PAT 464/564 (Winter 2026)

Generative AI for Music & Audio Creation

Lecture 10: Transformers

Instructor: Hao-Wen Dong

UNIVERSITY OF MICHIGAN

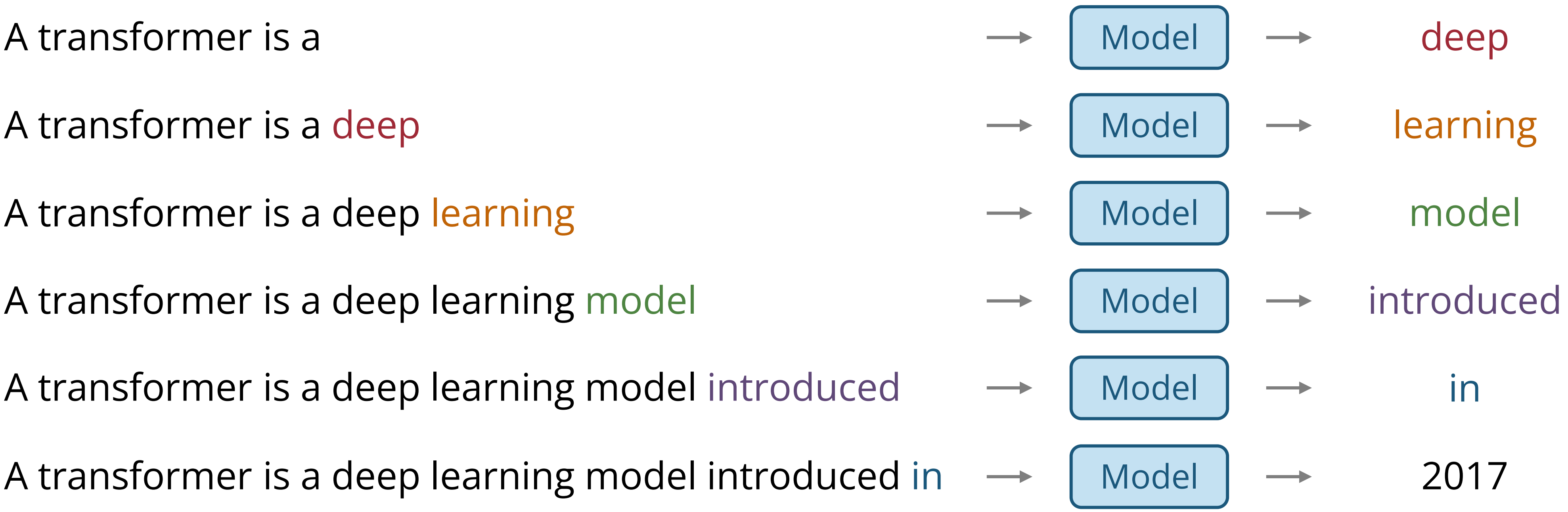
Representative Types of Deep Generative Models

- **Deep autoregressive models**
  - Recurrent neural network (RNN)
  - Long short-term memory (LSTM)
  - Transformer model **Today's topic!**
- **Deep latent variable models**
  - Generative adversarial network (GAN)
  - Variational autoencoder (VAE)
  - Diffusion model
  - Flow-based model
- *And many others…*

Deep Autoregressive Models

Deep Autoregressive Models

- **Intuition**: Decompose the generation of a sequence into generating one item after another



A transformer is a → Model → deep

A transformer is a deep → Model → learning

A transformer is a deep learning → Model → model

A transformer is a deep learning model → Model → introduced

A transformer is a deep learning model introduced → Model → in

A transformer is a deep learning model introduced in → Model → 2017

Deep Autoregressive Models

- **Intuition**: Decompose the generation of a sequence into generating one item after another

$$P(\underset{\swarrow \atop \text{Next word}}{x_i} \mid \underbrace{x_1, x_2, \ldots, x_{i-1}}_{\text{Previous words}})$$

$P($ electrical $\mid$ A transformer is a $)$ ⬆

$P($ character $\mid$ A transformer is a $)$ ⬆

$P($ gene $\mid$ A transformer is a $)$ ⬆

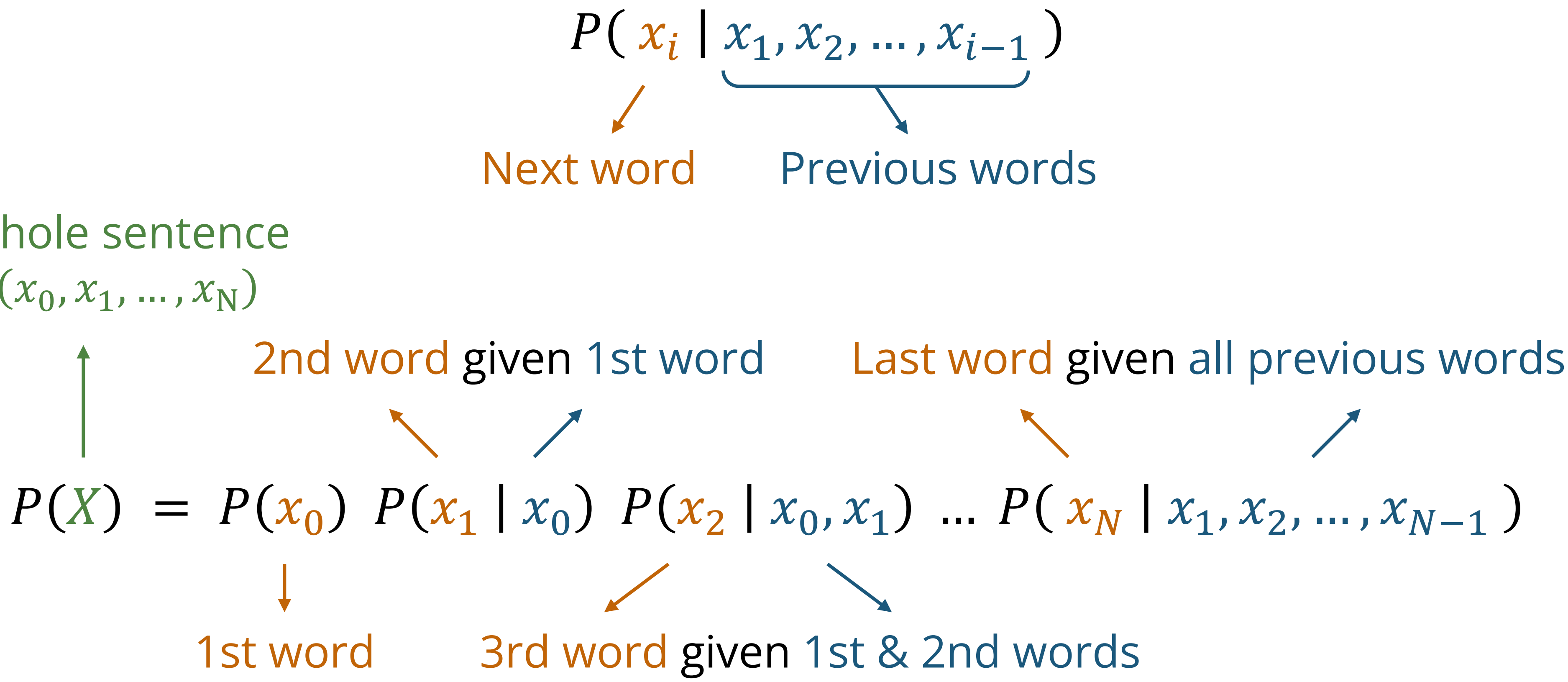
$P($ model $\mid$ A transformer is a $)$ ⬆

$P($ food $\mid$ A transformer is a $)$ ⬇

$P($ musical $\mid$ A transformer is a $)$ ⬇

Deep Autoregressive Models

- **Intuition**: Decompose the generation of a sequence into generating one item after another



$$P(\, x_i \mid \underbrace{x_1, x_2, \ldots, x_{i-1}} \,)$$

Next word

Previous words

The whole sentence

$X = (x_0, x_1, \ldots, x_{\mathrm{N}})$

2nd word given 1st word

Last word given all previous words

$$P(X) = P(x_0)\ P(x_1 \mid x_0)\ P(x_2 \mid x_0, x_1)\ \ldots\ P(\, x_N \mid x_1, x_2, \ldots, x_{N-1} \,)$$

1st word

3rd word given 1st & 2nd words

Deep Autoregressive Models

- **Intuition**: Decompose the generation of a sequence into generating one item after another

What we want the model to learn!

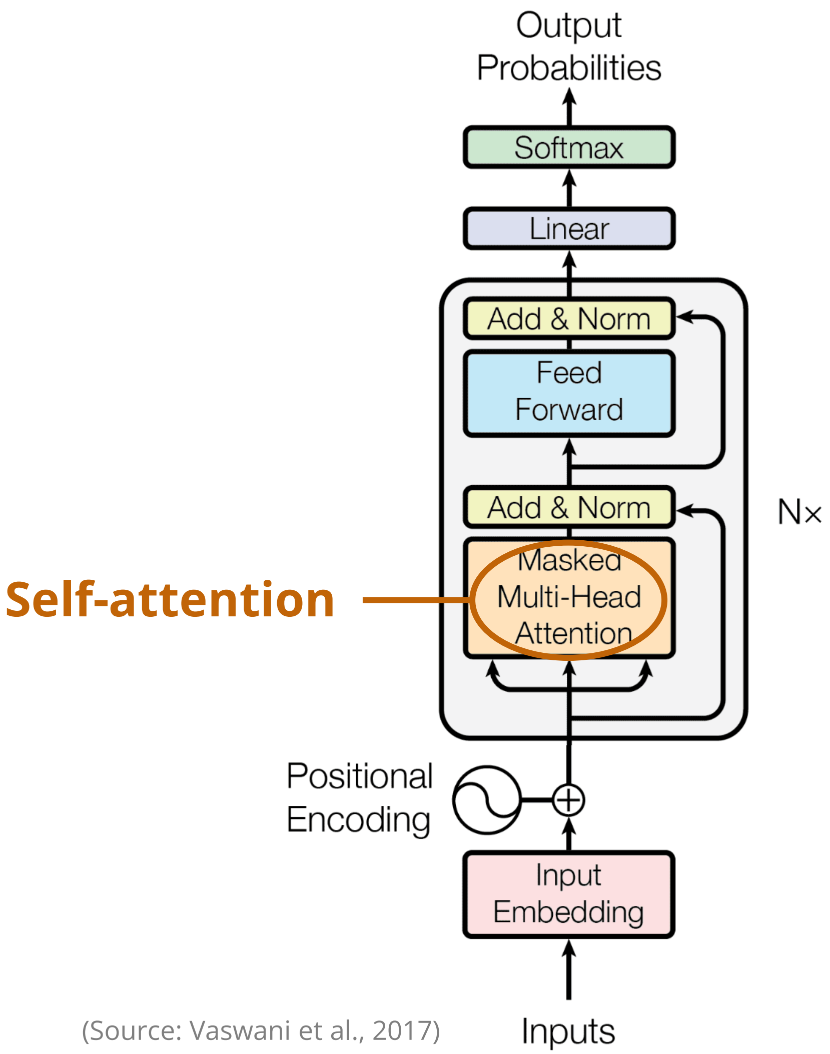
$$P(X) = \boxed{P(x_0)}\,\boxed{P(x_1 \mid x_0)}\,\boxed{P(x_2 \mid x_0, x_1)} \dots \boxed{P(x_N \mid x_1, x_2, \dots, x_{N-1})}$$

$$= P(x_0) \prod_{i=1}^{N} P(x_i \mid x_1, x_2, \dots, x_{i-1})$$

Transformers

What is a Transformer? (Vaswani et al., 2017)

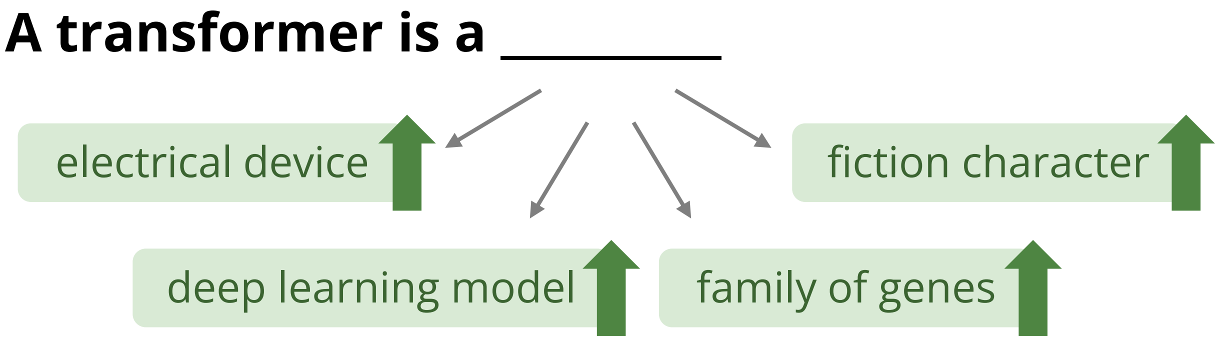
- A type of neural networks that use the **self-attention mechanism**

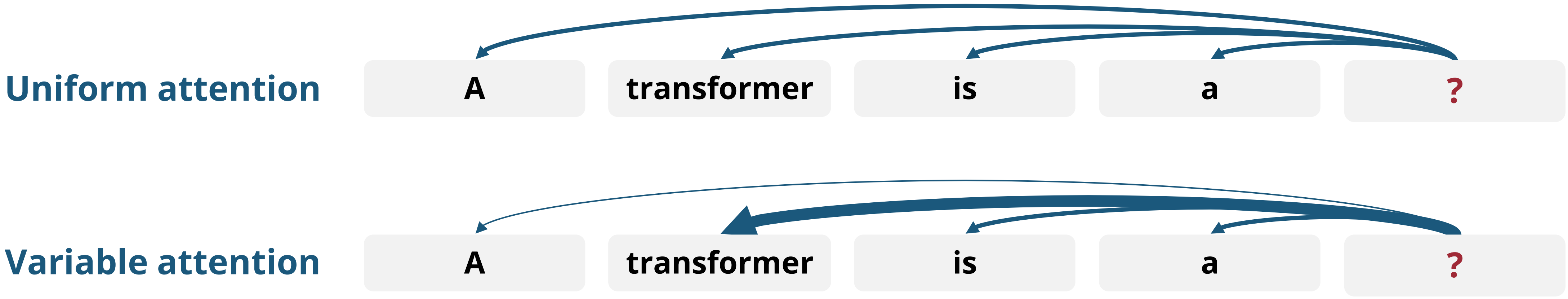


(Source: Vaswani et al., 2017)

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," *NeurIPS*, 2017.

Self-attention Mechanism (Cheng et al., 2016)

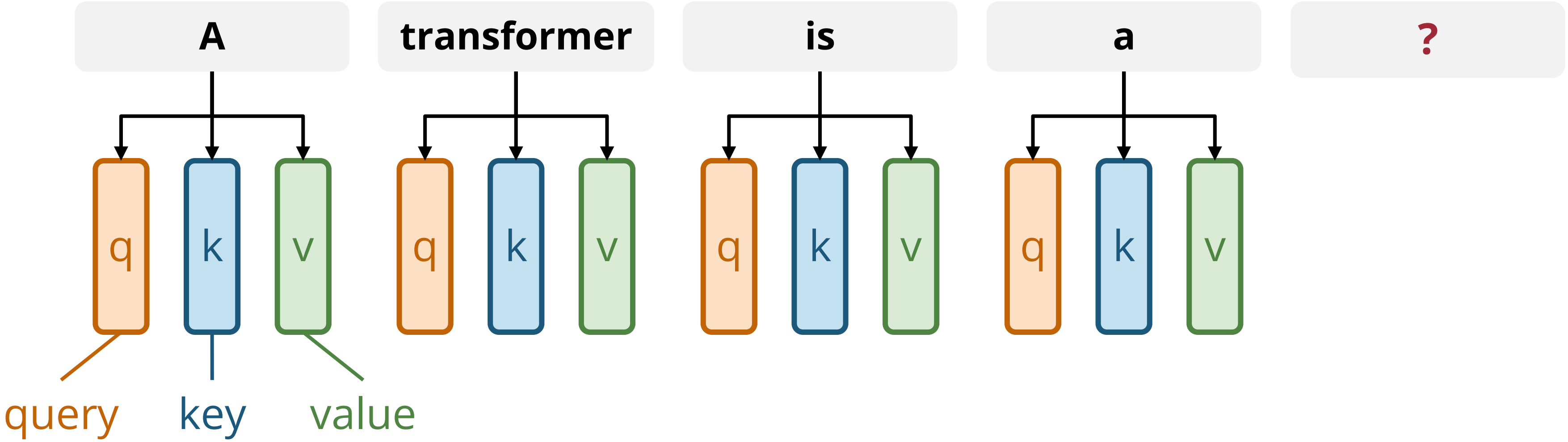




Transformers learn what to attend to from big data!

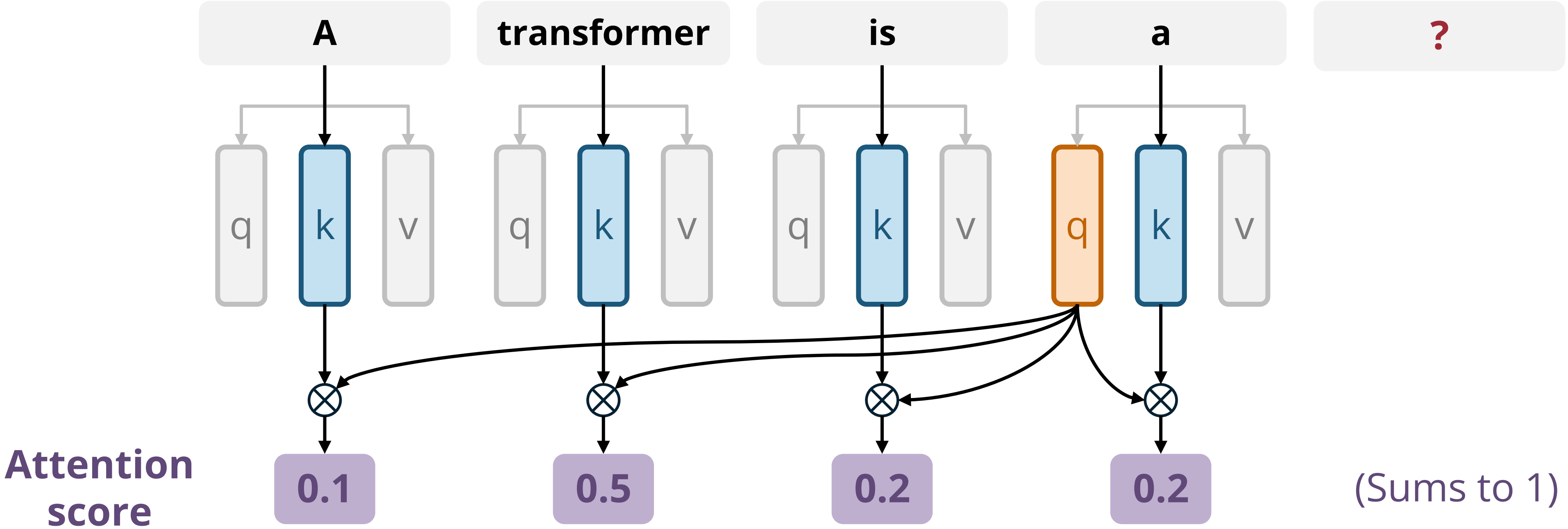
Jianpeng Cheng, Li Dong, and Mirella Lapata, "Long Short-Term Memory-Networks for Machine Reading," *EMNLP*, 2016.

Demystifying Transformers (Vaswani et al., 2017)



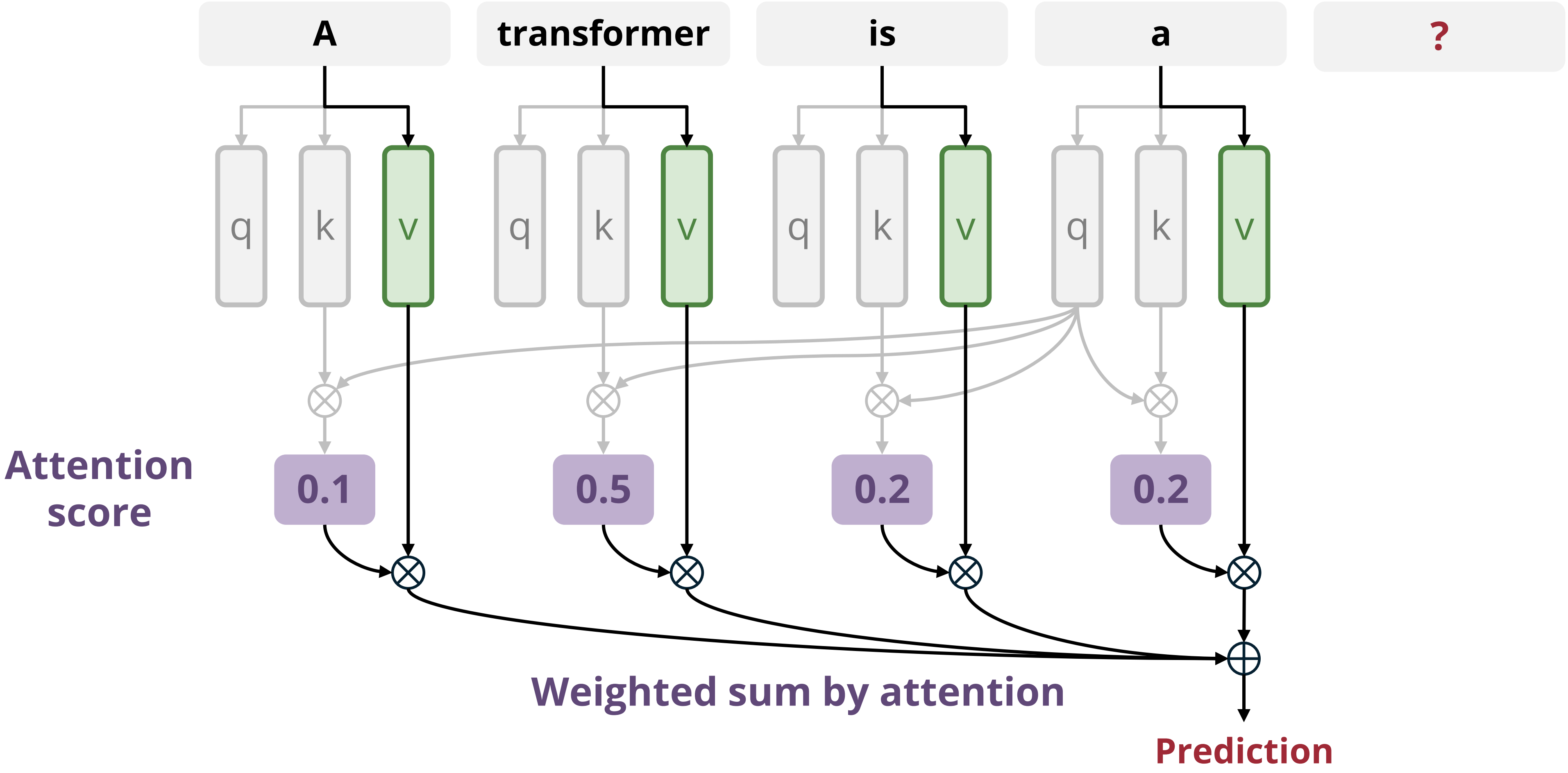
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention Is All You Need,” *NeurIPS*, 2017.

