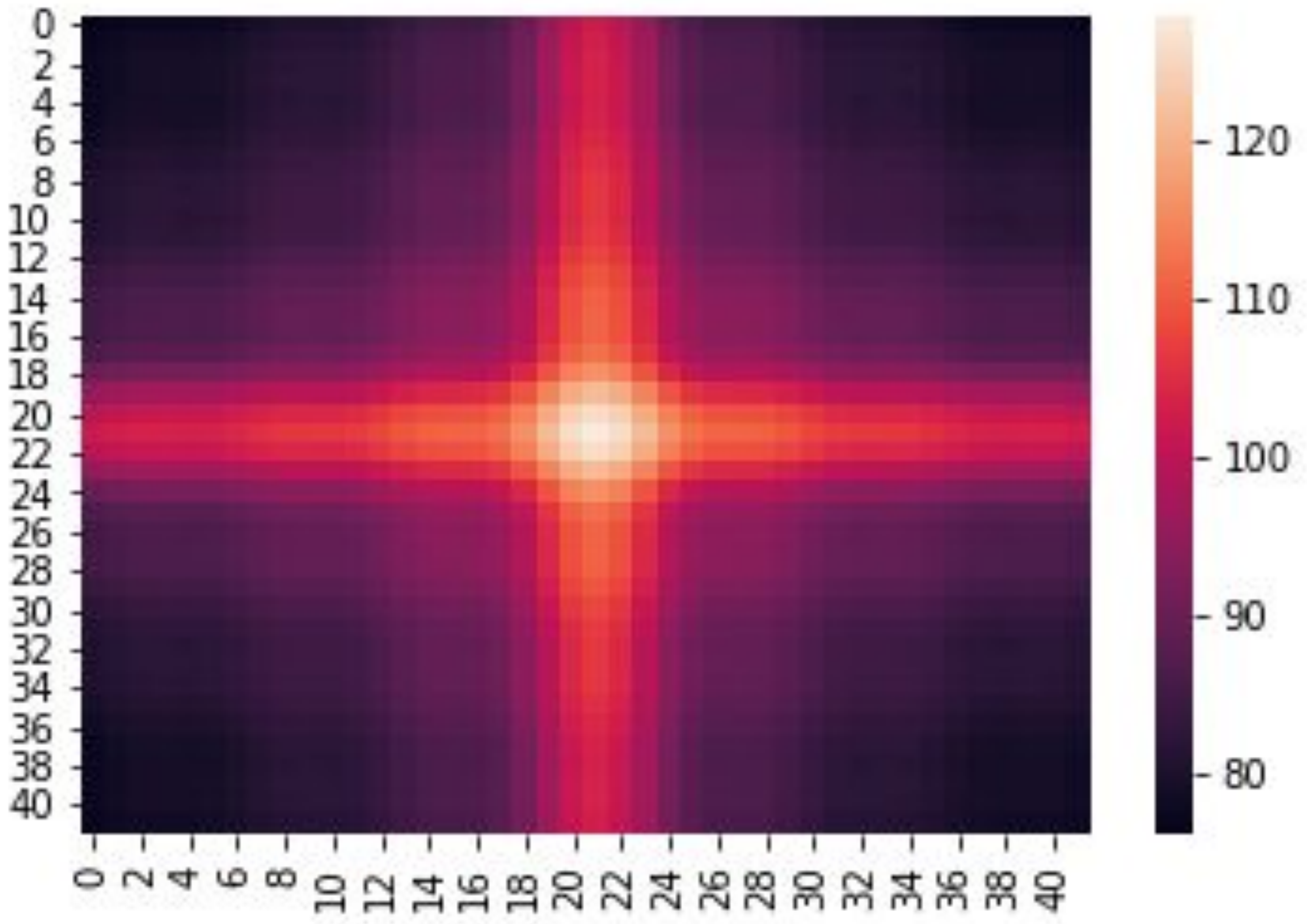
- Learnable embedding for discrete positions

- Sinusoidal positional encoding

$$PE(p, 2d) = \sin \frac{p}{10000^{2d/D}} \qquad PE(p, 2d+1) = \cos \frac{p}{10000^{2d/D}}$$