Demystifying Transformers (Vaswani et al., 2017)



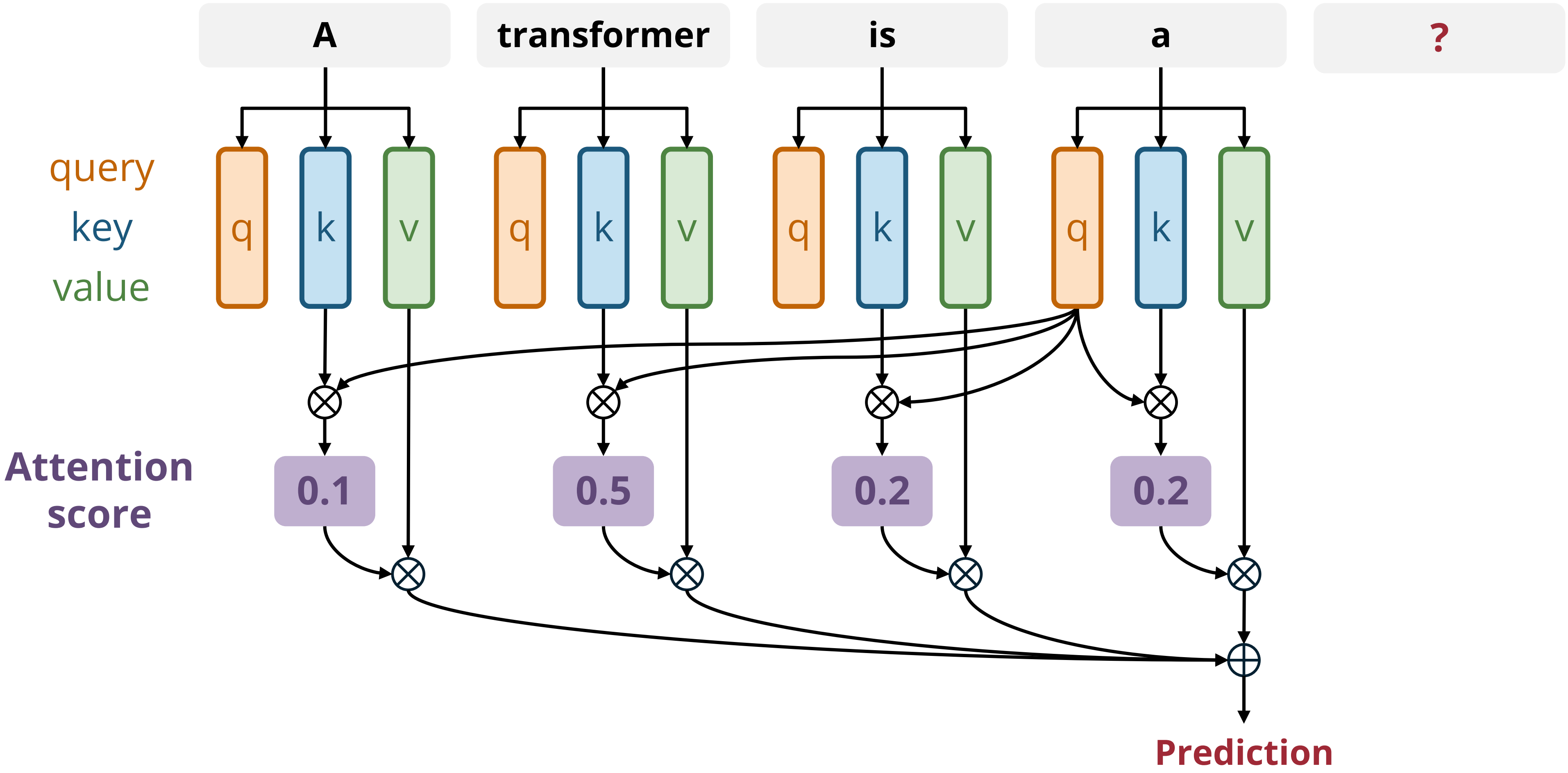
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," *NeurIPS*, 2017.

Demystifying Transformers (Vaswani et al., 2017)



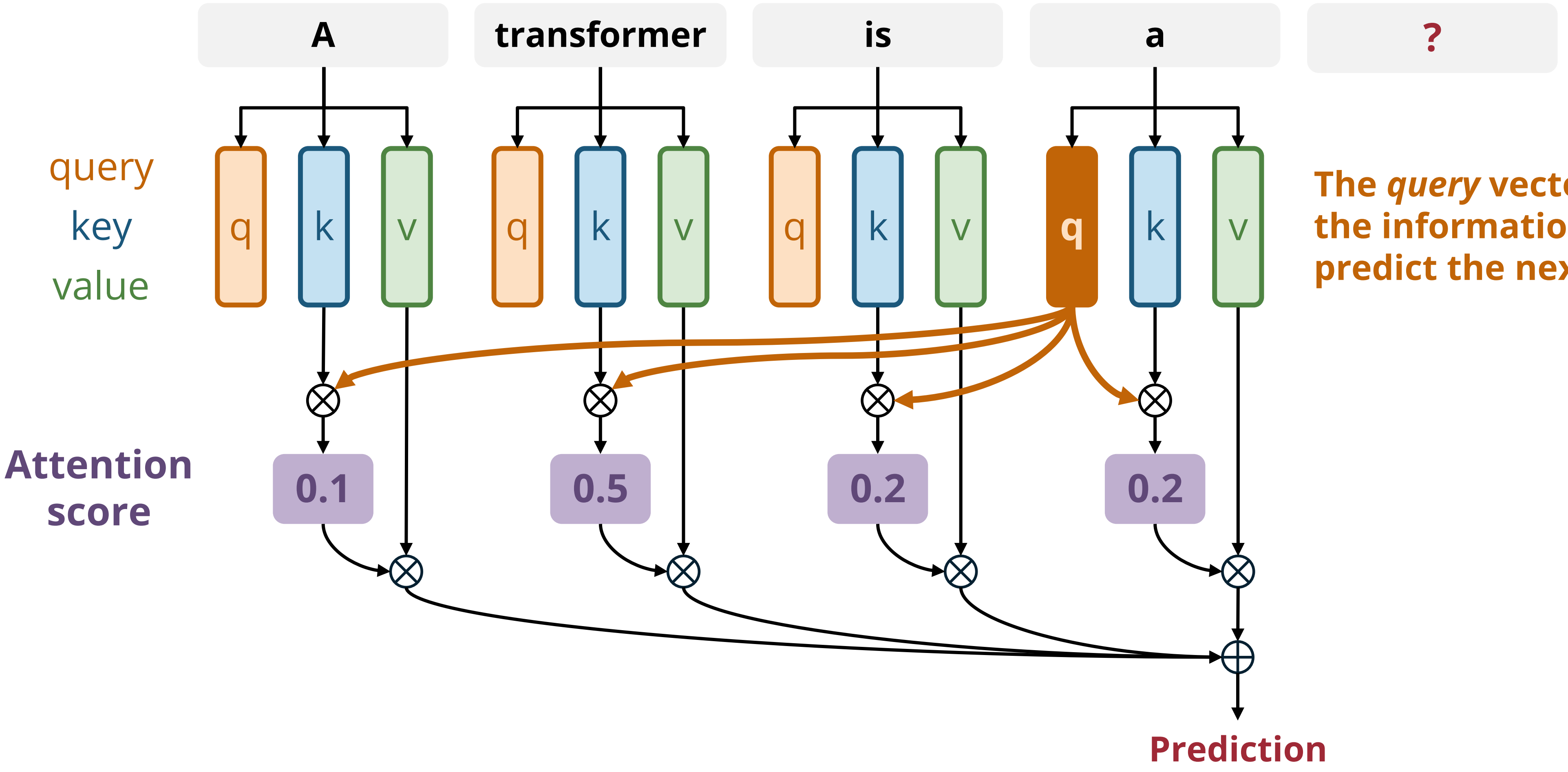
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," *NeurIPS*, 2017.

Demystifying Transformers (Vaswani et al., 2017)



Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," *NeurIPS*, 2017.

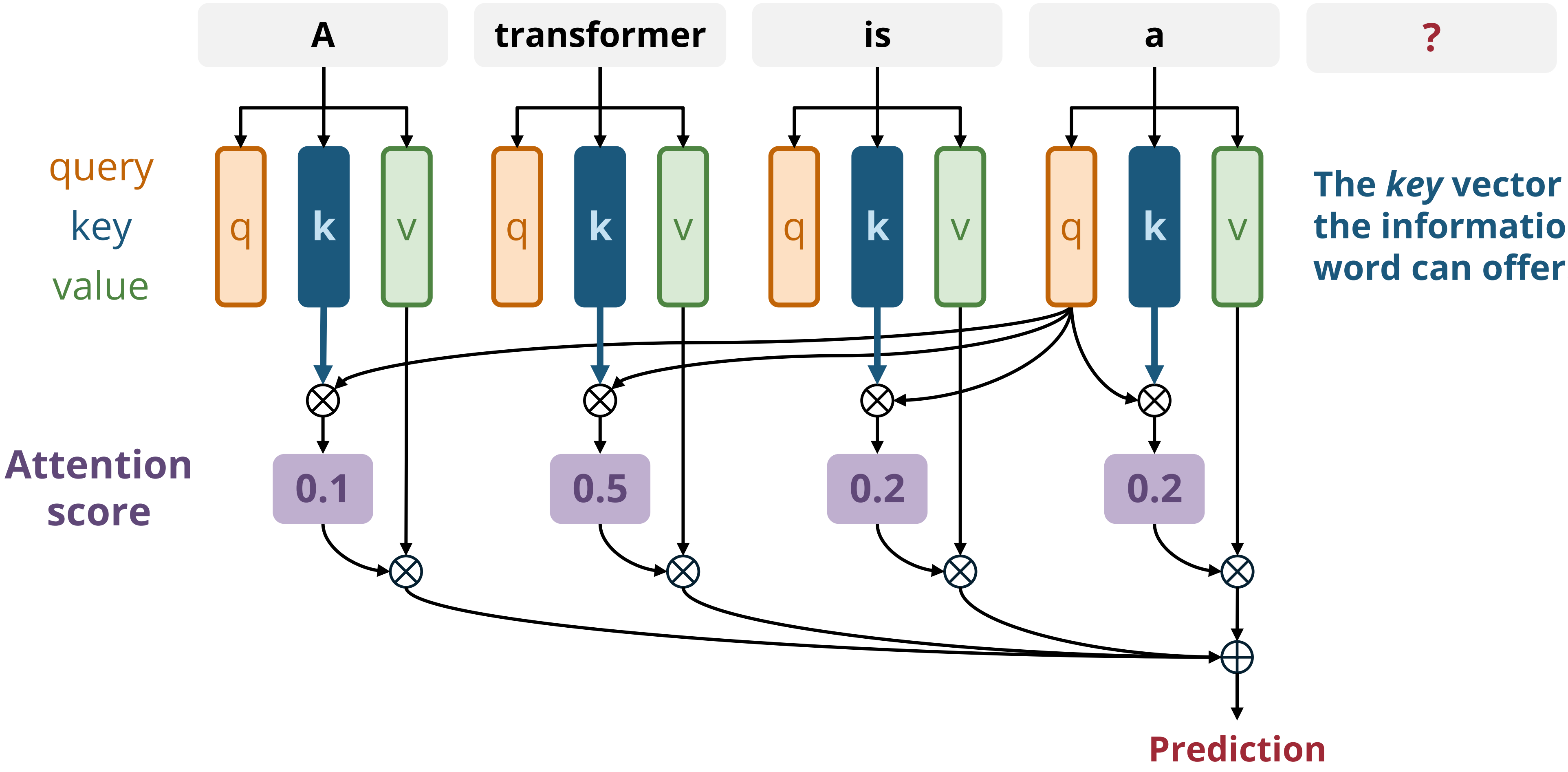
Demystifying Transformers (Vaswani et al., 2017)



The *query* vector captures the information needed to predict the next word.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," *NeurIPS*, 2017.

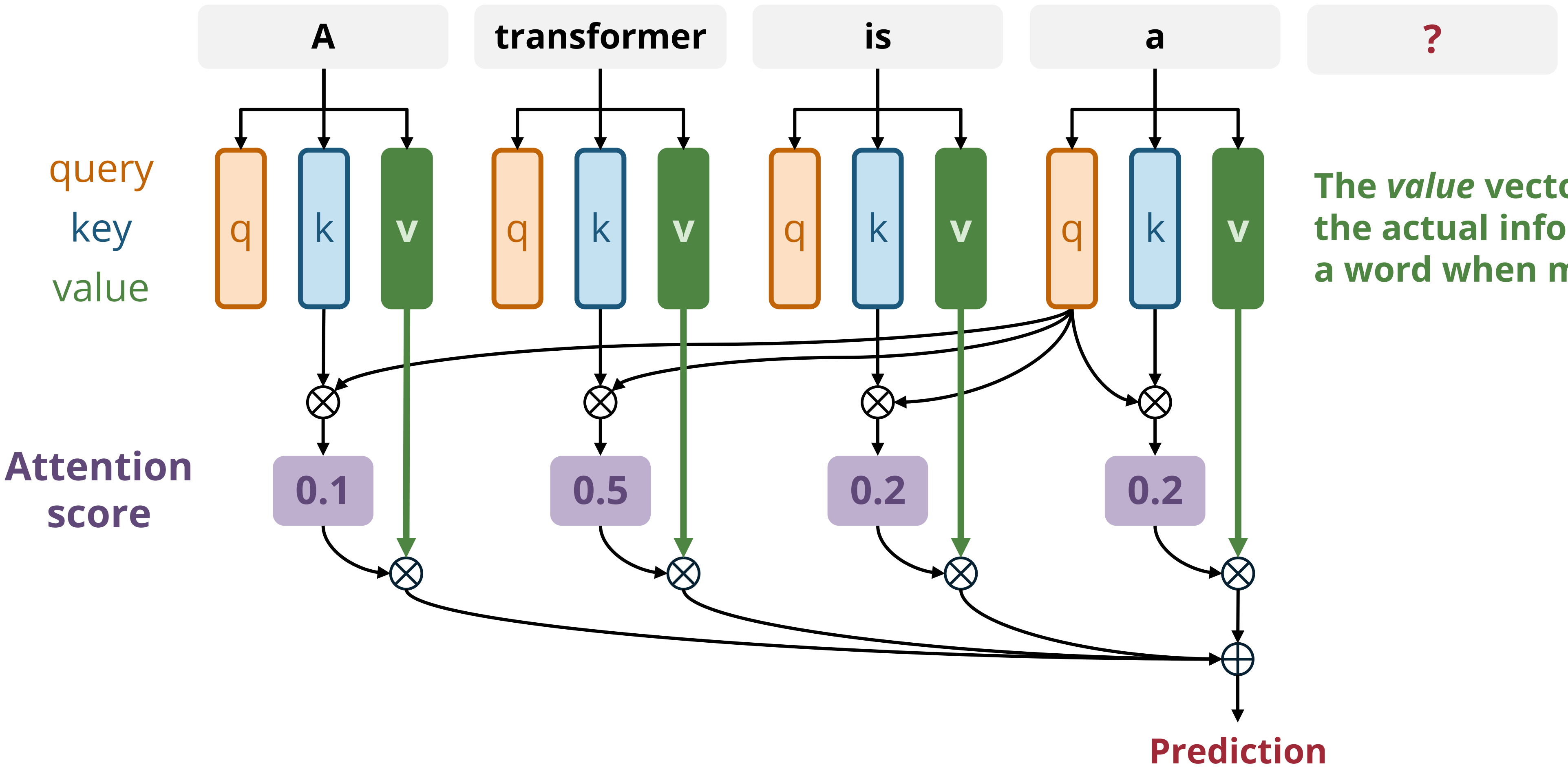
Demystifying Transformers (Vaswani et al., 2017)



The *key* vector captures the information that a word can offer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," *NeurIPS*, 2017.

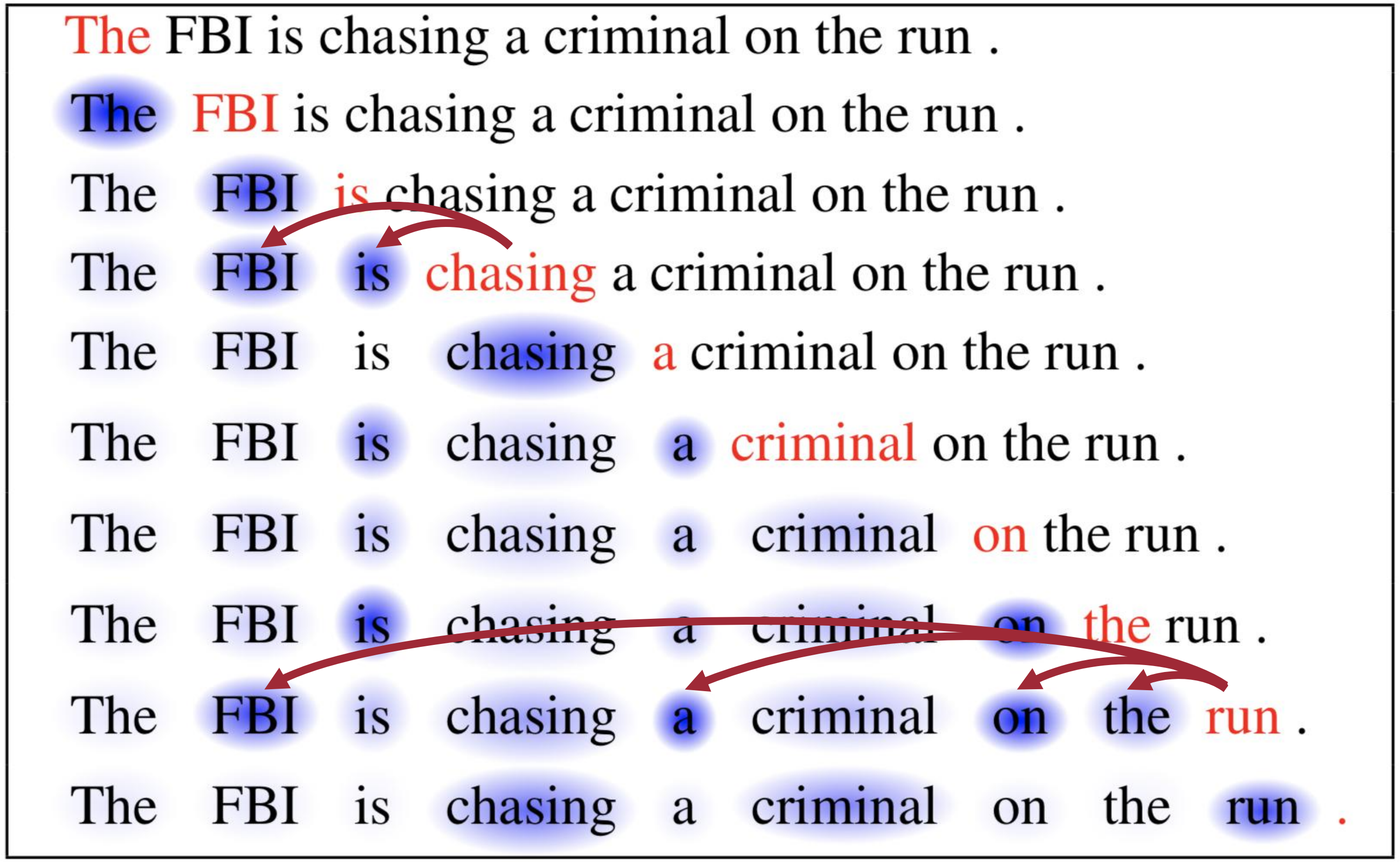
Demystifying Transformers (Vaswani et al., 2017)



The *value* vector captures the actual information of a word when matched.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," *NeurIPS*, 2017.

Why Self-Attention Mechanism?



The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

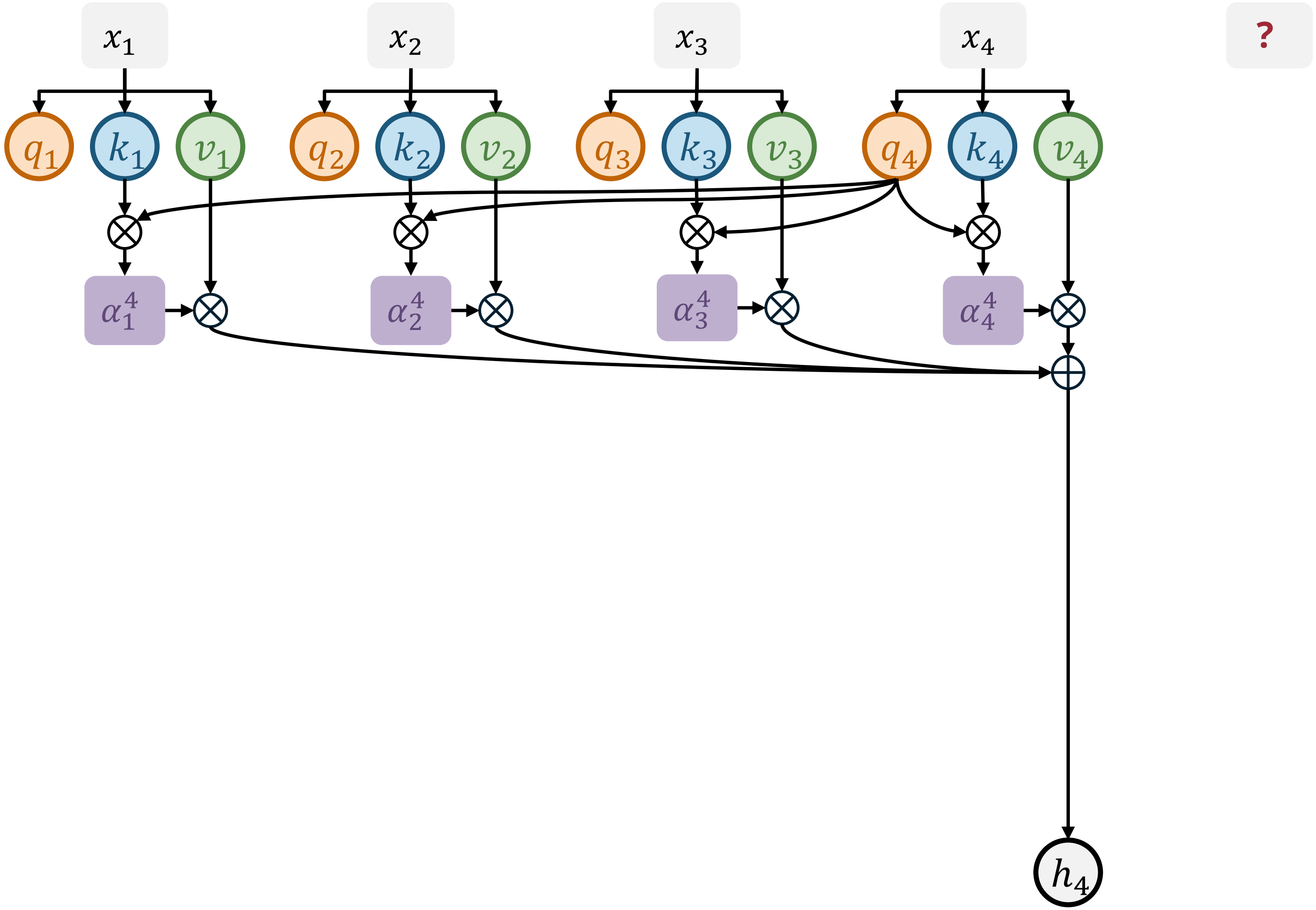
The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

(Source: Cheng et al., 2016)

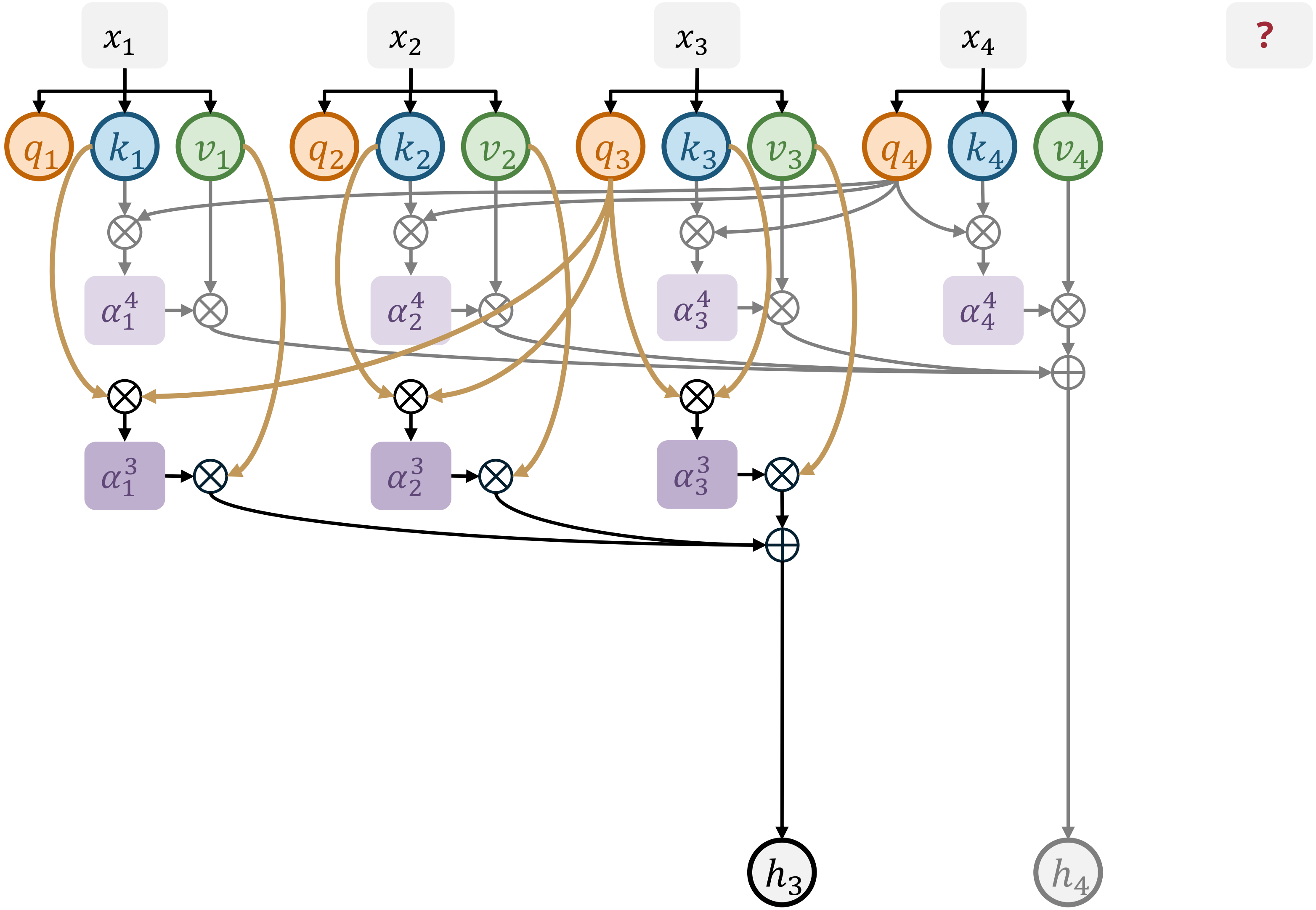
Jianpeng Cheng, Li Dong, and Mirella Lapata, "Long Short-Term Memory-Networks for Machine Reading," *EMNLP*, 2016.

Demystifying Transformer Layers (Vaswani et al., 2017)



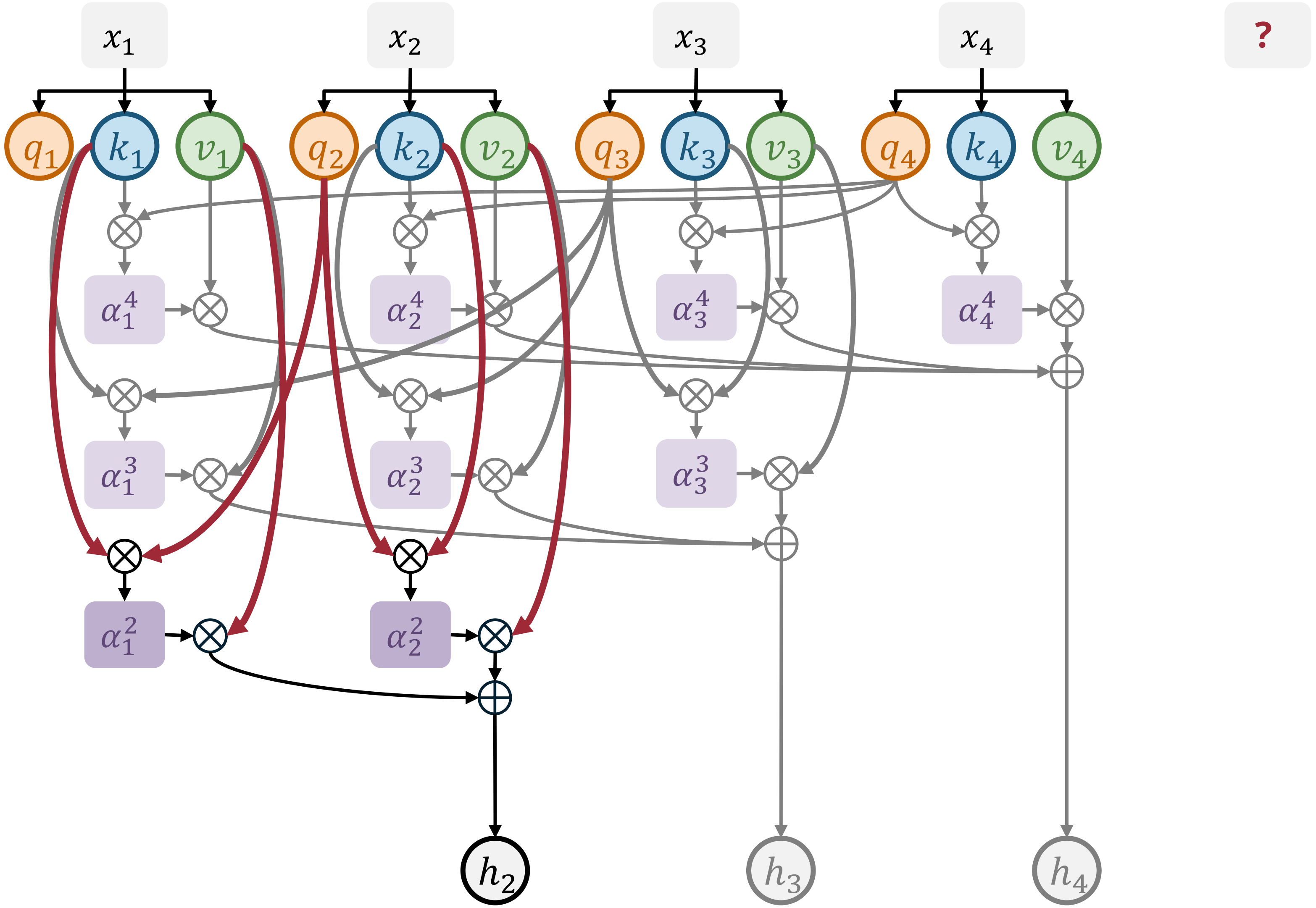
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention Is All You Need,” *NeurIPS*, 2017.

Demystifying Transformers (Vaswani et al., 2017)



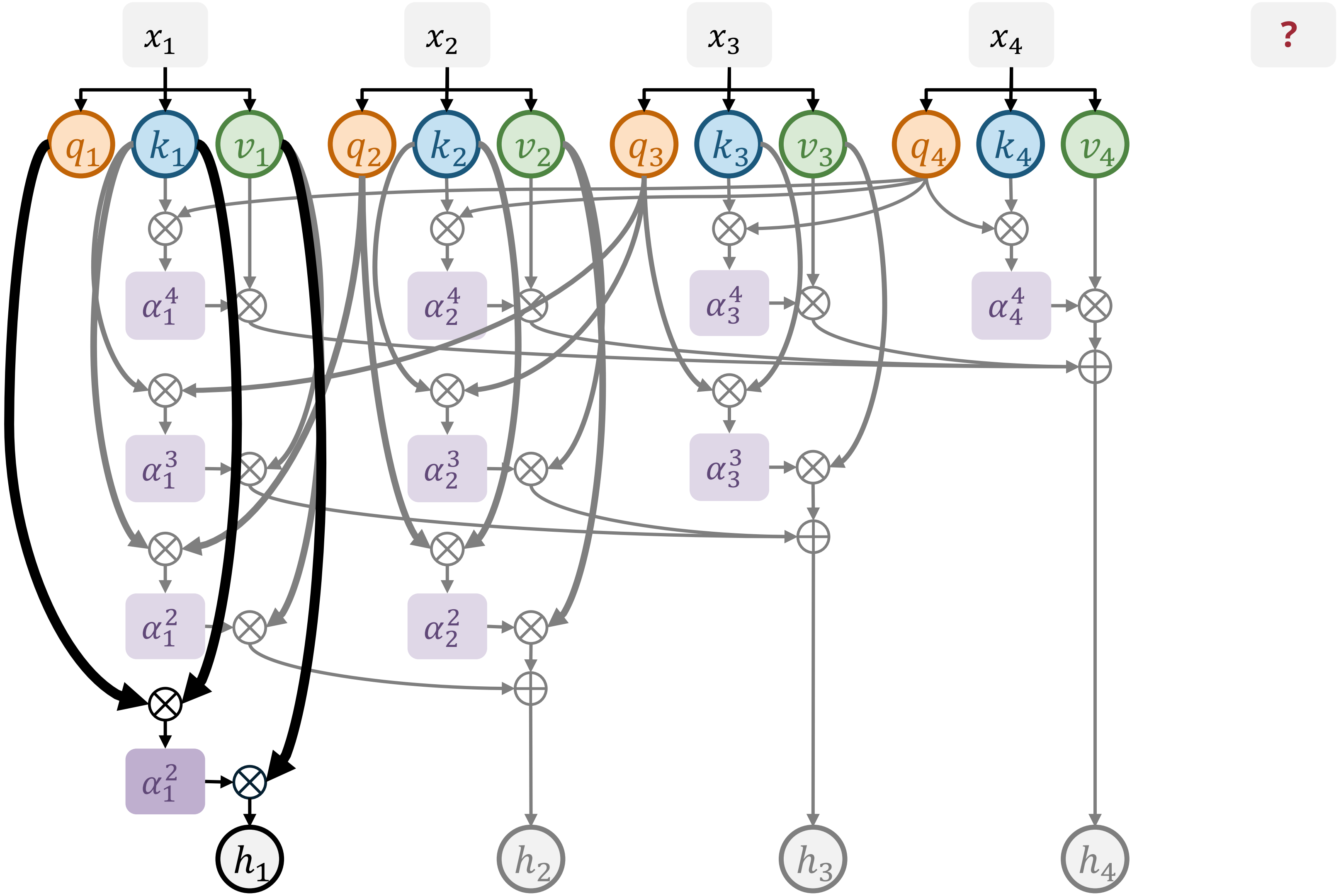
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," *NeurIPS*, 2017.

Demystifying Transformers (Vaswani et al., 2017)



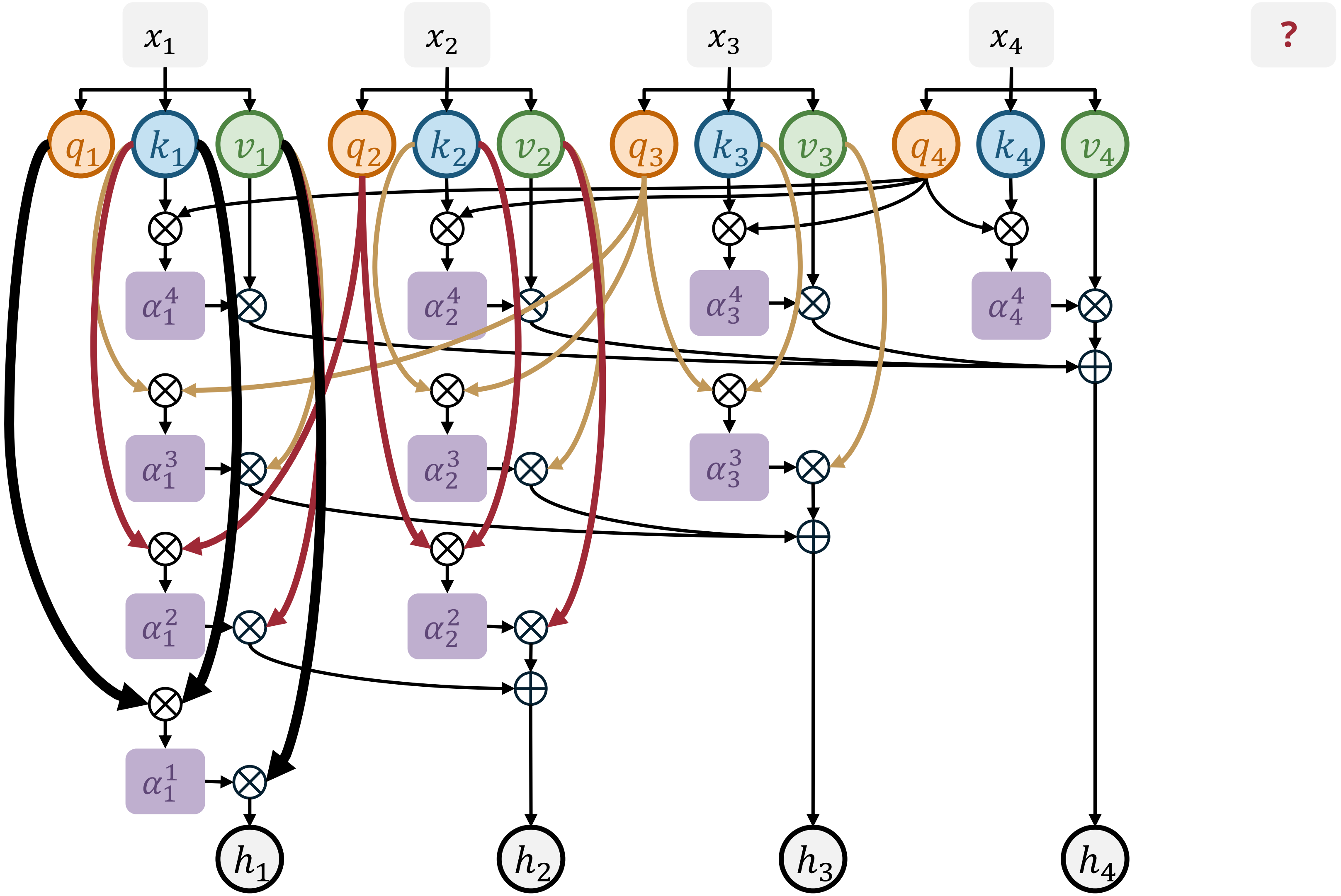
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention Is All You Need,” *NeurIPS*, 2017.

Demystifying Transformers (Vaswani et al., 2017)



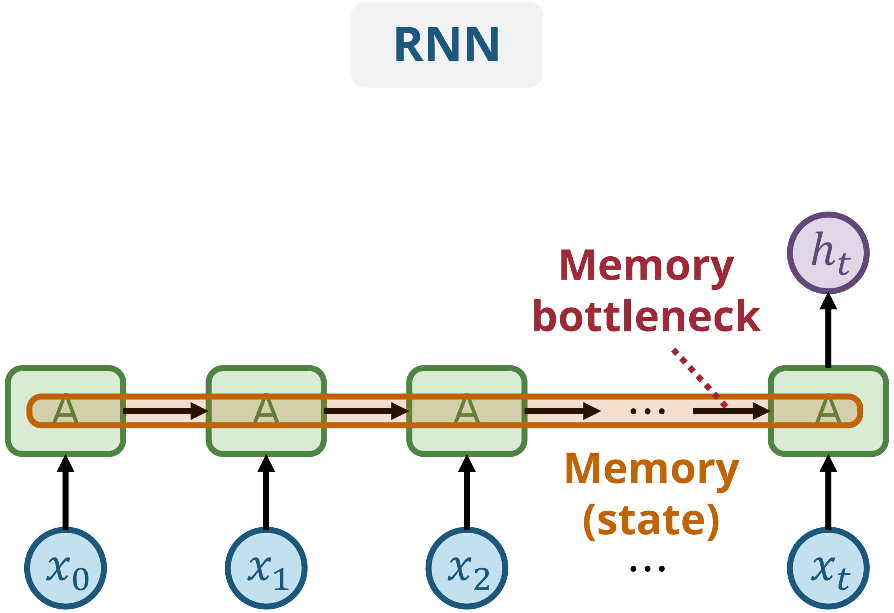
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention Is All You Need,” *NeurIPS*, 2017.

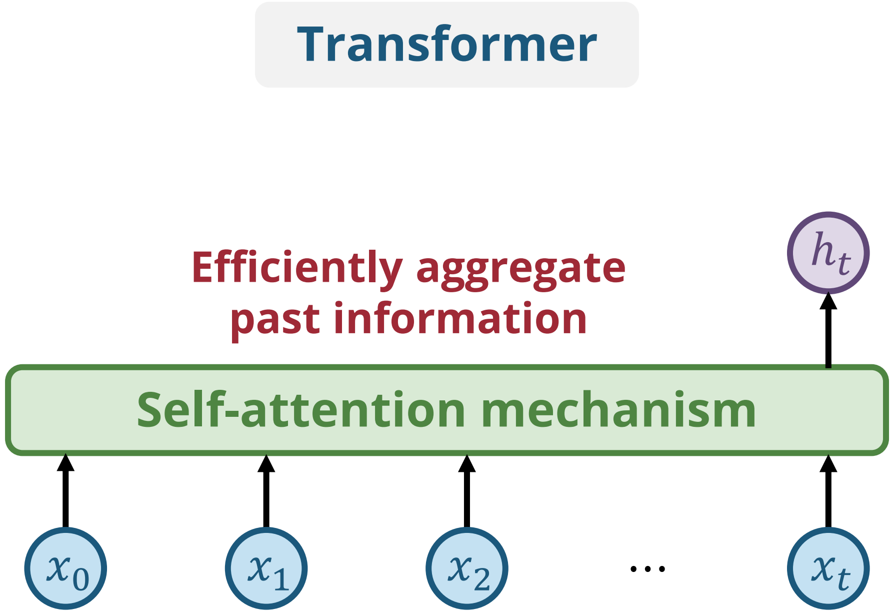
Demystifying Transformers (Vaswani et al., 2017)



Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," *NeurIPS*, 2017.