Sinusoidal encoding for 2D positions by concatenation

Vertical positions



Horizontal positions

Existing Positional Encoding Methods

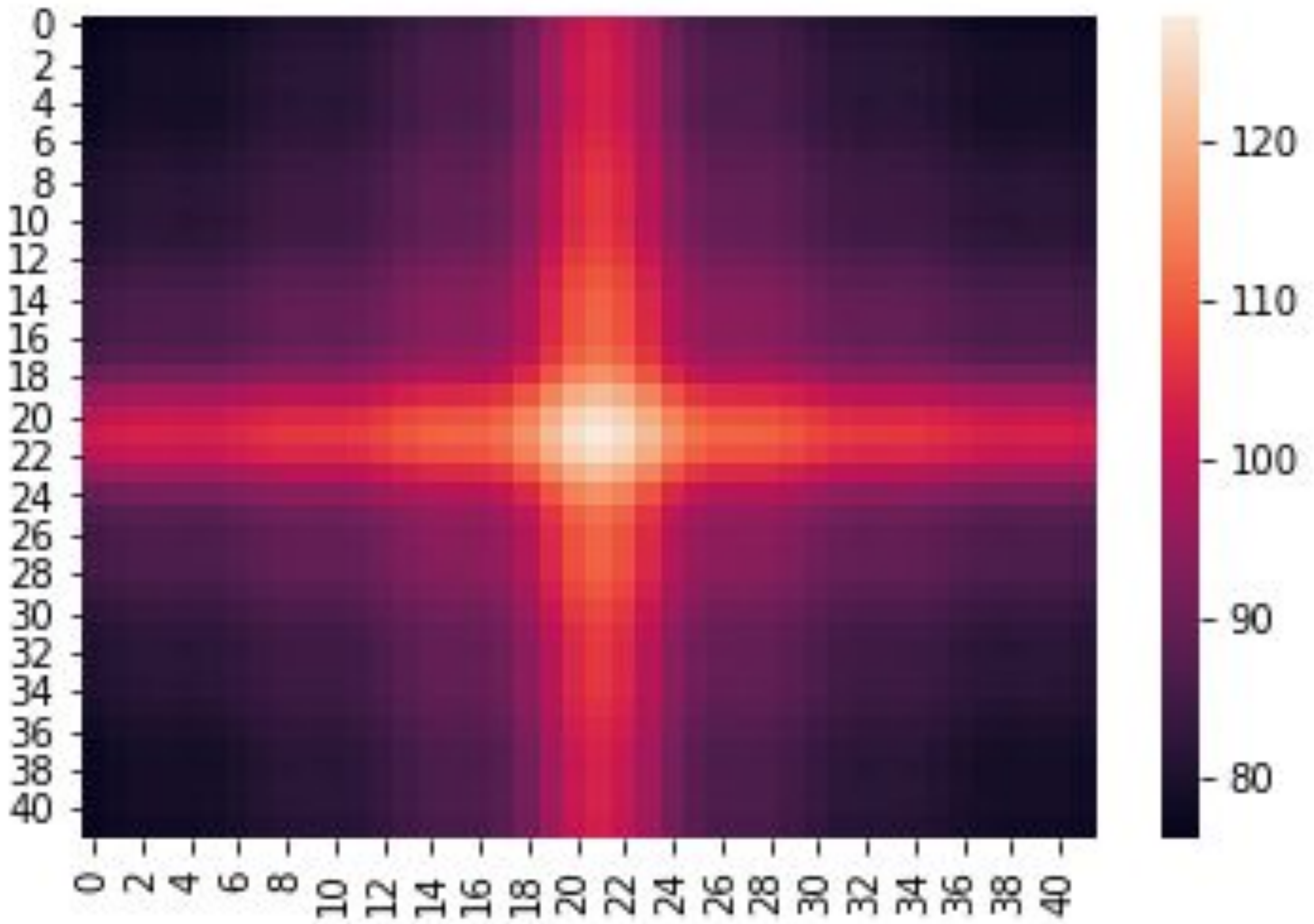
- Learnable embedding for discrete positions

- Sinusoidal positional encoding

$$PE(p, 2d) = \sin \frac{p}{10000^{2d/D}} \qquad PE(p, 2d+1) = \cos \frac{p}{10000^{2d/D}}$$