RNN vs. Transformer

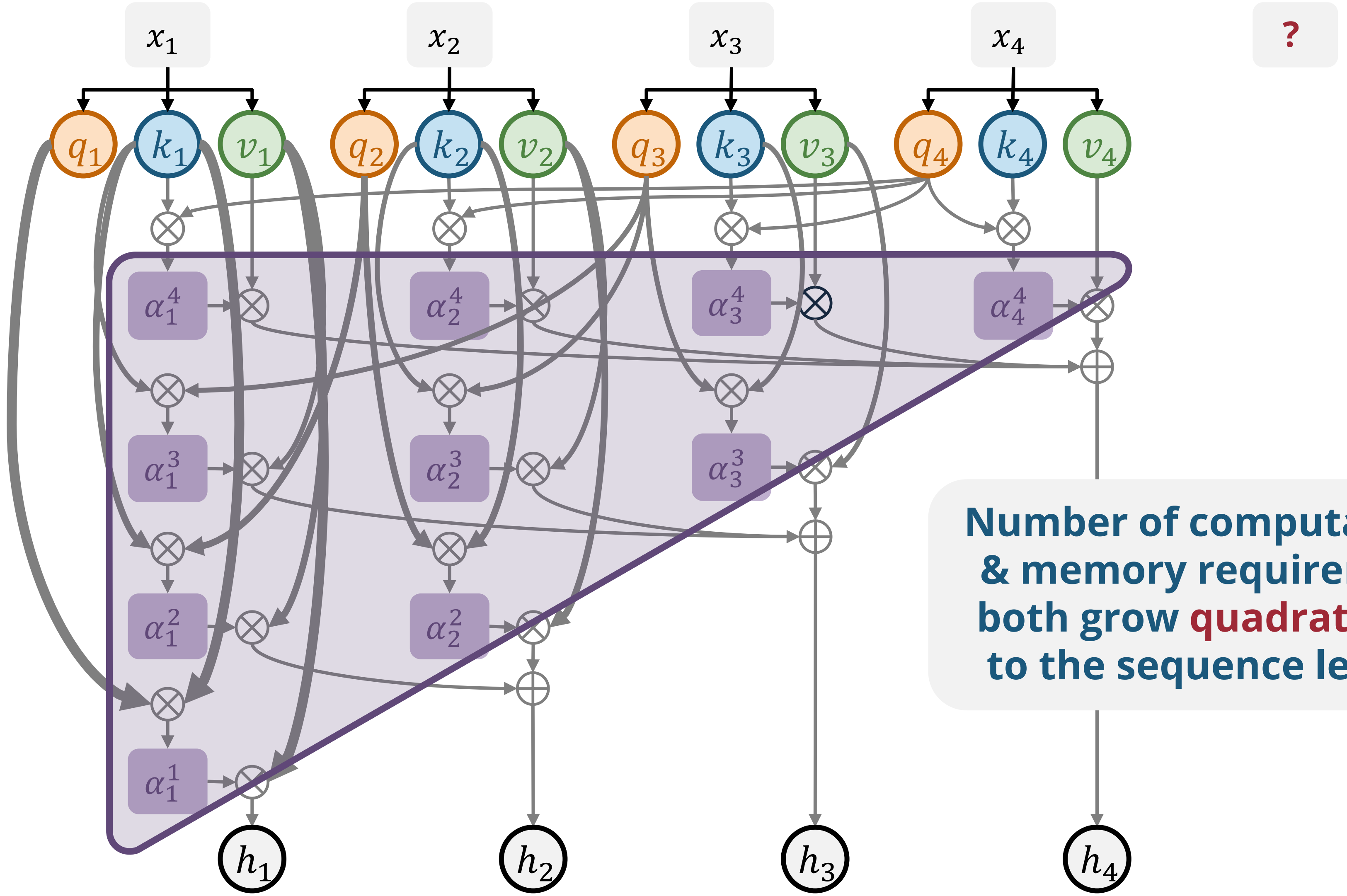




Pros: Requires less GPU memory
Cons: Memory bottleneck

Pros: Alleviate memory bottleneck constraints
Cons: Requires more GPU memory

Demystifying Transformers (Vaswani et al., 2017)

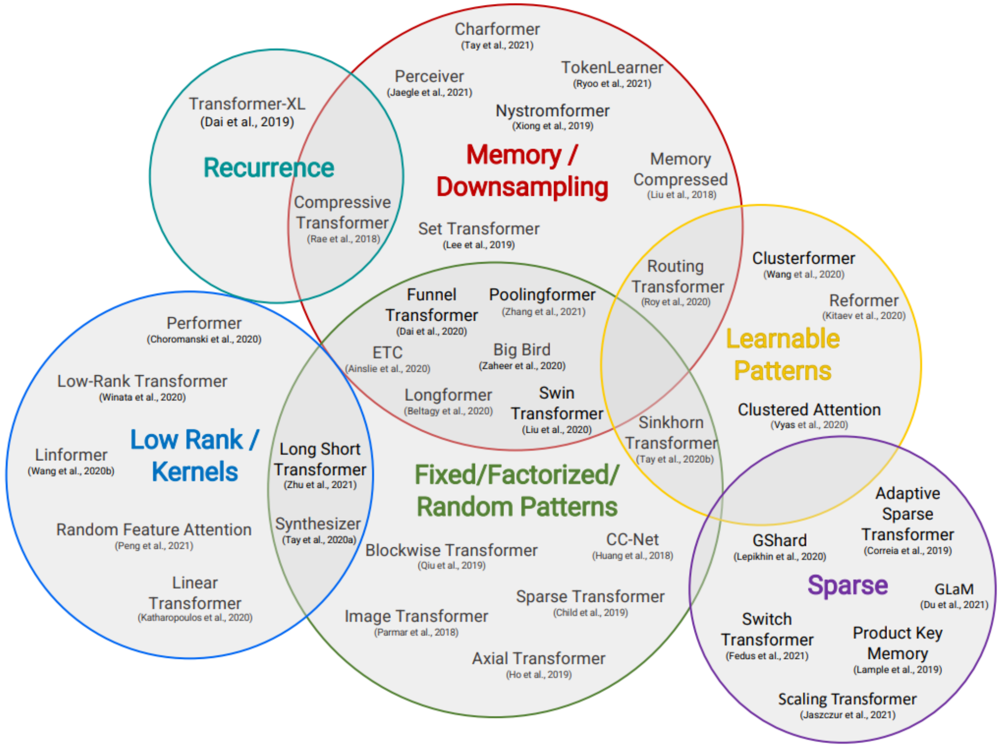


Number of computations & memory requirement both grow quadratically to the sequence length

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," *NeurIPS*, 2017.

Efficient Transformers

- The **memory requirement for self-attention** grows **quadratically**!
- There are many efficient transformer variants
  - Transformer-XL
  - Linear Transformer
  - Performer
  - Longformer
  - Reformer
  - Swin Transformer
  - … *just to name a few*



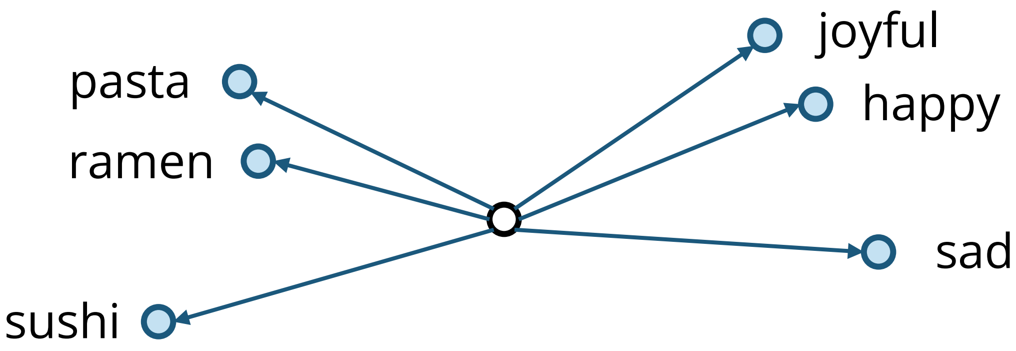
Charformer (Tay et al., 2021)
Perceiver (Jaegle et al., 2021)
TokenLearner (Ryoo et al., 2021)
Transformer-XL (Dai et al., 2019)
Nystromformer (Xiong et al., 2019)
Recurrence
Memory / Downsampling
Memory Compressed (Liu et al., 2018)
Compressive Transformer (Rae et al., 2018)
Set Transformer (Lee et al., 2019)
Routing Transformer (Roy et al., 2020)
Clusterformer (Wang et al., 2020)
Funnel Transformer (Dai et al., 2020)
Poolingformer (Zhang et al., 2021)
Reformer (Kitaev et al., 2020)
Performer (Choromanski et al., 2020)
Learnable Patterns
ETC (Ainslie et al., 2020)
Big Bird (Zaheer et al., 2020)
Low-Rank Transformer (Winata et al., 2020)
Longformer (Beltagy et al., 2020)
Swin Transformer (Liu et al., 2020)
Clustered Attention (Vyas et al., 2020)
Sinkhorn Transformer (Tay et al., 2020b)
Linformer (Wang et al., 2020b)
Low Rank / Kernels
Long Short Transformer (Zhu et al., 2021)
Fixed/Factorized/ Random Patterns
Adaptive Sparse Transformer (Correia et al., 2019)
Random Feature Attention (Peng et al., 2021)
Synthesizer (Tay et al., 2020a)
CC-Net (Huang et al., 2018)
GShard (Lepikhin et al., 2020)
Blockwise Transformer (Qiu et al., 2019)
Sparse
GLaM (Du et al., 2021)
Linear Transformer (Katharopoulos et al., 2020)
Sparse Transformer (Child et al., 2019)
Image Transformer (Parmar et al., 2018)
Switch Transformer (Fedus et al., 2021)
Product Key Memory (Lample et al., 2019)
Axial Transformer (Ho et al., 2019)
Scaling Transformer (Jaszczur et al., 2021)

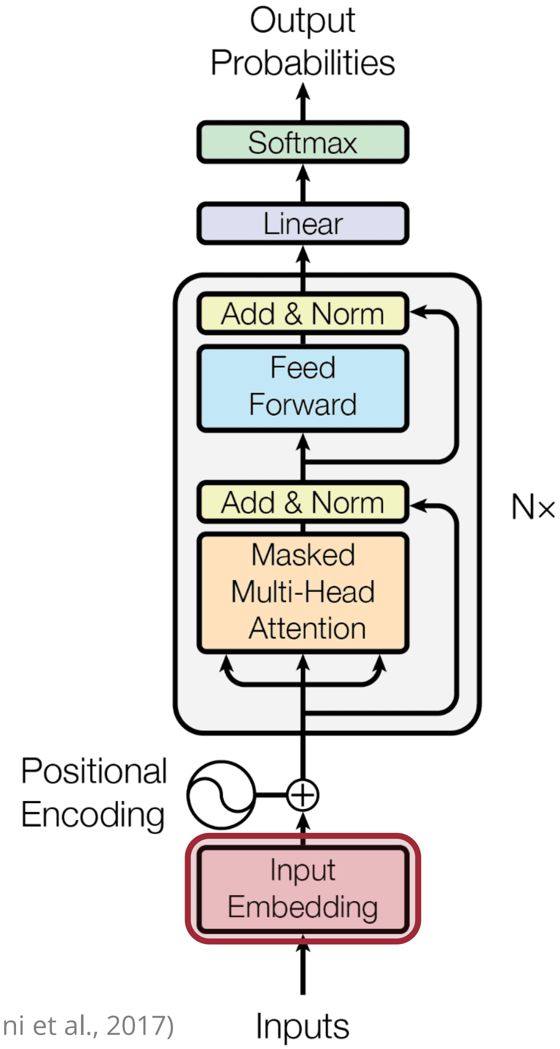
(Source: Tay et al., 2022)

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler, "Efficient Transformers: A Survey," *ACM Computing Surveys*, 55(6):109, 2022.

Word Embedding

- **Goal**: Learn to **represent words as vectors**
- **Intuition**: Synonyms should have close embeddings
- Should antonyms be far apart?
  - Not quite, antonyms usually fall in the same "topic"
  - For example, "happy" & "sad" are both emotions

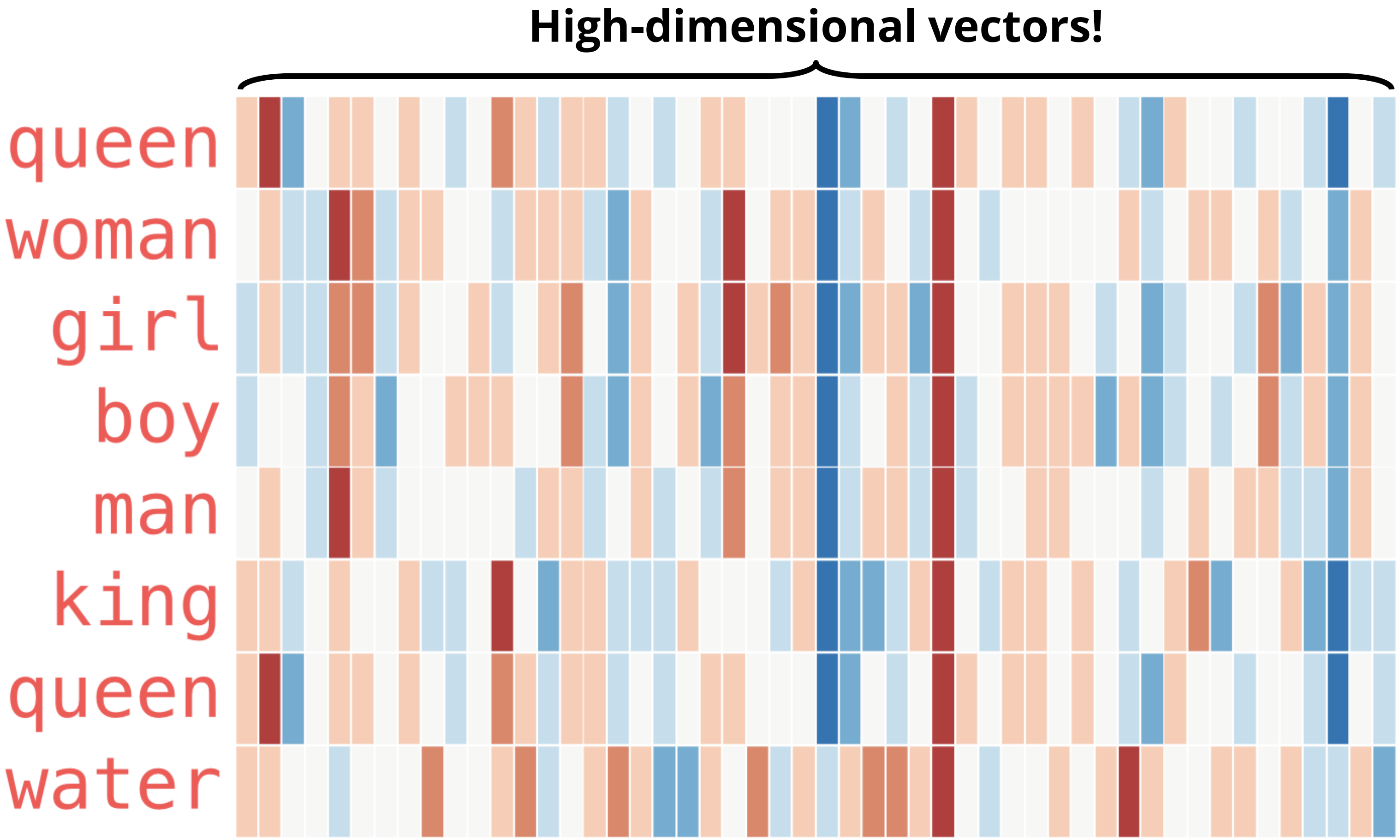




(Source: Vaswani et al., 2017)

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," *NeurIPS*, 2017.

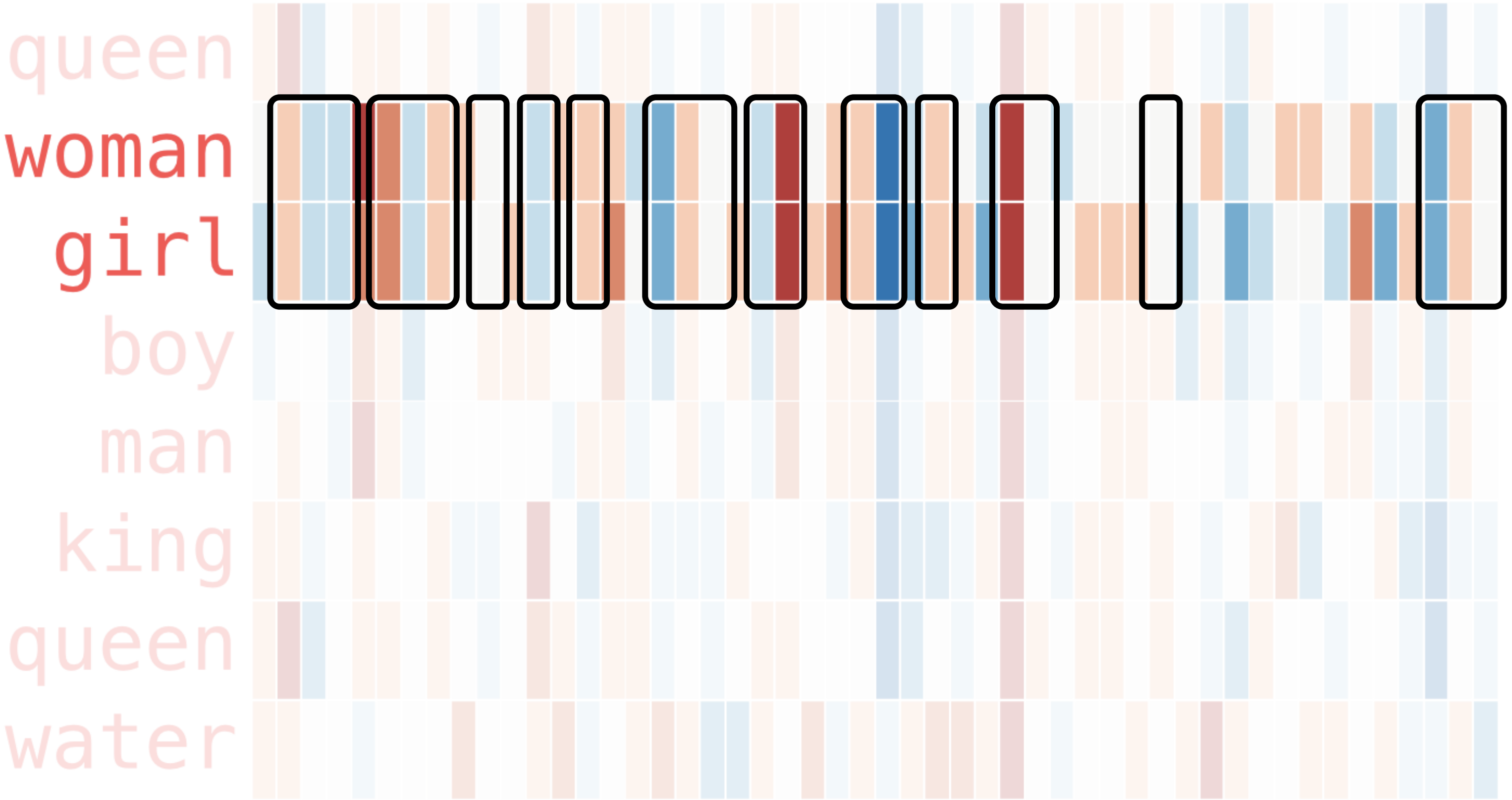
Word Embedding: A Toy Example



(Source: Alammar, 2019)

jalammar.github.io/illustrated-word2vec

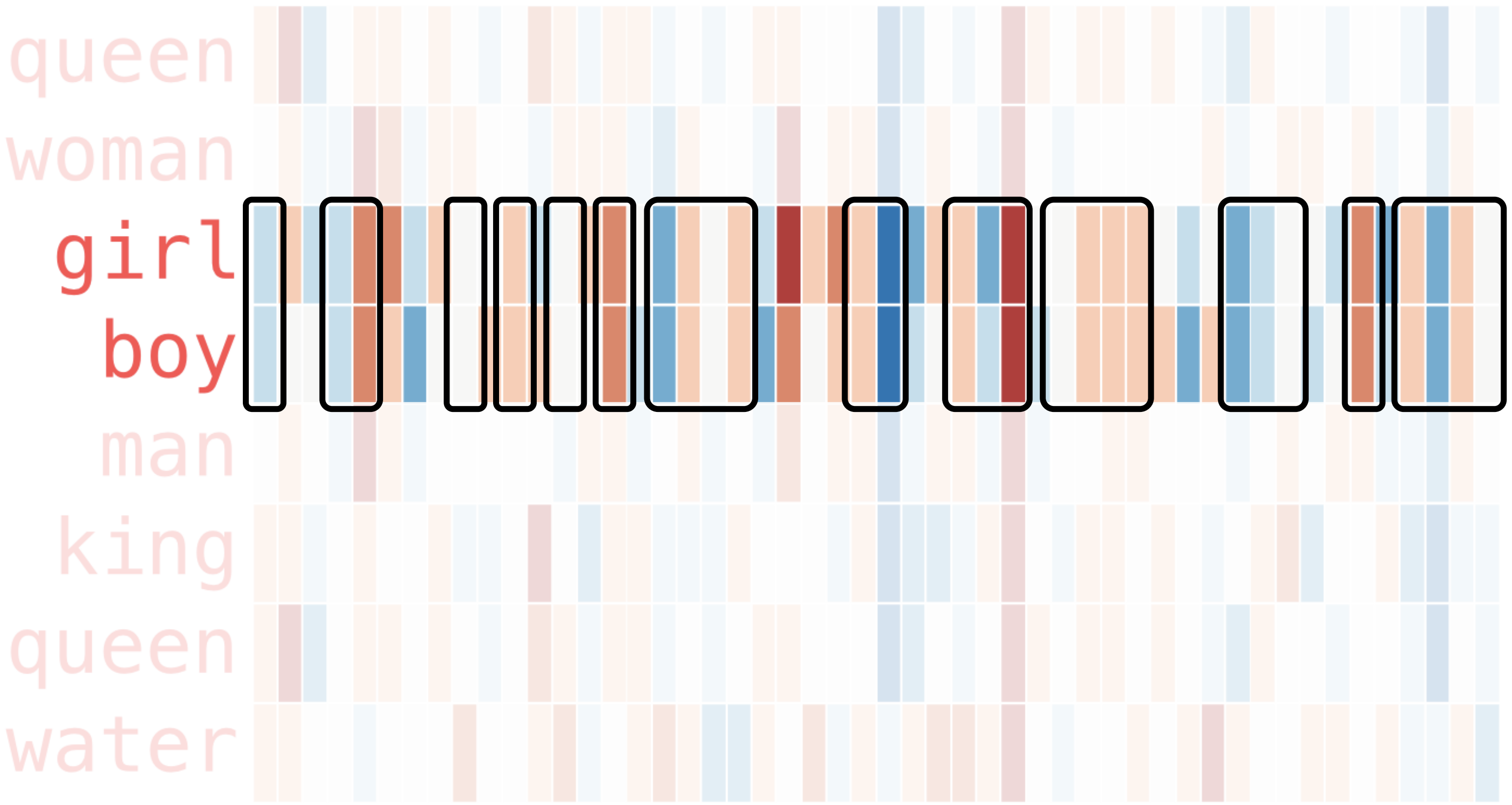
Word Embedding: A Toy Example



(Source: Alammar, 2019)

jalammar.github.io/illustrated-word2vec

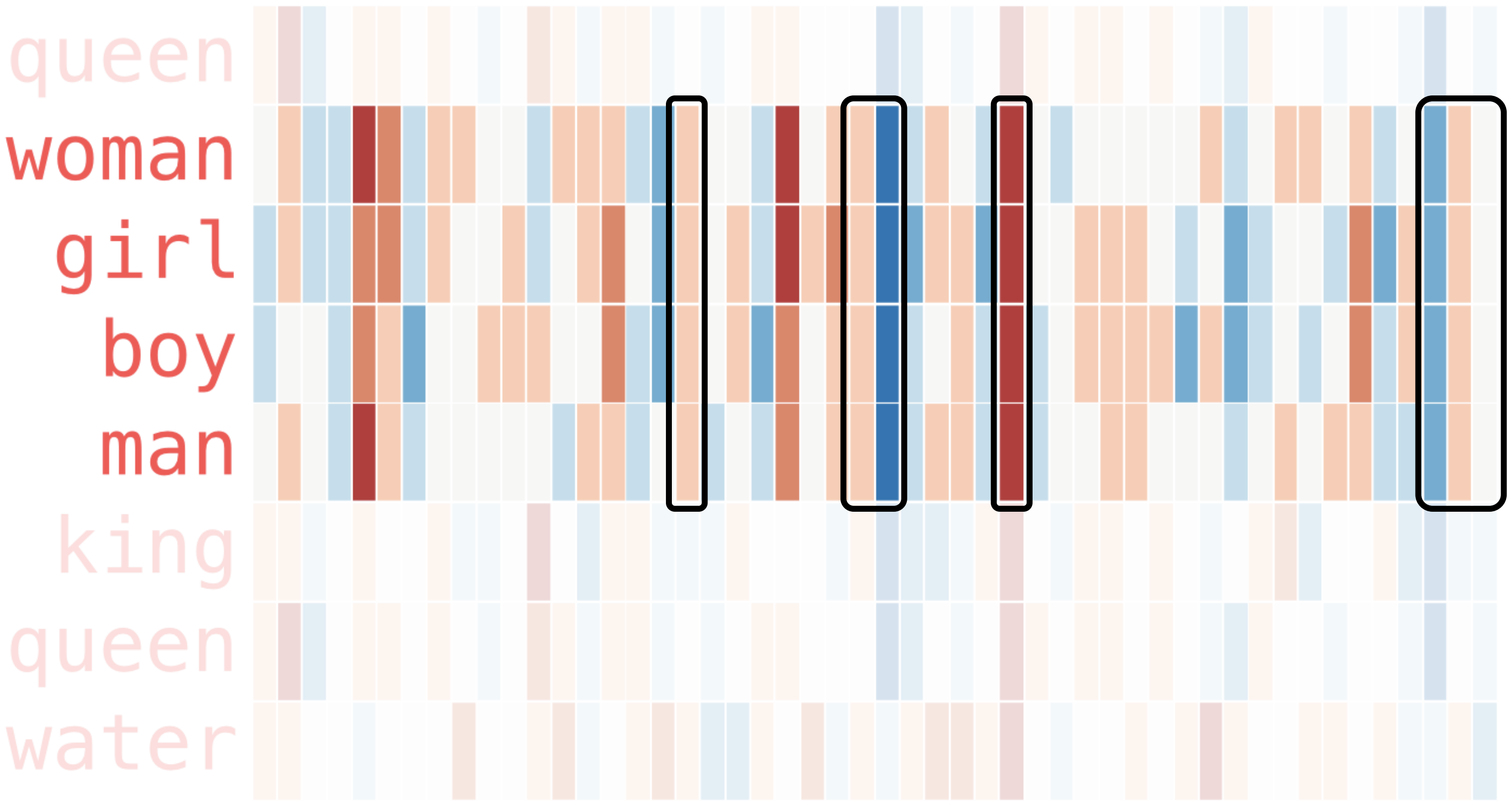
Word Embedding: A Toy Example



(Source: Alammar, 2019)

jalammar.github.io/illustrated-word2vec

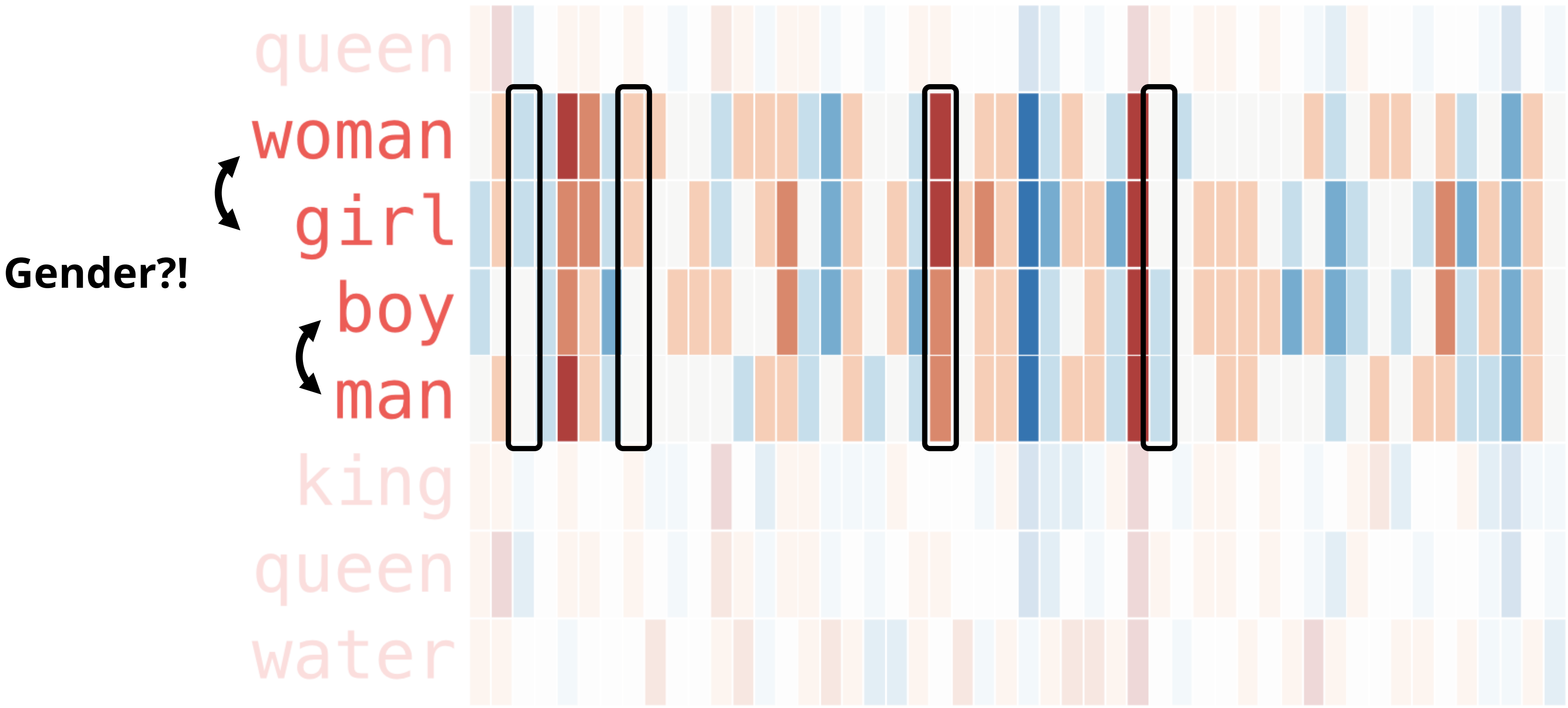
Word Embedding: A Toy Example



(Source: Alammar, 2019)

jalammar.github.io/illustrated-word2vec

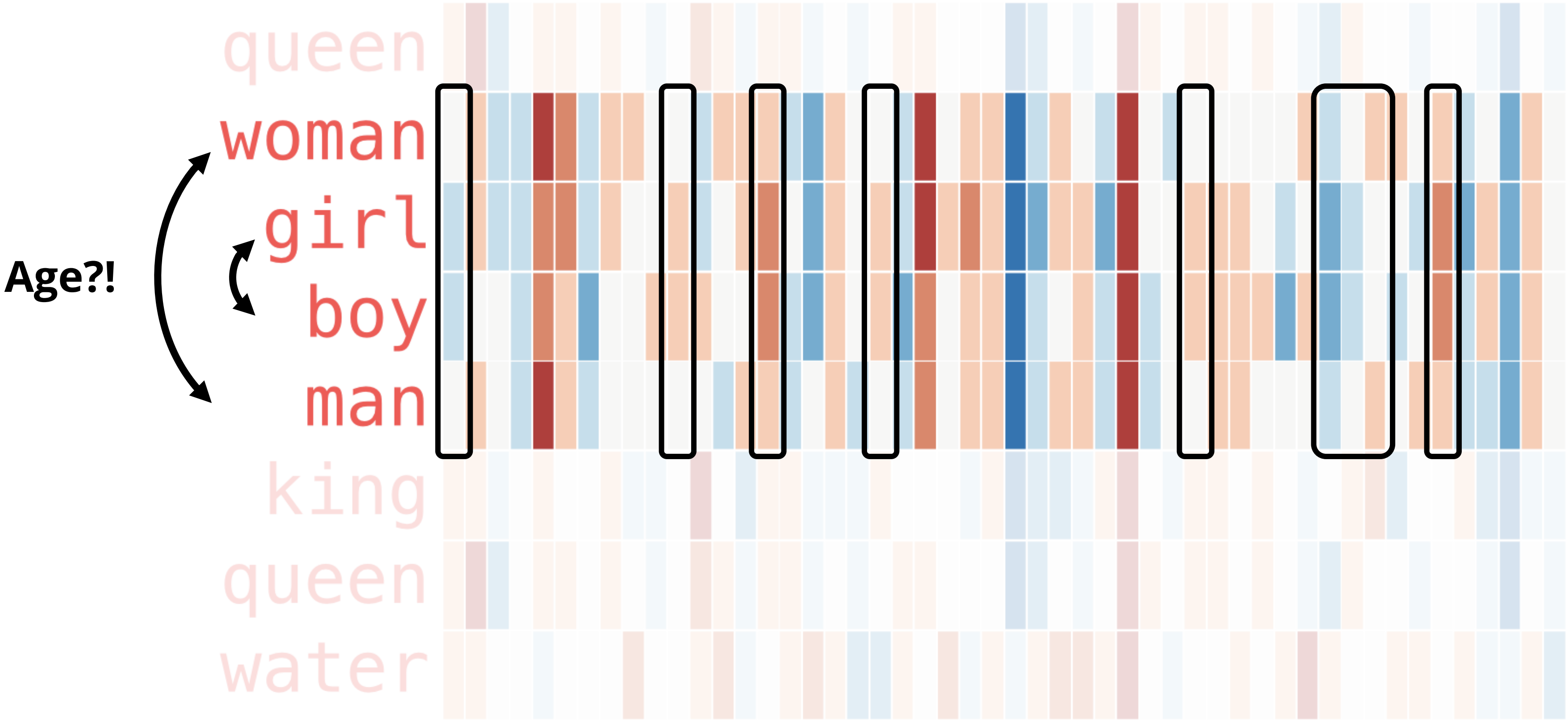
Word Embedding: A Toy Example



(Source: Alammar, 2019)

jalammar.github.io/illustrated-word2vec

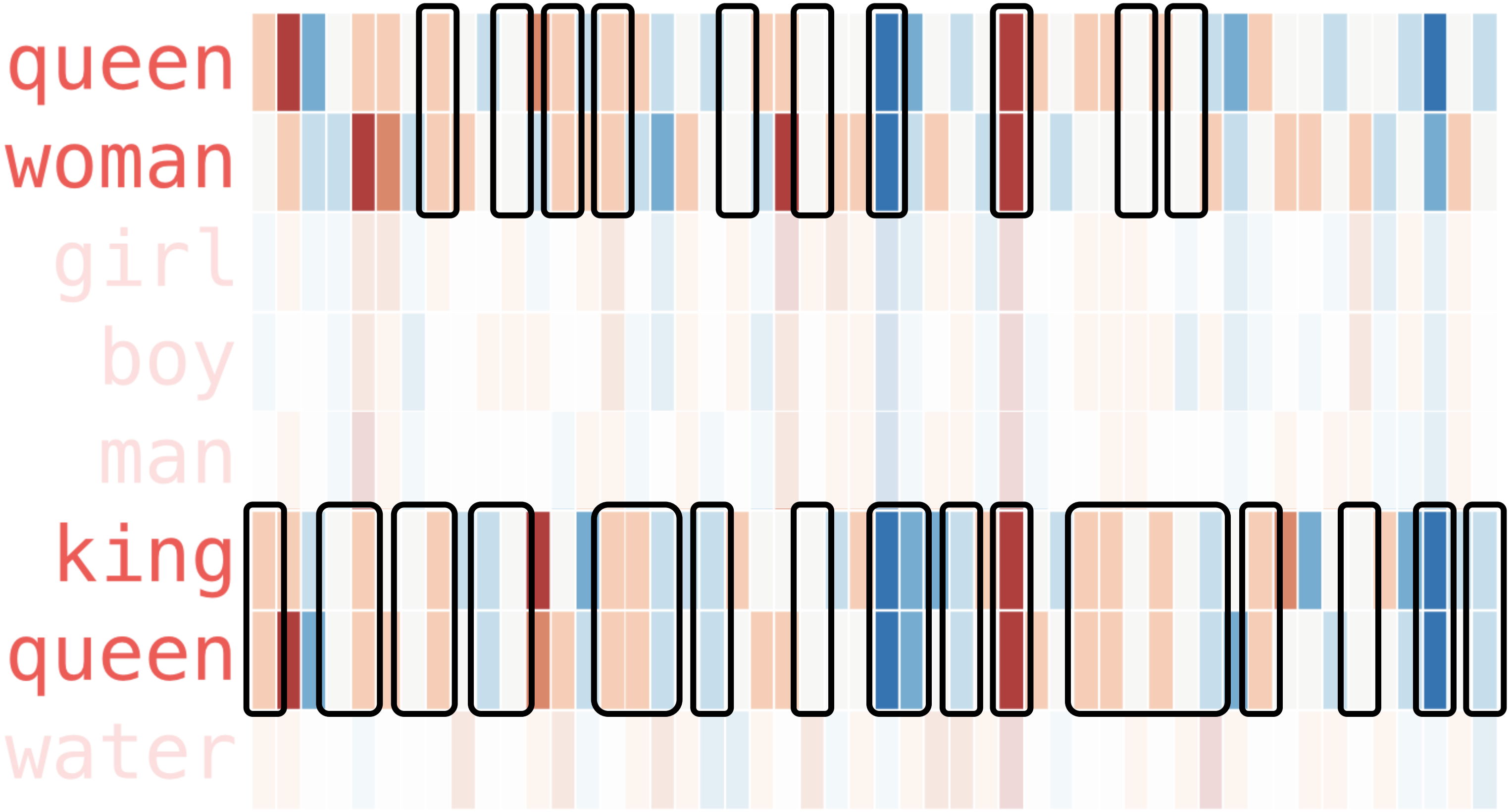
Word Embedding: A Toy Example



(Source: Alammar, 2019)

jalammar.github.io/illustrated-word2vec

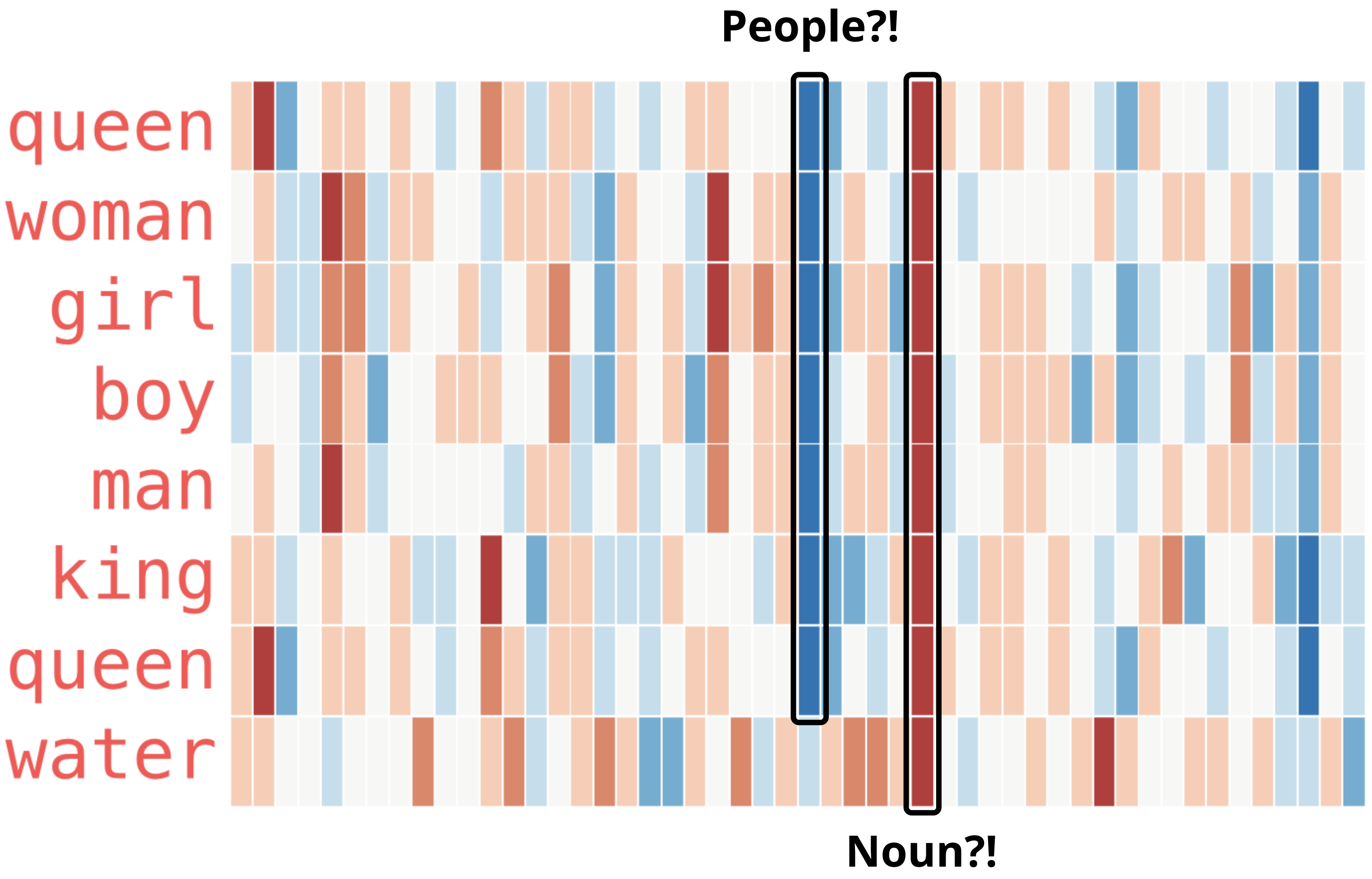
Word Embedding: A Toy Example



(Source: Alammar, 2019)

jalammar.github.io/illustrated-word2vec

Word Embedding: A Toy Example

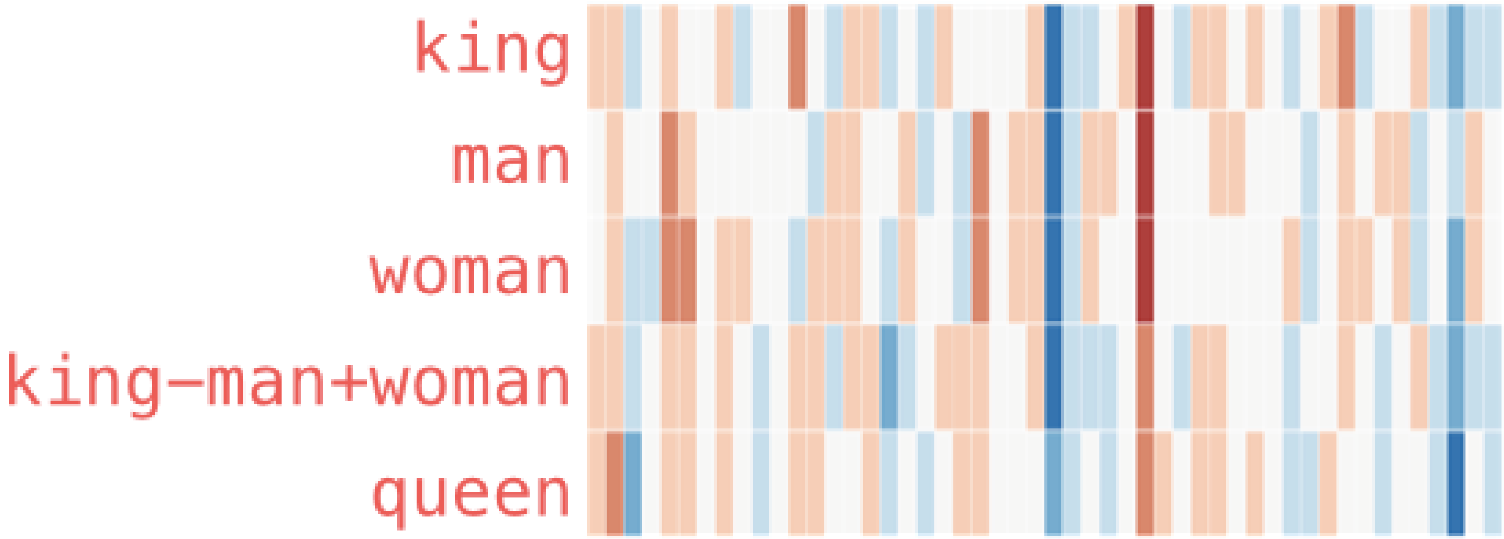


(Source: Alammar, 2019)

jalammar.github.io/illustrated-word2vec

Word Embedding: Arithmetic

king – man + woman ~= queen

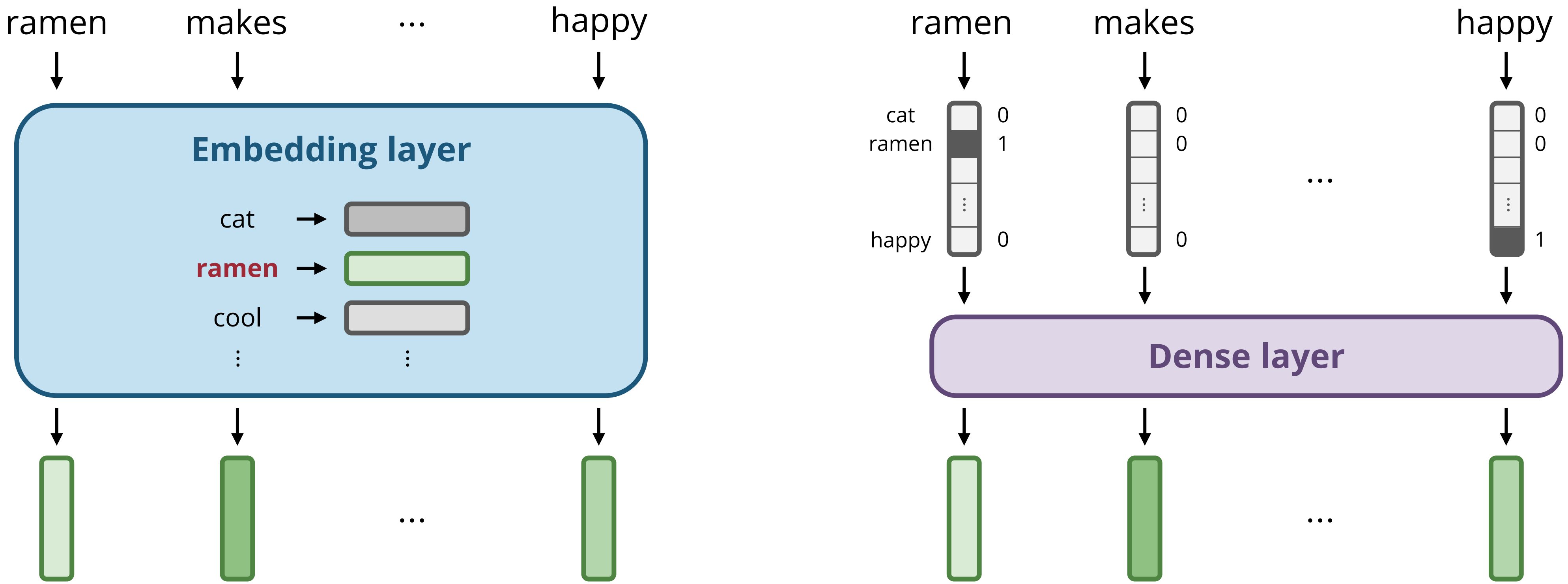


(Source: Alammar, 2019)

jalammar.github.io/illustrated-word2vec

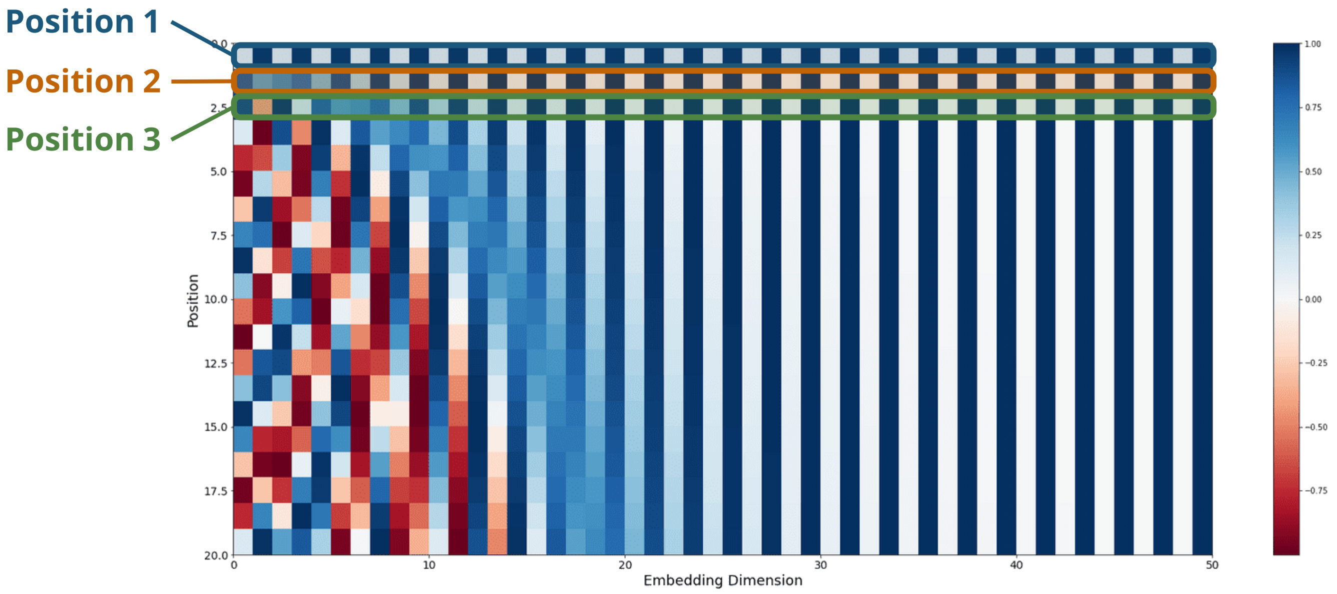
Word Embedding Layer

- A **word embedding layer** is functionally equivalent to **one-hot encoded words** followed by a **dense layer** → **But way faster thanks to hashing!**