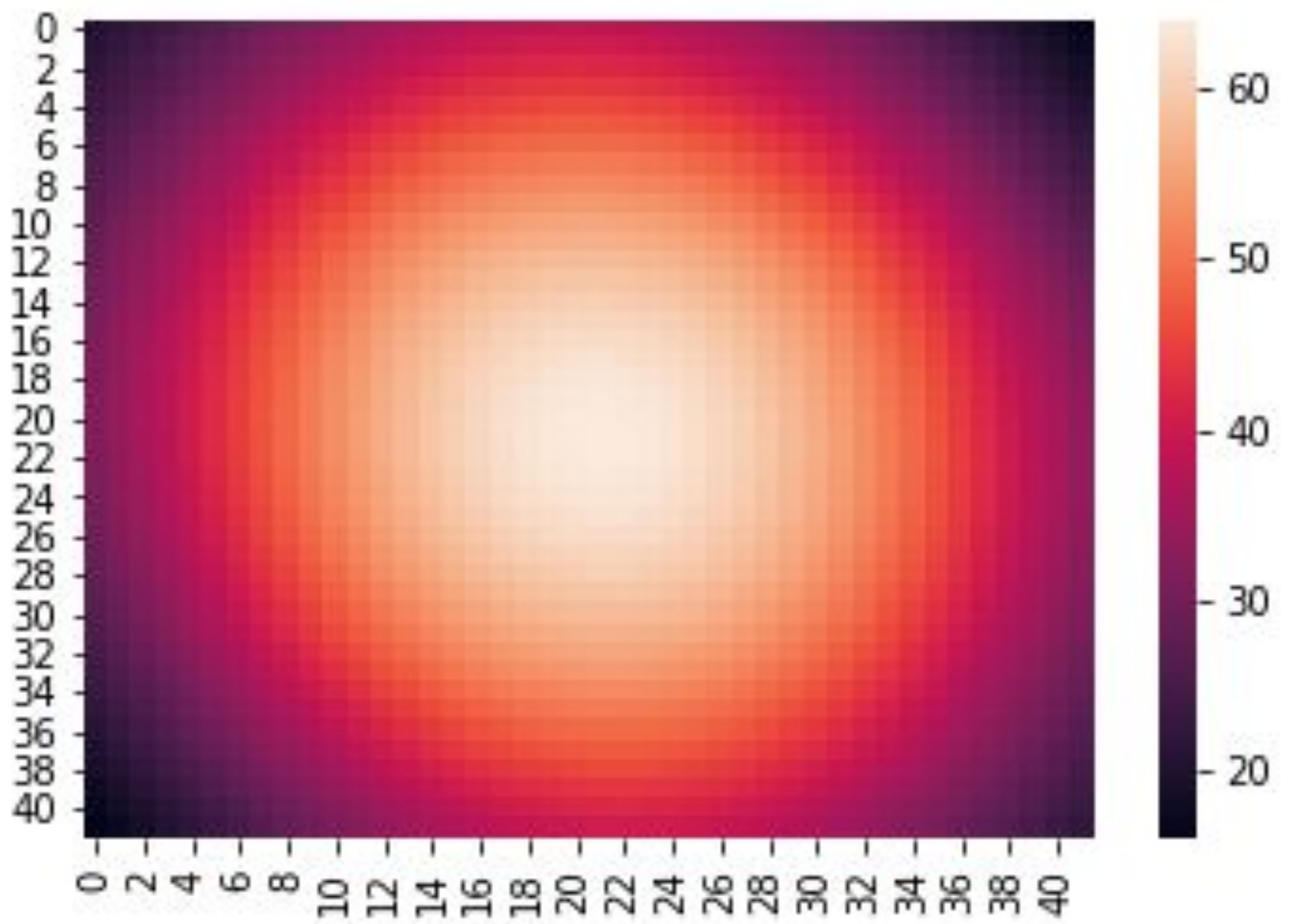
Sinusoidal encoding for 2D positions by concatenation

Vertical positions



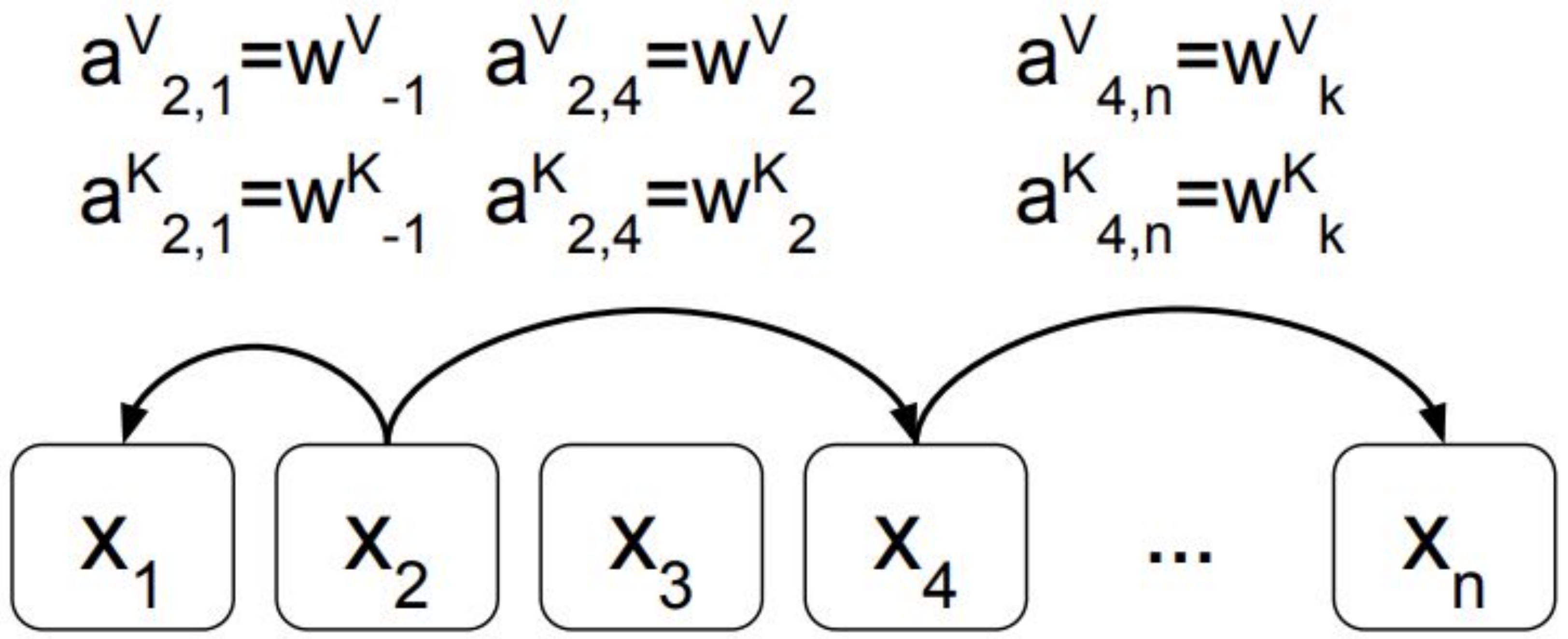
Horizontal positions

Ideal similarity for L2 distances



Existing Positional Encoding Methods

- Learnable embedding for discrete positions
- Sinusoidal positional encoding
- Relative positional encoding



Shaw et al. NAACL 2018

Learnable Fourier Feature Positional Encoding

Design objectives

- Positions as continuous-valued vectors
- Including inductive bias such as L2 distances
- Learnable & composable

Learnable Fourier Feature Positional Encoding

Given an M-dimensional position: $x \in R^M$

Acquire D-dimensional Fourier features: $r_x = \frac{1}{\sqrt{D}}[\cos xW_r^T \parallel \sin xW_r^T]$

Trainable parameters: $W_r \in R^{\frac{D}{2} \times M}$

Learnable Fourier Feature Positional Encoding

Given an M-dimensional position: $x \in R^M$

Acquire D-dimensional Fourier features: $r_x = \frac{1}{\sqrt{D}}[\cos xW_r^T \parallel \sin xW_r^T]$

Trainable parameters: $W_r \in R^{\frac{D}{2} \times M}$

Shift invariance: $r_x \cdot r_y = \frac{1}{D}\text{sum}\big(\cos((x-y)W_r^T)\big) := h_{W_r}(x-y)$