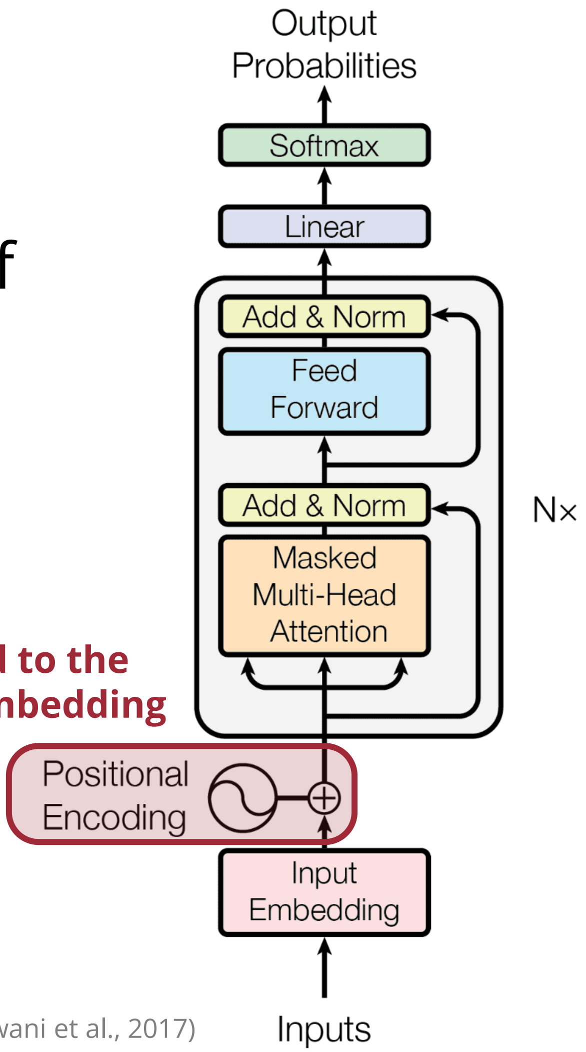


Positional Encoding

- **Intuition**: A word could have **different meanings at different positions**
- Positional encoding provides positional information of the words to the model



(Source: erdem.pl)



Added to the word embedding

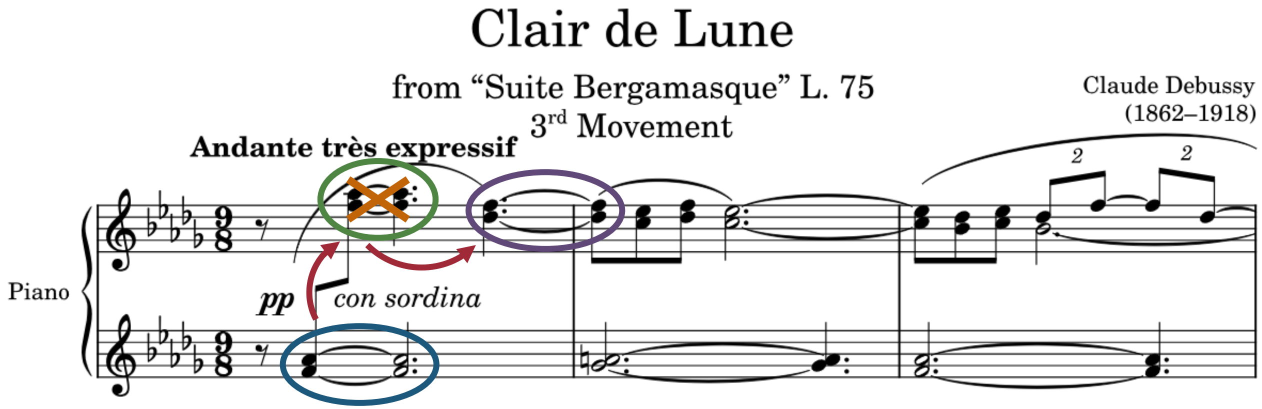
(Source: Vaswani et al., 2017)

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," *NeurIPS*, 2017.
erdem.pl/2021/05/understanding-positional-encoding-in-transformers

Transformers for Music Generation

Representing Polyphonic Music

- We can now handle music with multi-pitch at the same time
  - In the literature, “polyphonic” & “multi-pitch” are often used interchangeably



```
Note_on_65, Note_on_68, Time_shift_eighth_note, Note_on_77, Note_on_80,
Time_shift_half_note, Note_off_77, Note_off_80, Note_on_73, Note_on_77,
Time_shift_dotted_quarter_note, Note_off_65, Note_off_68, ...
```

Music Transformer (Huang et al., 2019)

- **Data**
  - Yamaha e-Piano Competition dataset (MAESTRO)
- **Representation**
  - 128 Note-On events
  - 128 Note-Off events
  - 100 Time-Shift events (10ms–1s)
  - 32 Set-Velocity events
- **Model**
  - Transformer

Almost the same representation as PerformanceRNN

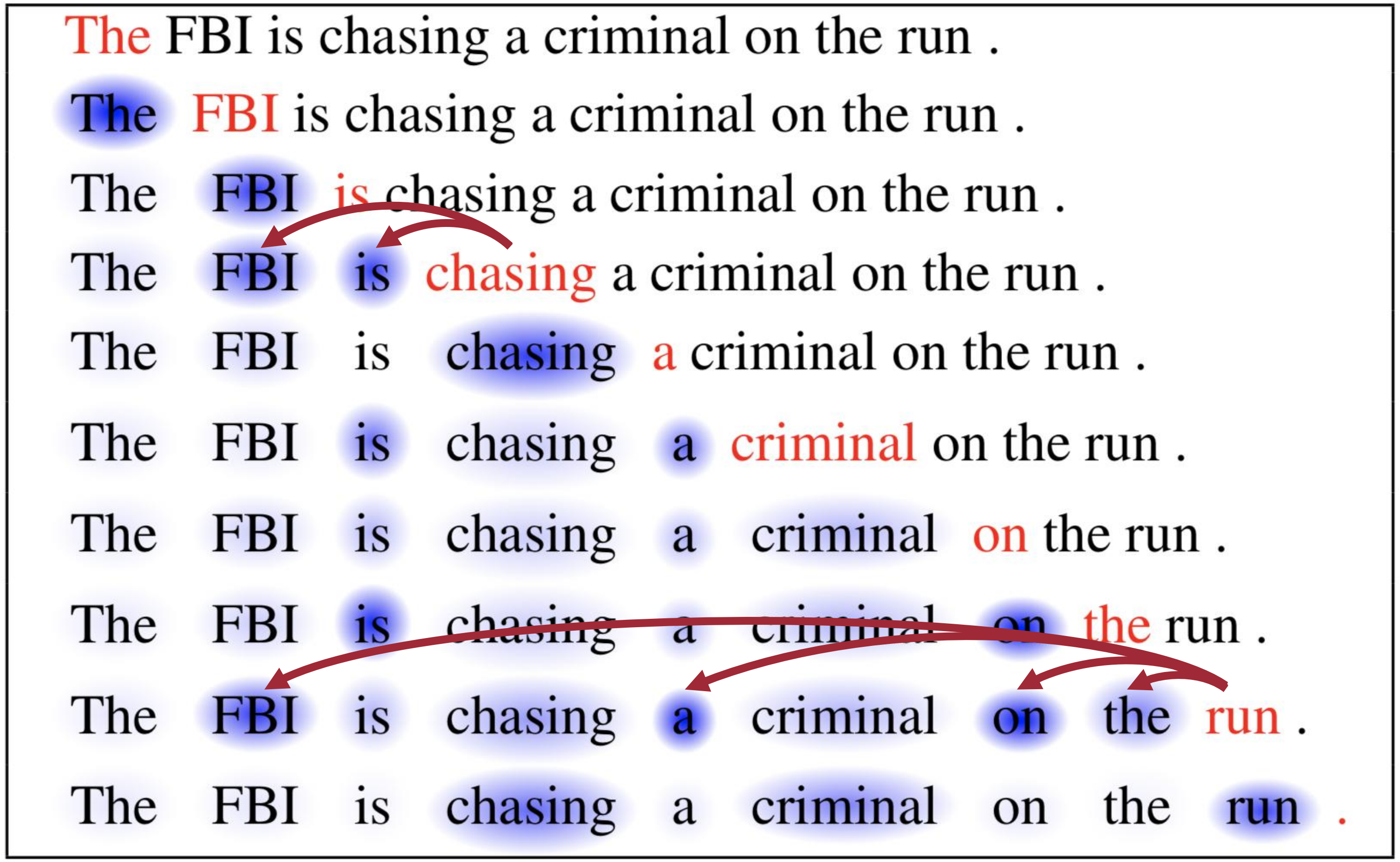
Expressive timing

Expressive dynamics

Examples of generated music

Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck, "Music Transformer: Generating Music with Long-Term Structure," *ICLR*, 2019.

Why Self-Attention Mechanism?



The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

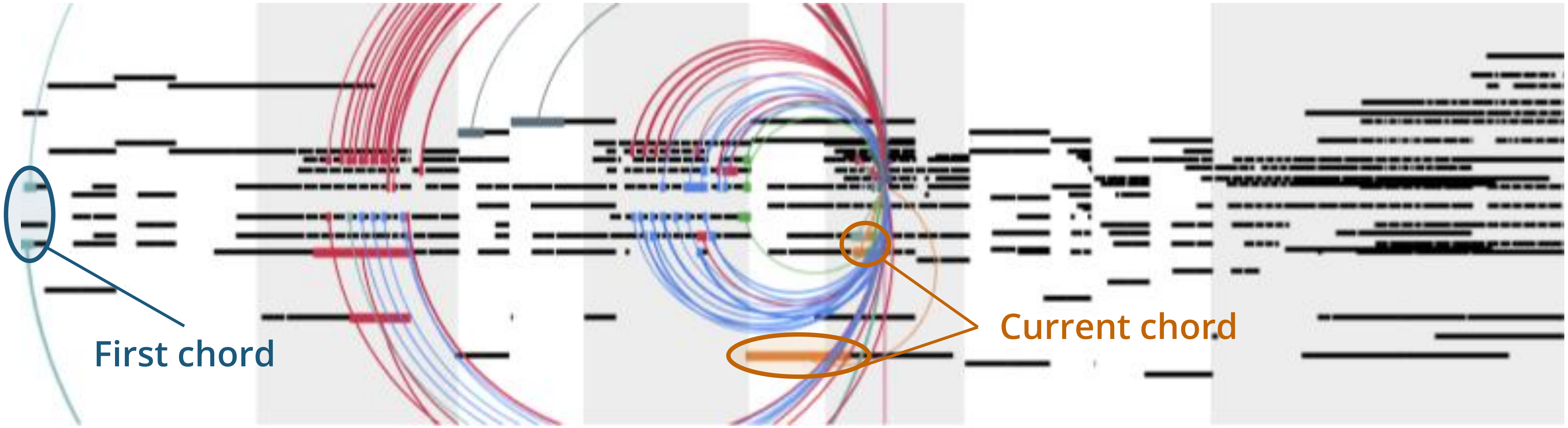
The FBI is chasing a criminal on the run .

(Source: Cheng et al., 2016)

Jianpeng Cheng, Li Dong, and Mirella Lapata, "Long Short-Term Memory-Networks for Machine Reading," *EMNLP*, 2016.

Visualizing Musical Self-attention (Huang et al., 2018)

(Each color represents an attention head)

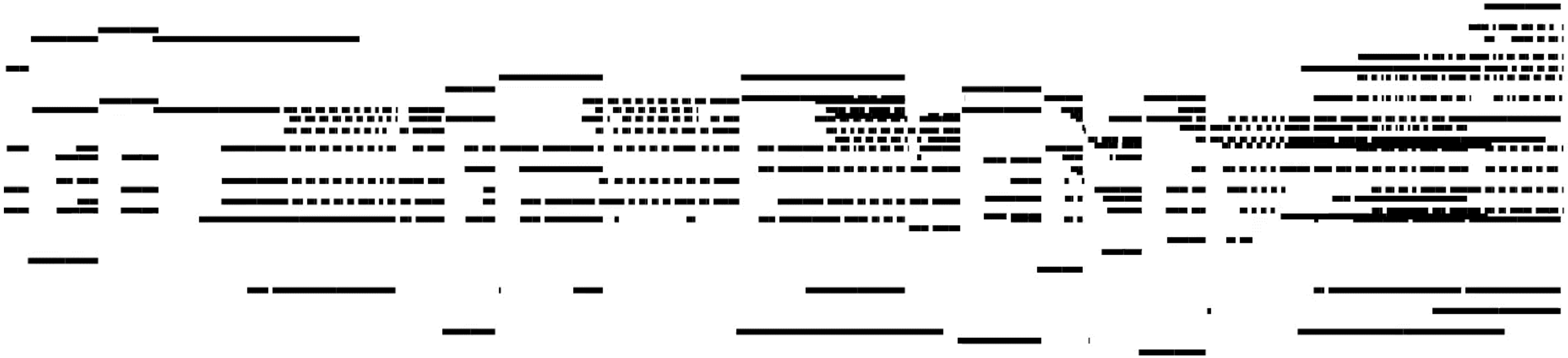


(Source: Huang et al., 2018)

Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck, "Music Transformer: Generating Music with Long-Term Structure," *ICLR*, 2019.
Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck, "Music Transformer: Generating Music with Long-Term Structure," *Magenta Blog*, December 13, 2018.

Visualizing Musical Self-attention (Huang et al., 2018)

(Each color represents an attention head)



(Source: Huang et al., 2018)

Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck, "Music Transformer: Generating Music with Long-Term Structure," *ICLR*, 2019.
Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck, "Music Transformer: Generating Music with Long-Term Structure," *Magenta Blog*, December 13, 2018.

Analyzing Musical Self-attention (Dong et al., 2023)

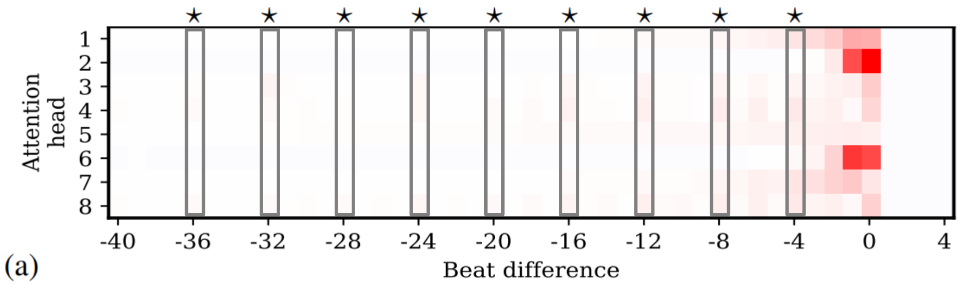
- Measuring **mean relative attention**

$$\gamma_k^{(d)} = \frac{\sum_{\mathbf{x}\in\mathcal{D}} \sum_{s>t} a_{s,t}(\mathbf{x})\ \mathbb{1}_{x_t^{(d)} - x_s^{(d)} = k}}{\sum_{\mathbf{x}\in\mathcal{D}} \sum_{s>t} a_{s,t}(\mathbf{x})} \qquad \tilde{\gamma}_k^{(d)} = \gamma_k^{(d)} - \frac{\sum_{\mathbf{x}\in\mathcal{D}} \sum_{s>t} \mathbb{1}_{x_t^{(d)} - x_s^{(d)} = k}}{\sum_{\mathbf{x}\in\mathcal{D}} \sum_{s>t} 1}$$

- The MMT model attends more to notes

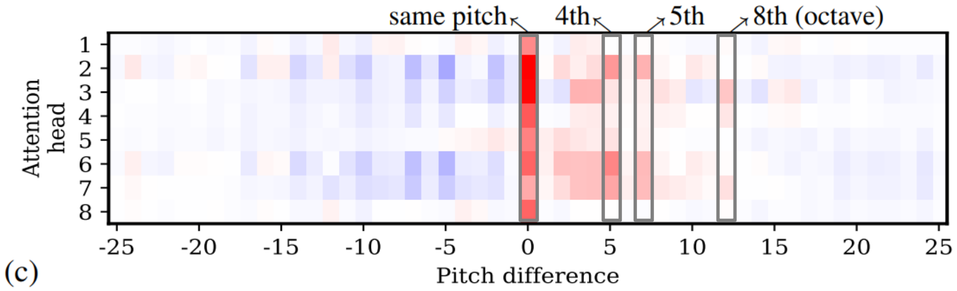
that are ***4N* beats away** in the past

that has a pitch in an octave above which **forms a consonant interval**



Positive/negative gain

(Source: Dong et al., 2023)

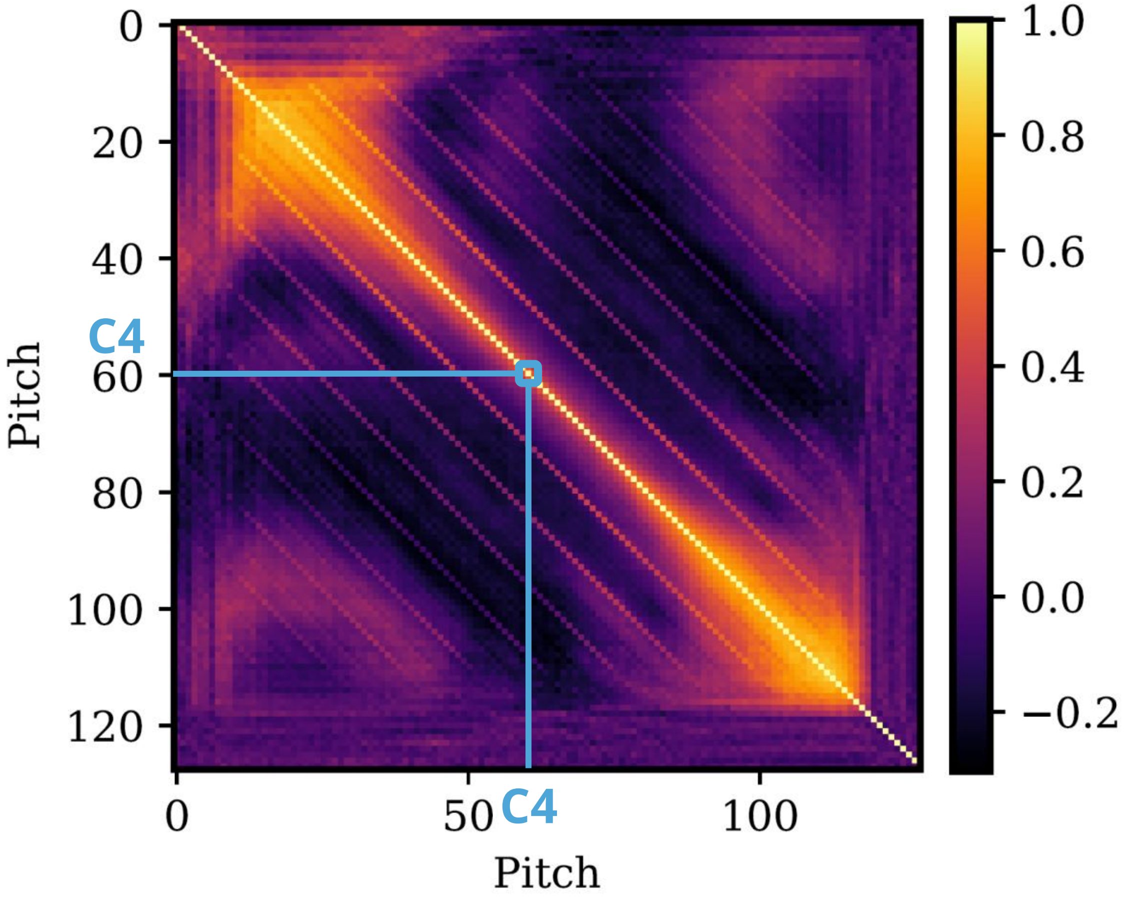


Positive/negative gain

(Source: Dong et al., 2023)

Hao-Wen Dong, Ke Chen, Shlomo Dubnov, Julian McAuley, and Taylor Berg-Kirkpatrick, "Multitrack Music Transformer," *ICASSP*, 2023.

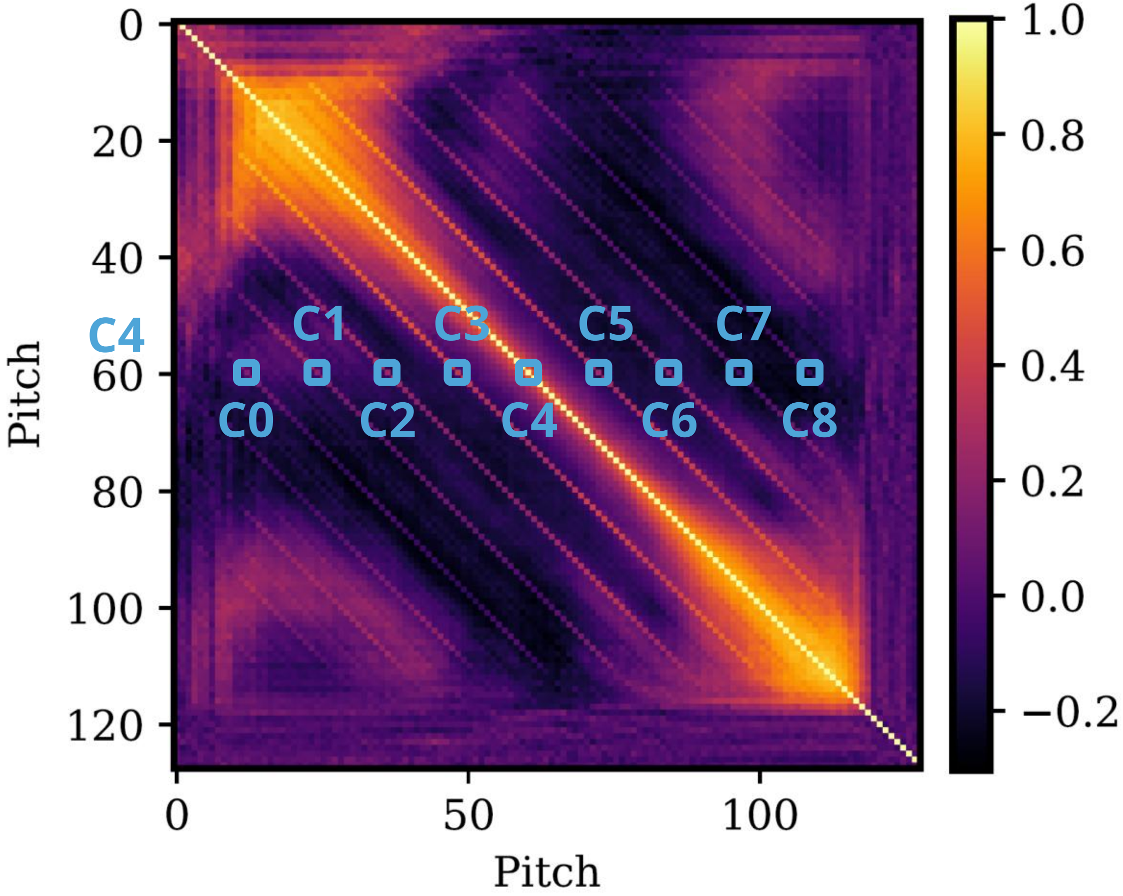
Learned Pitch Embedding (Dong et al., 2023)



(Source: Dong et al., 2023)

Hao-Wen Dong, Ke Chen, Shlomo Dubnov, Julian McAuley, and Taylor Berg-Kirkpatrick, "Multitrack Music Transformer," *ICASSP*, 2023.

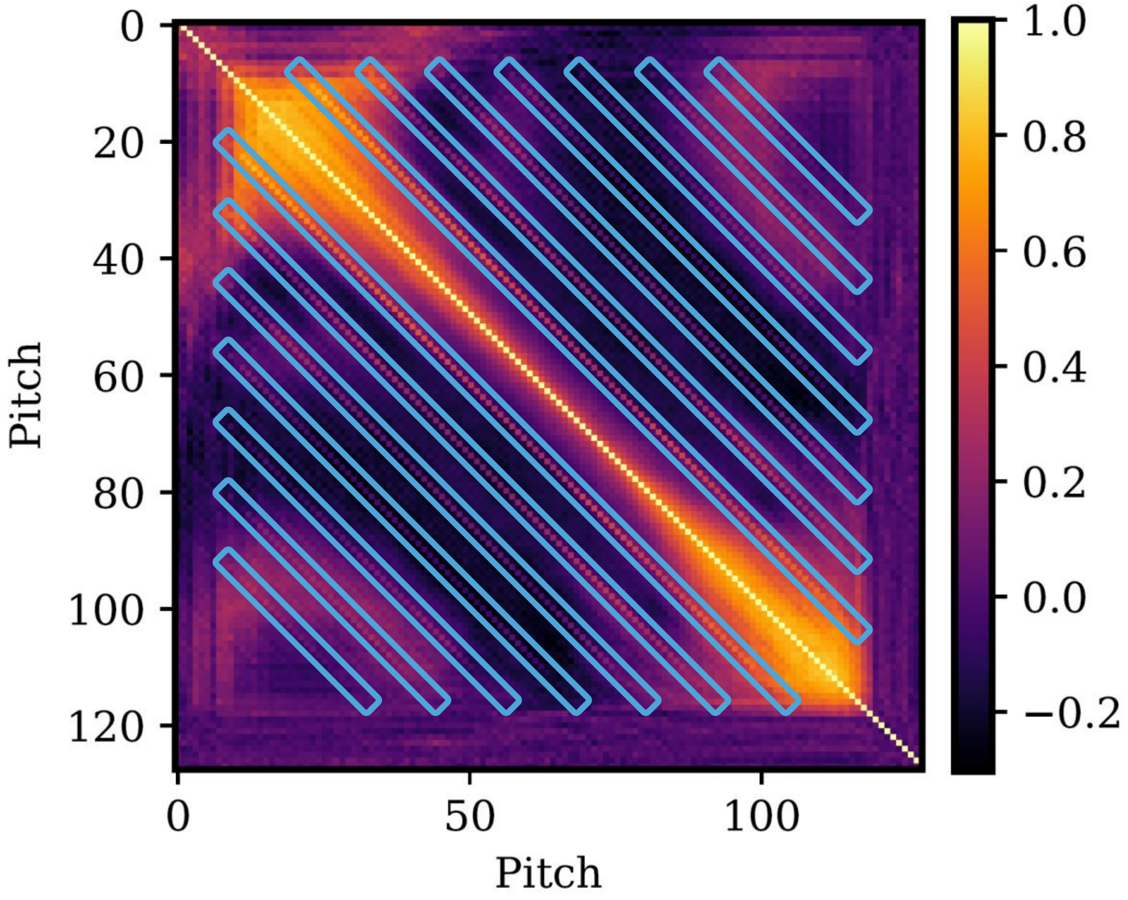
Learned Pitch Embedding (Dong et al., 2023)



(Source: Dong et al., 2023)

Hao-Wen Dong, Ke Chen, Shlomo Dubnov, Julian McAuley, and Taylor Berg-Kirkpatrick, "Multitrack Music Transformer," *ICASSP*, 2023.

Learned Pitch Embedding (Dong et al., 2023)



Transformer models can learn the concept of octaves!

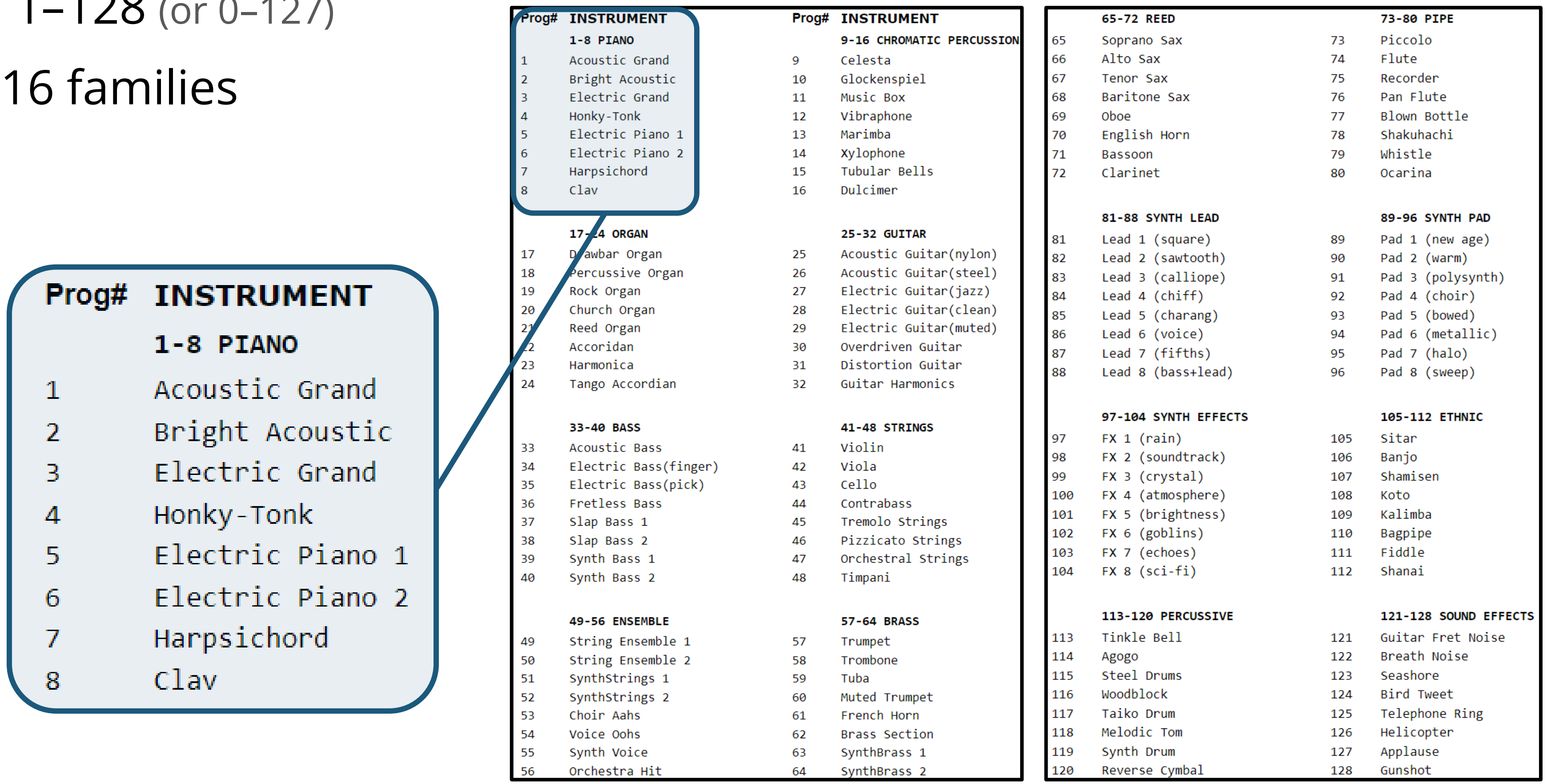
(Source: Dong et al., 2023)

Hao-Wen Dong, Ke Chen, Shlomo Dubnov, Julian McAuley, and Taylor Berg-Kirkpatrick, "Multitrack Music Transformer," *ICASSP*, 2023.

Transformers for Multitrack Music Generation

Representing Multiple Instruments

- Using **MIDI program change** messages
  - Program numbers: 1–128 (or 0–127)
  - 128 instruments in 16 families



Prog#	INSTRUMENT
	1-8 PIANO
1	Acoustic Grand
2	Bright Acoustic
3	Electric Grand
4	Honky-Tonk
5	Electric Piano 1
6	Electric Piano 2
7	Harpsichord
8	Clav

Prog#	INSTRUMENT	Prog#	INSTRUMENT
	1-8 PIANO		9-16 CHROMATIC PERCUSSION
1	Acoustic Grand	9	Celesta
2	Bright Acoustic	10	Glockenspiel
3	Electric Grand	11	Music Box
4	Honky-Tonk	12	Vibraphone
5	Electric Piano 1	13	Marimba
6	Electric Piano 2	14	Xylophone
7	Harpsichord	15	Tubular Bells
8	Clav	16	Dulcimer
	17-24 ORGAN		25-32 GUITAR
17	Drawbar Organ	25	Acoustic Guitar(nylon)
18	Percussive Organ	26	Acoustic Guitar(steel)
19	Rock Organ	27	Electric Guitar(jazz)
20	Church Organ	28	Electric Guitar(clean)
21	Reed Organ	29	Electric Guitar(muted)
22	Accoridan	30	Overdriven Guitar
23	Harmonica	31	Distortion Guitar
24	Tango Accordian	32	Guitar Harmonics
	33-40 BASS		41-48 STRINGS
33	Acoustic Bass	41	Violin
34	Electric Bass(finger)	42	Viola
35	Electric Bass(pick)	43	Cello
36	Fretless Bass	44	Contrabass
37	Slap Bass 1	45	Tremolo Strings
38	Slap Bass 2	46	Pizzicato Strings
39	Synth Bass 1	47	Orchestral Strings
40	Synth Bass 2	48	Timpani
	49-56 ENSEMBLE		57-64 BRASS
49	String Ensemble 1	57	Trumpet
50	String Ensemble 2	58	Trombone
51	SynthStrings 1	59	Tuba
52	SynthStrings 2	60	Muted Trumpet
53	Choir Aahs	61	French Horn
54	Voice Oohs	62	Brass Section
55	Synth Voice	63	SynthBrass 1
56	Orchestra Hit	64	SynthBrass 2

	65-72 REED		73-80 PIPE
65	Soprano Sax	73	Piccolo
66	Alto Sax	74	Flute
67	Tenor Sax	75	Recorder
68	Baritone Sax	76	Pan Flute
69	Oboe	77	Blown Bottle
70	English Horn	78	Shakuhachi
71	Bassoon	79	Whistle
72	Clarinet	80	Ocarina
	81-88 SYNTH LEAD		89-96 SYNTH PAD
81	Lead 1 (square)	89	Pad 1 (new age)
82	Lead 2 (sawtooth)	90	Pad 2 (warm)
83	Lead 3 (calliope)	91	Pad 3 (polysynth)
84	Lead 4 (chiff)	92	Pad 4 (choir)
85	Lead 5 (charang)	93	Pad 5 (bowed)
86	Lead 6 (voice)	94	Pad 6 (metallic)
87	Lead 7 (fifths)	95	Pad 7 (halo)
88	Lead 8 (bass+lead)	96	Pad 8 (sweep)
	97-104 SYNTH EFFECTS		105-112 ETHNIC
97	FX 1 (rain)	105	Sitar
98	FX 2 (soundtrack)	106	Banjo
99	FX 3 (crystal)	107	Shamisen
100	FX 4 (atmosphere)	108	Koto
101	FX 5 (brightness)	109	Kalimba
102	FX 6 (goblins)	110	Bagpipe
103	FX 7 (echoes)	111	Fiddle
104	FX 8 (sci-fi)	112	Shanai
	113-120 PERCUSSIVE		121-128 SOUND EFFECTS
113	Tinkle Bell	121	Guitar Fret Noise
114	Agogo	122	Breath Noise
115	Steel Drums	123	Seashore
116	Woodblock	124	Bird Tweet
117	Taiko Drum	125	Telephone Ring
118	Melodic Tom	126	Helicopter
119	Synth Drum	127	Applause
120	Reverse Cymbal	128	Gunshot

(Source: Roger Dannenberg)

www.cs.cmu.edu/~music/cmsip/readings/GMSpecs_Patches.htm

MuseNet (Payne et al., 2019)

- **Data**: ClassicalArchives + BitMidi + MAESTRO
- **Representation**: "**instrument:velocity:pitch**"
  - Time shifts in real time (sec)
- **Model**: Transformer

```
bach piano_strings start tempo90
piano:v72:G1 piano:v72:G2 piano:v72:B4
piano:v72:D4 violin:v80:G4 piano:v72:G4
piano:v72:B5 piano:v72:D5 wait:12
piano:v0:B5 wait:5 piano:v72:D5 wait:12
...
```

Example of generated music

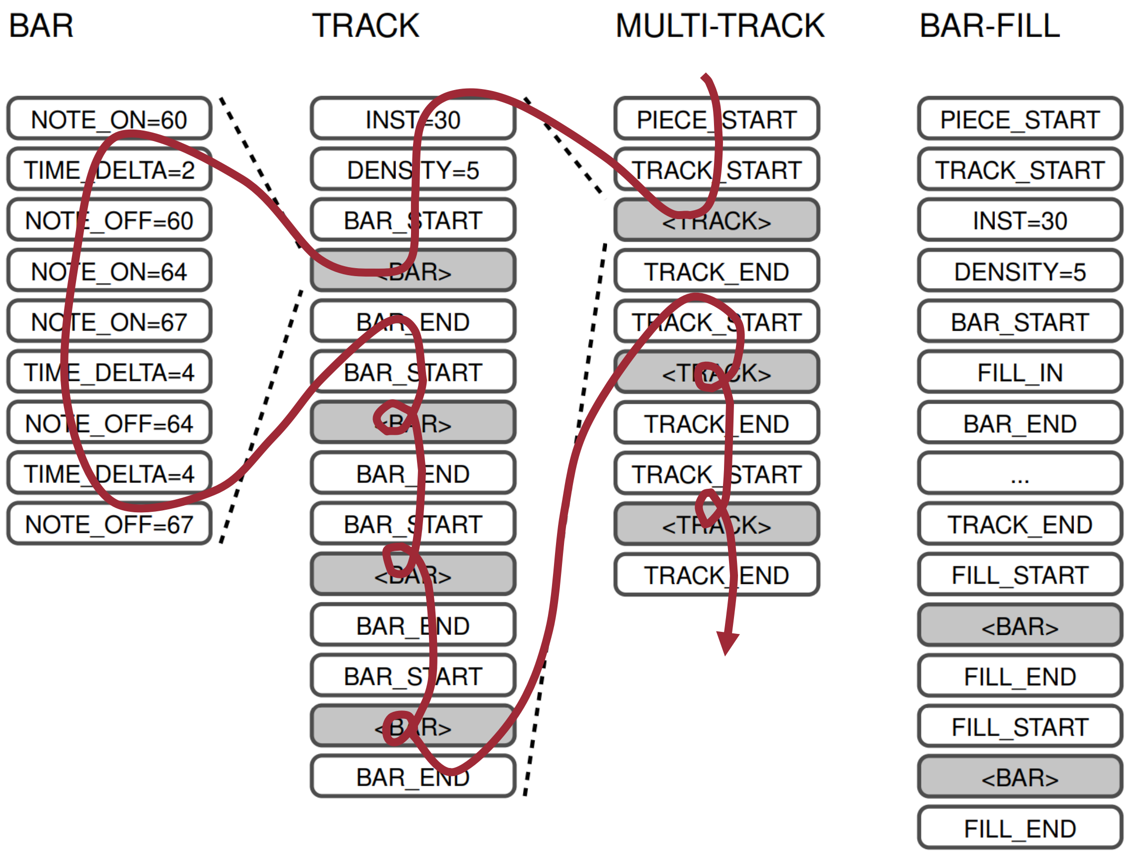
Christine Payne, "MuseNet," *OpenAI*, 2019.

Multitrack Music Machine (Ens & Pasquier, 2020)

- **Data**: Lakh MIDI Dataset (LMD)
- **Representation**: as shown →
- **Model**: Transformer

LETS START WITH SOME U2

youtu.be/NdeMZ3y-84Q



BAR	TRACK	MULTI-TRACK	BAR-FILL
NOTE_ON=60	INST=30	PIECE_START	PIECE_START
TIME_DELTA=2	DENSITY=5	TRACK_START	TRACK_START
NOTE_OFF=60	BAR_START	<TRACK>	INST=30
NOTE_ON=64	<BAR>	TRACK_END	DENSITY=5
NOTE_ON=67	BAR_END	TRACK_START	BAR_START
TIME_DELTA=4	BAR_START	<TRACK>	FILL_IN
NOTE_OFF=64	<BAR>	TRACK_END	BAR_END
TIME_DELTA=4	BAR_END	TRACK_START	...
NOTE_OFF=67	BAR_START	<TRACK>	TRACK_END
	<BAR>	TRACK_END	FILL_START
	BAR_END		<BAR>
	BAR_START		FILL_END
	<BAR>		FILL_START
	BAR_END		<BAR>
			FILL_END

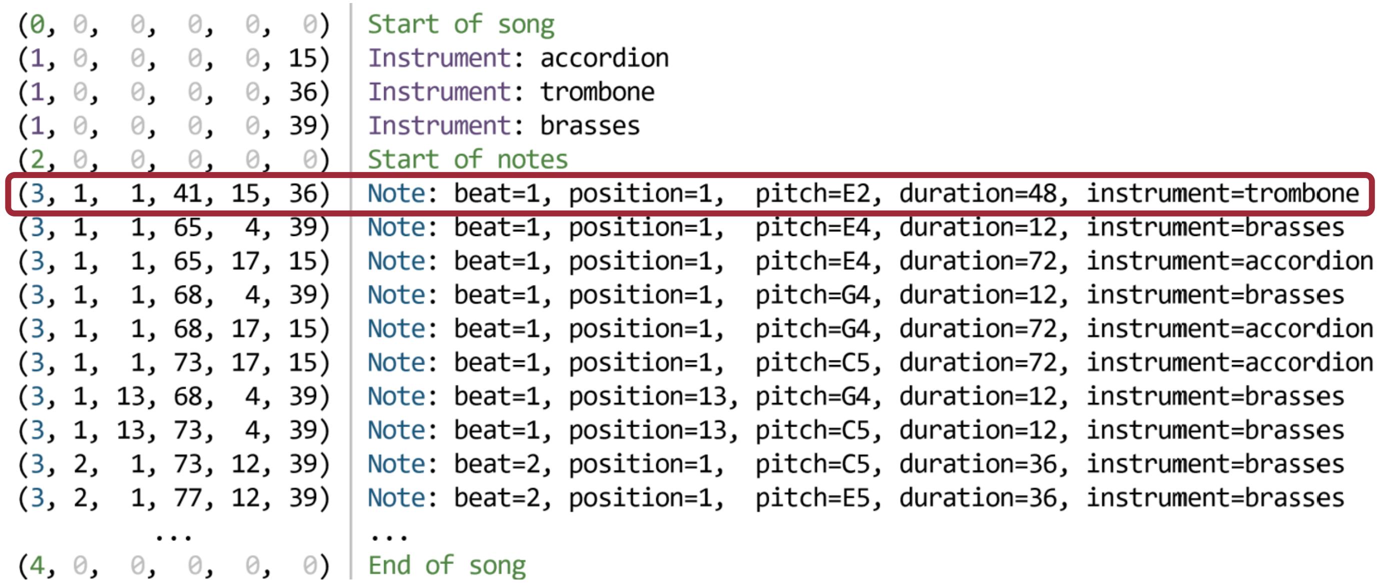
Fig. 1. The MultiTrack and BarFill representations are shown. The <bar> tokens correspond to complete bars, and the <track> tokens correspond to complete tracks.

(Ens & Pasquier, 2020)

Jeff Ens and Philippe Pasquier, "MMM : Exploring Conditional Multi-Track Music Generation with the Transformer," *arXiv preprint arXiv:2008:06048*, 2020.