Learnable Fourier Feature Positional Encoding

Given an M-dimensional position: $x \in R^M$

Acquire D-dimensional Fourier features: $r_x = \frac{1}{\sqrt{D}}[\cos xW_r^T \parallel \sin xW_r^T]$

Trainable parameters: $W_r \in R^{\frac{D}{2} \times M}$

Shift invariance: $r_x \cdot r_y = \frac{1}{D}\mathrm{sum}\big(\cos((x-y)W_r^T)\big) := h_{W_r}(x-y)$

Approximate Gaussian kernel: $W_r \sim \mathcal{N}(0, \gamma^{-2})$ $\quad r_x \cdot r_y \approx \exp(-\frac{\|x-y\|^2}{\gamma^2})$

Learnable Fourier Feature Positional Encoding

Given an M-dimensional position: $x \in R^M$

Acquire D-dimensional Fourier features: $r_x = \frac{1}{\sqrt{D}}[\cos xW_r^T \parallel \sin xW_r^T]$

MLP Modulator: $PE_x = \phi(r_x, \theta)W_p$

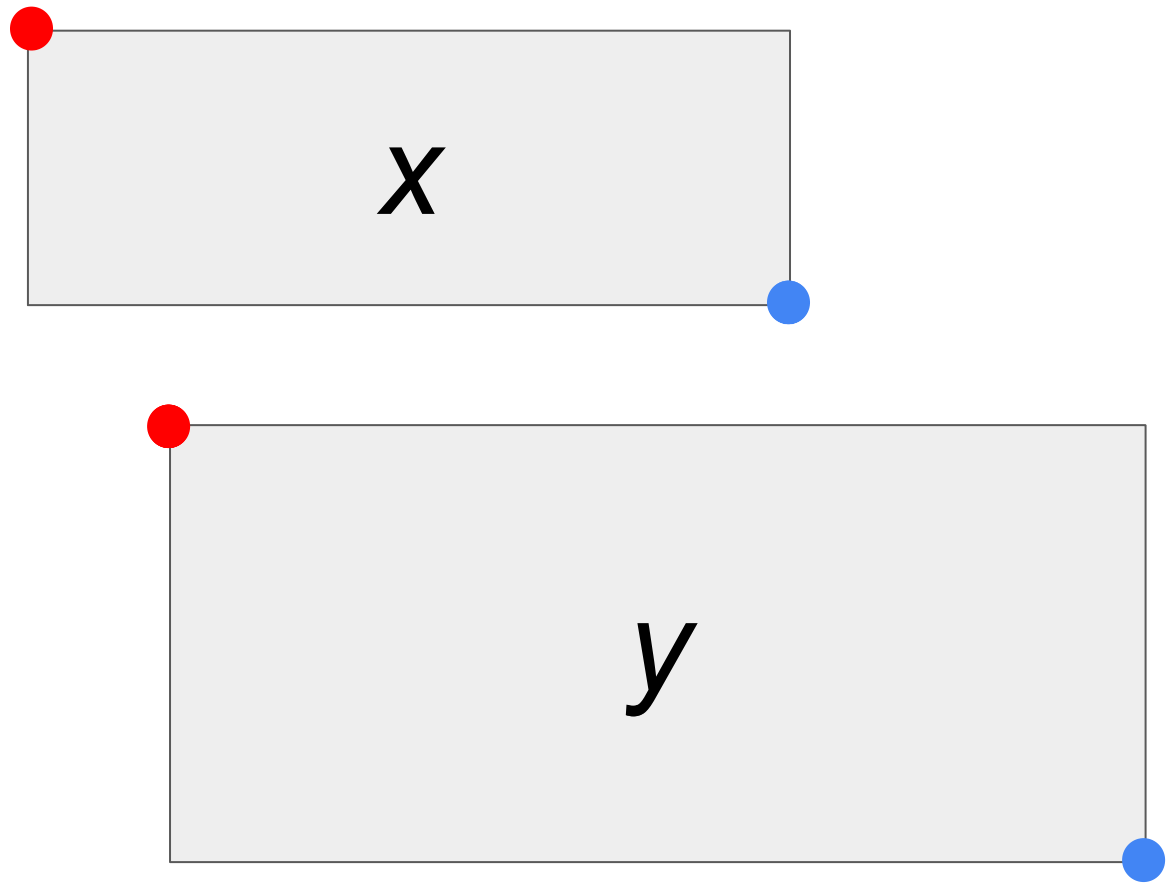
Learnable Fourier Feature Positional Encoding

Given an M-dimensional position: $x \in R^M$

Acquire D-dimensional Fourier features: $r_x = \frac{1}{\sqrt{D}}[\cos xW_r^T \parallel \sin xW_r^T]$

MLP Modulator: $PE_x = \phi(r_x, \theta)W_p$

Composability:



One group

`[(top, left, bottom, right)]`

Two groups

`[(top, left), (bottom, right)]`

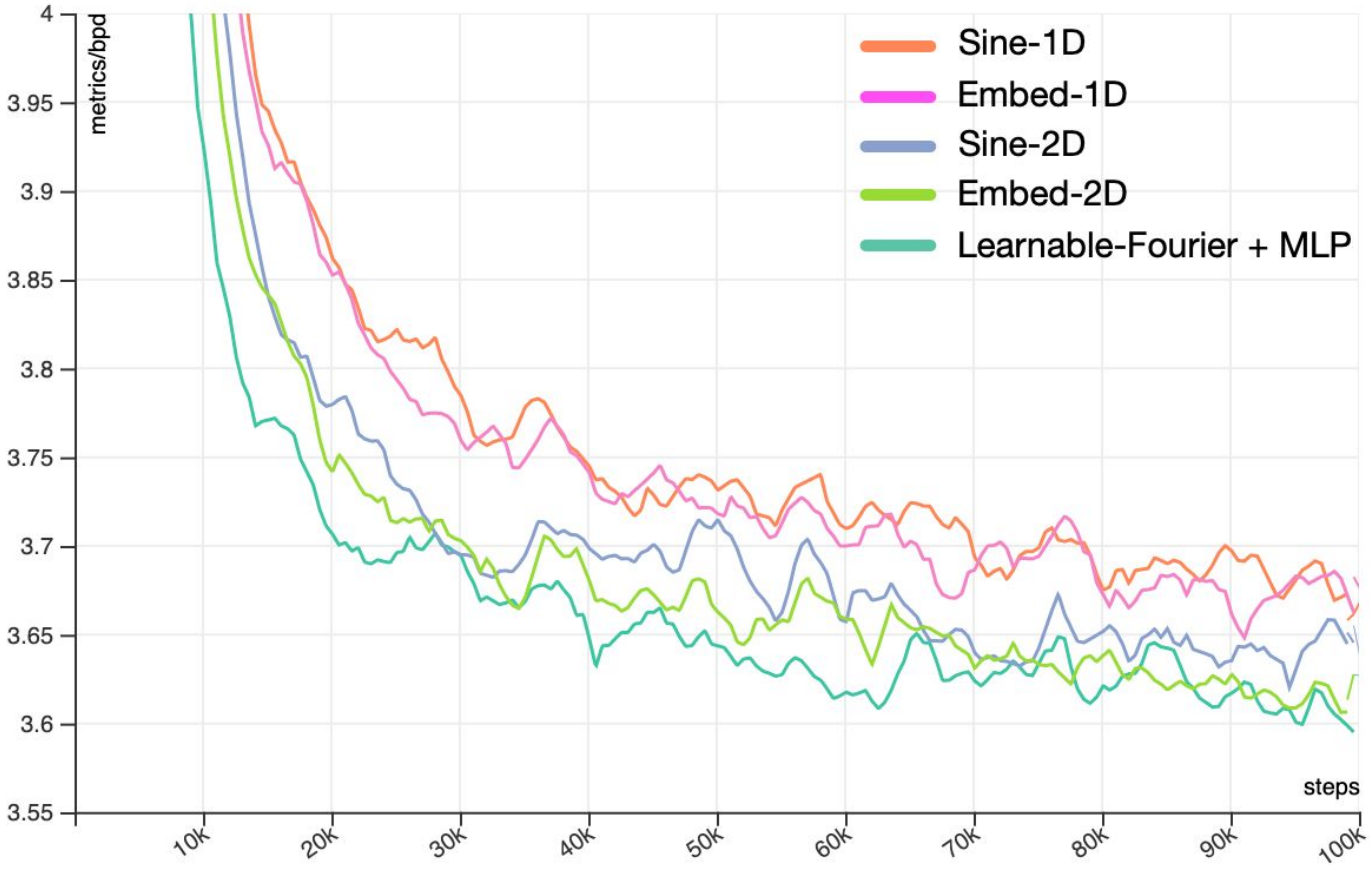
Experiments

- Image generation
- Object detection
- Image classification
- Widget captioning

Image Generation

Benchmark:

- Reformer on ImageNet64 [Kitaev et al. ICLR 2020]
- Images with 64x64 unique 2D pixel positions



Object Detection

Benchmark:

- DETR on MS COCO 2017 [Carion et al. ECCV 2020]
- Image feature maps with 42x42 unique 2D positions

Method	AP	AP_{50}	AP_{75}	AP_{small}	AP_{medium}	AP_{large}
Sine-2D	40.1	60.4	42.6	18.5	43.6	58.8
Embed-2D	39.3	59.8	41.4	18.7	42.5	57.5
MLP	40.0	60.3	42.2	18.6	43.7	58.1
Learnable-Fourier+MLP	**40.2**	**60.7**	**42.7**	**18.8**	**43.8**	**59.1**

Generalization on unseen image sizes

Method	AP	AP_{50}	AP_{75}	AP_{small}	AP_{medium}	AP_{large}
Sine-2D	38.9	59.6	40.9	17.5	42.5	57.5
Embed-2D	36.6	58.2	37.7	15.9	40.0	55.3
MLP	38.6	59.5	40.3	17.1	42.1	57.1
Learnable-Fourier+MLP	**39.5**	**60.0**	**41.6**	**18.9**	**43.0**	**58.0**

Image Classification

Benchmark:

- ViT-B/16 on ImageNet and JFT(300M) [Dosovitskiy et al. ICLR 2021]
- Image feature maps with 14x14 unique 2D positions

Trained & validated on ImageNet
Embed-1D: Precision@1=73.6%
Learnable-Fourier+MLP: Precision@1=74.5%

Pretrained on JFT and 5-Shot Learning on ImageNet
Embed-1D: 64.206%
Learnable-Fourier+MLP: 74.732%

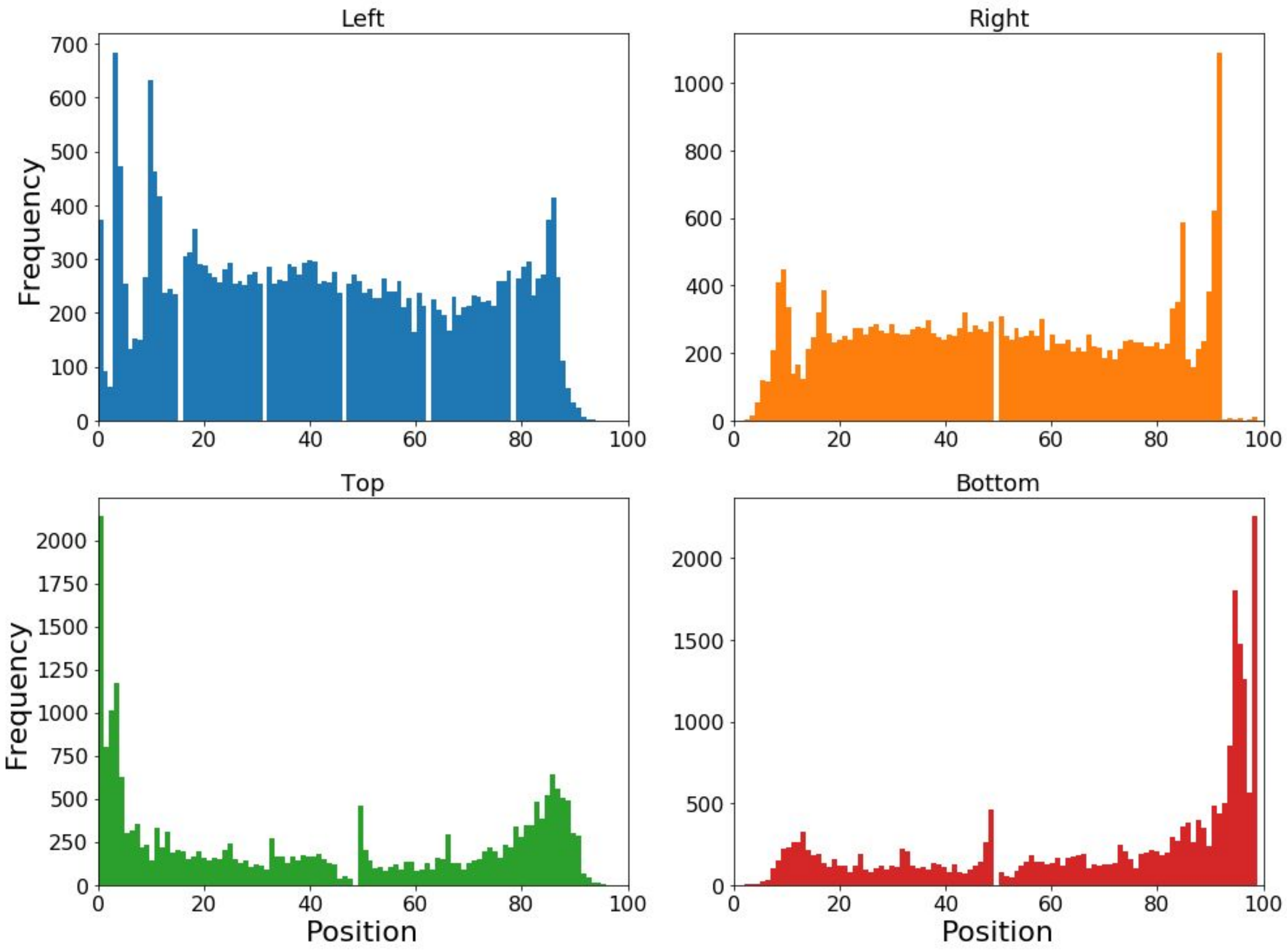
Widget Captioning

Benchmark:

- Widget captioning [Li et al. EMNLP 2020]
- Sparse spatial UI layouts with 100x100x100x100 4D positions

Positional Embedding	BLEU-1	BLEU-2	ROUGE	CIDEr	METOER	SPICE
SOTA [20]	44.9	32.2	44.7	97.0	31.7	17.6
Embed-4D	45.2	31.9	45.0	97.0	31.7	17.3
MLP	34.0	23.5	33.7	70.3	23.7	10.2
Sine-4D	44.9	31.9	43.9	94.9	31.0	16.7
Learnable-Fourier-2/2	44.9	31.6	44.3	95.3	31.6	17.7
Fixed-Fourier+MLP-1/4	45.0	32.1	44.2	95.4	31.2	17.1
Fixed-Fourier+MLP-2/2	46.1	32.5	45.8	100.2	32.5	18.4
Fixed-Fourier+MLP-4/1	45.5	32.1	45.1	97.2	31.7	17.6
Learnable-Fourier+MLP-1/4	45.6	32.7	45.2	99.1	32.2	17.1
Learnable-Fourier+MLP-2/2	46.1	32.7	45.9	98.0	**32.6**	**17.9**
Learnable-Fourier+MLP-4/1	**46.8**	**33.4**	**46.1**	**100.7**	32.4	17.8

Performance on Unseen Positions in Widget Captioning



Positional Embedding	Seen CIDEr	Unseen CIDEr
Embed-4D	**123.4**	78.5
Sine-4D	121.3	76.4
Learnable-Fourier+MLP-4/1	**123.4**	**82.2**

Conclusions

- A novel approach for positional encoding based on learnable Fourier features.
  - Positions as continuous-valued vectors
  - Bringing in inductive bias such as L2 distances
  - Learnable & composable
- Extensive experiments based on a range of multi-dimensional spatial tasks.
  - Image generation
  - Object detection
  - Image classification
  - Widget captioning

Learnable Fourier Features for Multi-Dimensional Spatial Positional Encoding

Yang Li, Si Si, Gang Li, Cho-Jui Hsieh*, Samy Bengio

Google Research UCLA*