Multitrack Music Transformer (Dong et al., 2023)

- **Data**: Symbolic Orchestral Database (SOD)
- **Representation**: “(**beat**, **position**, **pitch**, **duration**, **instrument**)”
- **Model**: Multi-dimensional Transformer



```
(0, 0,  0,  0,  0,  0) | Start of song
(1, 0,  0,  0,  0, 15) | Instrument: accordion
(1, 0,  0,  0,  0, 36) | Instrument: trombone
(1, 0,  0,  0,  0, 39) | Instrument: brasses
(2, 0,  0,  0,  0,  0) | Start of notes
(3, 1,  1, 41, 15, 36) | Note: beat=1, position=1,  pitch=E2, duration=48, instrument=trombone
(3, 1,  1, 65,  4, 39) | Note: beat=1, position=1,  pitch=E4, duration=12, instrument=brasses
(3, 1,  1, 65, 17, 15) | Note: beat=1, position=1,  pitch=E4, duration=72, instrument=accordion
(3, 1,  1, 68,  4, 39) | Note: beat=1, position=1,  pitch=G4, duration=12, instrument=brasses
(3, 1,  1, 68, 17, 15) | Note: beat=1, position=1,  pitch=G4, duration=72, instrument=accordion
(3, 1,  1, 73, 17, 15) | Note: beat=1, position=1,  pitch=C5, duration=72, instrument=accordion
(3, 1, 13, 68,  4, 39) | Note: beat=1, position=13, pitch=G4, duration=12, instrument=brasses
(3, 1, 13, 73,  4, 39) | Note: beat=1, position=13, pitch=C5, duration=12, instrument=brasses
(3, 2,  1, 73, 12, 39) | Note: beat=2, position=1,  pitch=C5, duration=36, instrument=brasses
(3, 2,  1, 77, 12, 39) | Note: beat=2, position=1,  pitch=E5, duration=36, instrument=brasses
         ...           | ...
(4, 0,  0,  0,  0,  0) | End of song
```

(Source: Dong et al., 2023)

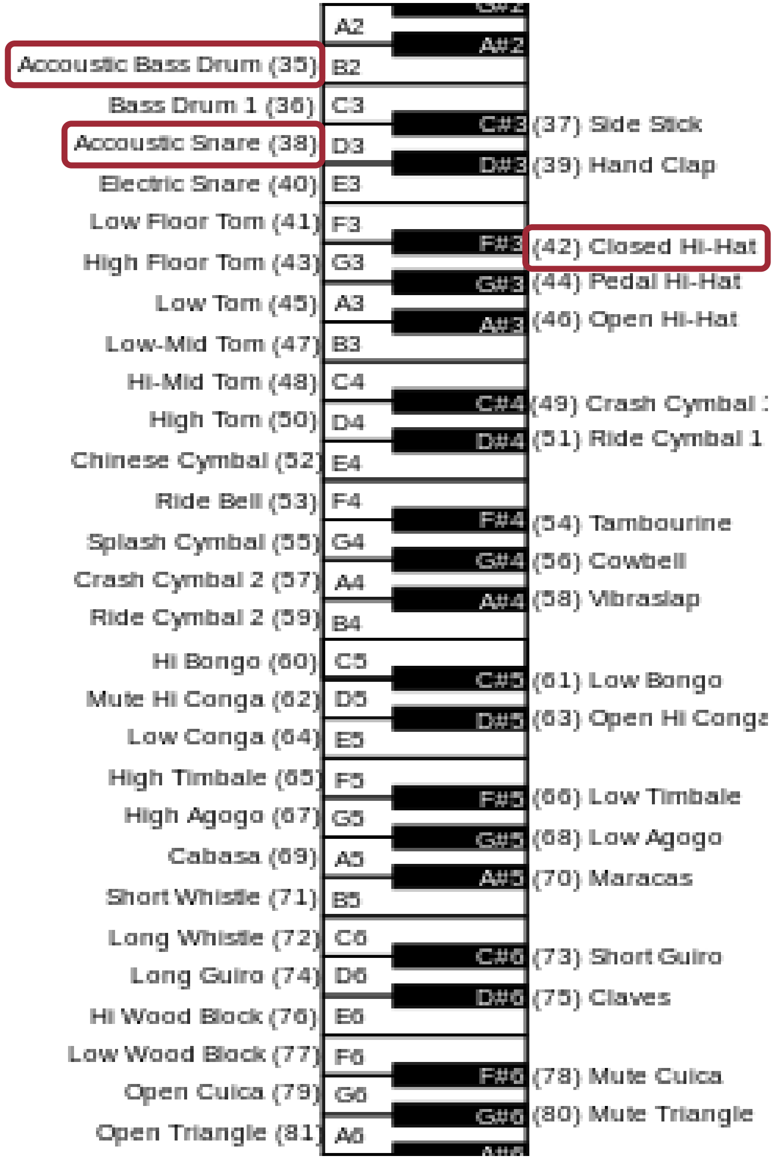
Example of generated music

🔈

Hao-Wen Dong, Ke Chen, Shlomo Dubnov, Julian McAuley, and Taylor Berg-Kirkpatrick, “Multitrack Music Transformer,” *ICASSP*, 2023.

Drums in MIDI

- **Channel 10** is reserved for drums
- Encoded by MIDI pitches 35–81
- Models that support drums
  - **MuseNet** (Payne et al., 2019)
  - **Song from PI** (Chu et al., 2017)
  - **MMM** (Ens and Pasquier, 2019)
  - *and many more…*



(Source: Wikipedia)

en.wikipedia.org/wiki/General_MIDI
Christine Payne, "MuseNet," *OpenAI*, 2019.
Hang Chu, Raquel Urtasun, and Sanja Fidler, "Song From PI: A Musically Plausible Network for Pop Music Generation," *ICLR Workshop*, 2017.
Jeff Ens and Philippe Pasquier, "MMM : Exploring Conditional Multi-Track Music Generation with the Transformer," *arXiv preprint arXiv:2008.06048*, 2020.

The Many Representations for Music Generation

- **PerformanceRNN** (Oore et al., 2020)
- **REMI** (Huang et al., 2020)
- **MuMIDI** (Ren et al., 2020)
- **Compound Word** (Hsiao et al., 2021)
- **REMI+** (von Rütte et al., 2023)
- **TSD** (Fradet et al., 2023)
- *and so on...*

MIDITok

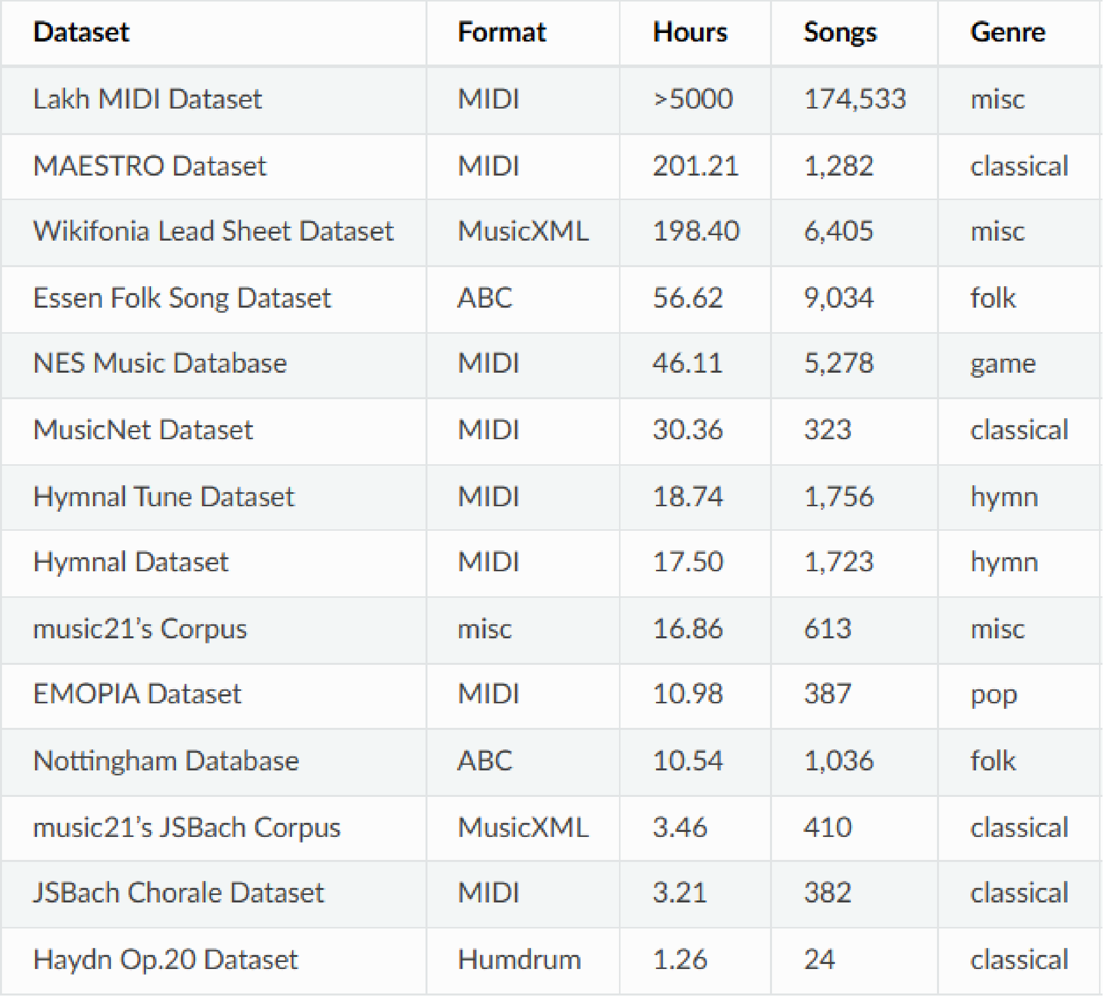
github.com/Natooz/MidiTok

Sageev Oore, Ian Simon, Sander Dieleman, Douglas Eck, and Karen Simonyan, “This Time with Feeling: Learning Expressive Musical Performance”, *Neural Computing and Applications*, 32, 2020.
Yu-Siang Huang and Yi-Hsuan Yang, “Pop Music Transformer: Beat-based Modeling and Generation of Expressive Pop Piano Compositions,” *MM*, 2020.
Yi Ren, Jinzheng He, Xu Tan, Tao Qin, Zhou Zhao, and Tie-Yan Liu, “PopMAG: Pop Music Accompaniment Generation,” *MM*, 2020.
Wen-Yi Hsiao, Jen-Yu Liu, Yin-Cheng Yeh, and Yi-Hsuan Yang, “Compound Word Transformer: Learning to Compose Full-Song Music over Dynamic Directed Hypergraphs,” *AAAI*, 2021.
Dimitri von Rütte, Luca Biggio, Yannic Kilcher, and Thomas Hofmann, “FIGARO: Generating Symbolic Music with Fine-Grained Artistic Control,” *ICLR*, 2023.
Nathan Fradet, Nicolas Gutowski, Fabien Chhel, and Jean-Pierre Briot, “Byte Pair Encoding for Symbolic Music,” *EMNLP*, 2023.

Symbolic Music Datasets

- JSBach Chorale
- MusicNet
- Essen Folk Song Dataset
- Wikifonia
- Lakh MIDI Dataset
- MetaMIDI
- MAESTRO

Symbolic Music Datasets



Dataset	Format	Hours	Songs	Genre
Lakh MIDI Dataset	MIDI	>5000	174,533	misc
MAESTRO Dataset	MIDI	201.21	1,282	classical
Wikifonia Lead Sheet Dataset	MusicXML	198.40	6,405	misc
Essen Folk Song Dataset	ABC	56.62	9,034	folk
NES Music Database	MIDI	46.11	5,278	game
MusicNet Dataset	MIDI	30.36	323	classical
Hymnal Tune Dataset	MIDI	18.74	1,756	hymn
Hymnal Dataset	MIDI	17.50	1,723	hymn
music21's Corpus	misc	16.86	613	misc
EMOPIA Dataset	MIDI	10.98	387	pop
Nottingham Database	ABC	10.54	1,036	folk
music21's JSBach Corpus	MusicXML	3.46	410	classical
JSBach Chorale Dataset	MIDI	3.21	382	classical
Haydn Op.20 Dataset	Humdrum	1.26	24	classical

(Source: MusPy Documentation)

muspy.readthedocs.io/en/stable/datasets/datasets.html

Four Paradigms of Music Generation

Symbolic music generation

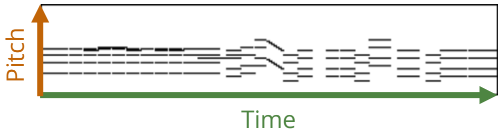
Text-based

```
Program_change_0,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_76, Time_shift_2, Note_off_67,
Note_on_67, Time_shift_2, Note_off_67,
...
```

MIDI

So far!

Image-based



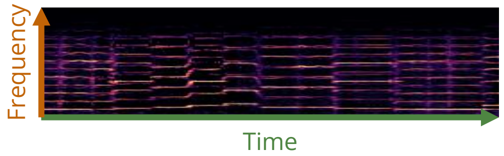
Piano roll

Audio-domain music generation

Time series-based

Waveform

Image-based

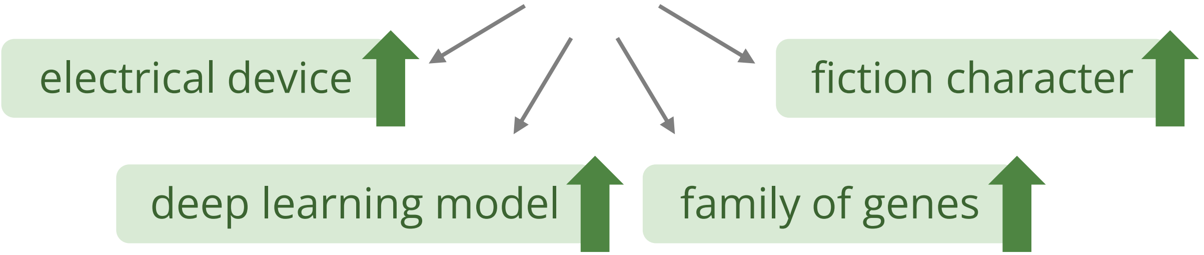


Spectrogram

Recap

Self-attention Mechanism (Cheng et al., 2016)

A transformer is a ________

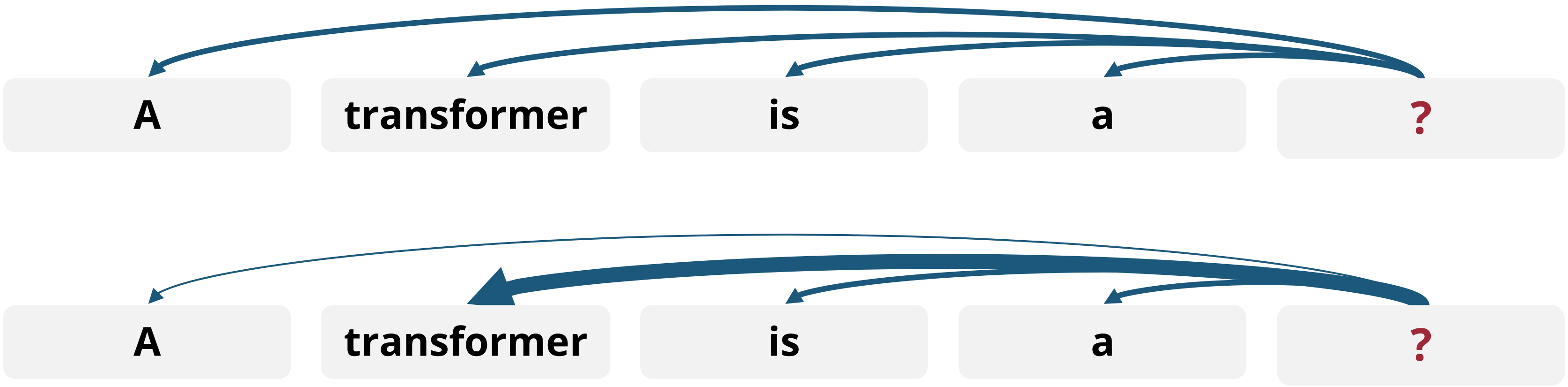


electrical device

fiction character

deep learning model

family of genes



Uniform attention

A	transformer	is	a	?

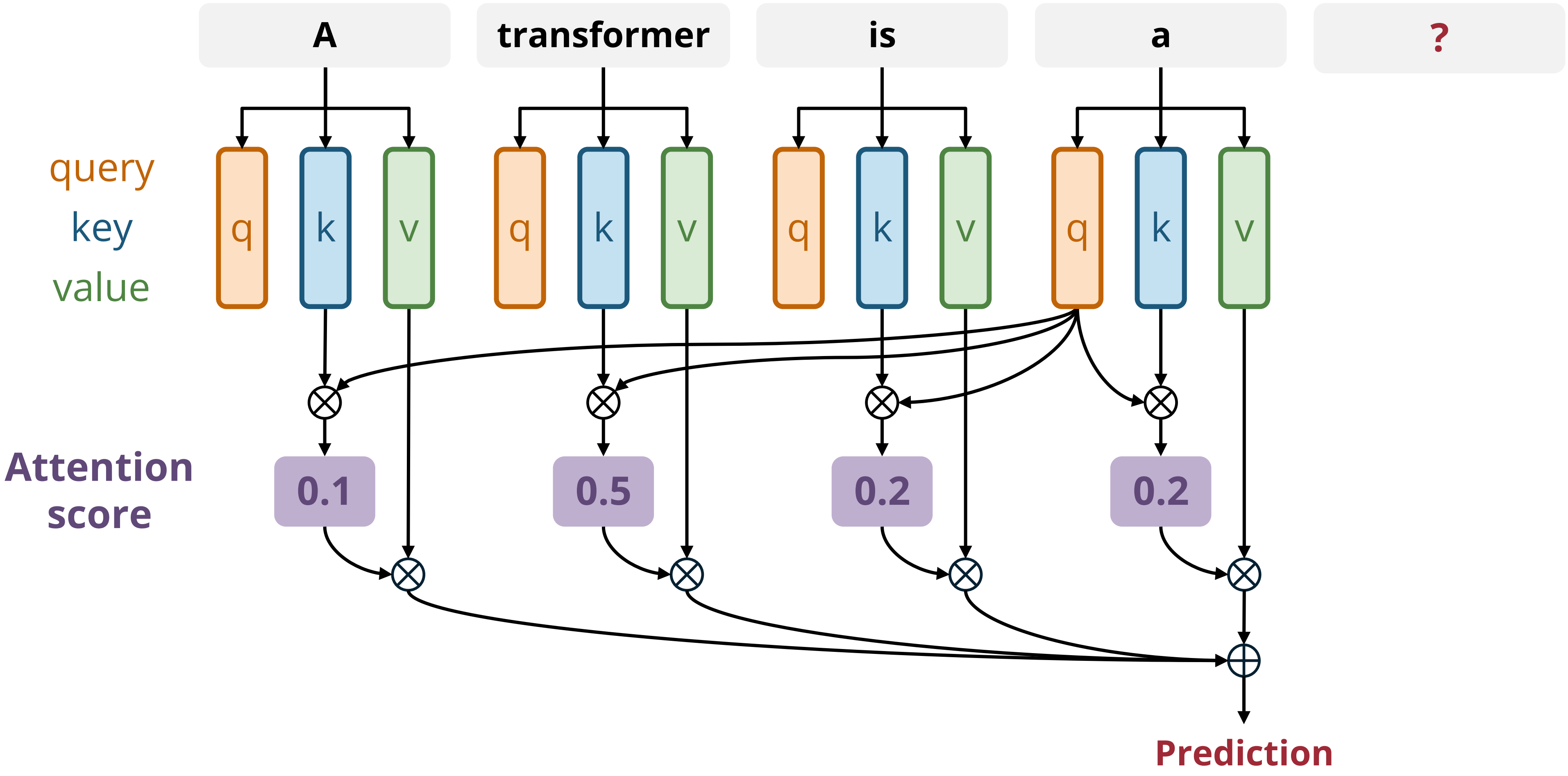
Variable attention

A	transformer	is	a	?

Transformers learn what to attend to from big data!

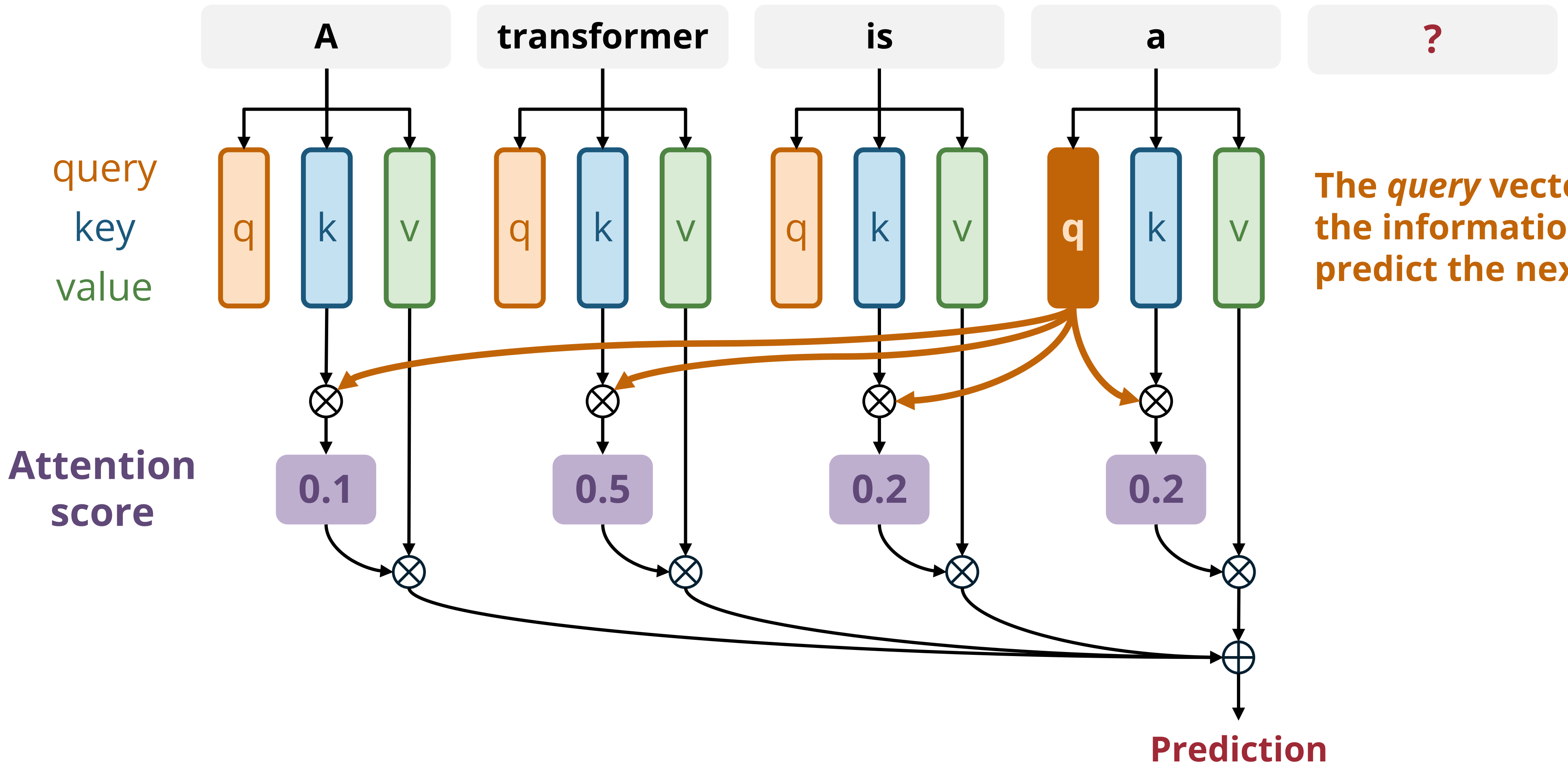
Jianpeng Cheng, Li Dong, and Mirella Lapata, "Long Short-Term Memory-Networks for Machine Reading," *EMNLP*, 2016.

Demystifying Transformers (Vaswani et al., 2017)



Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," *NeurIPS*, 2017.

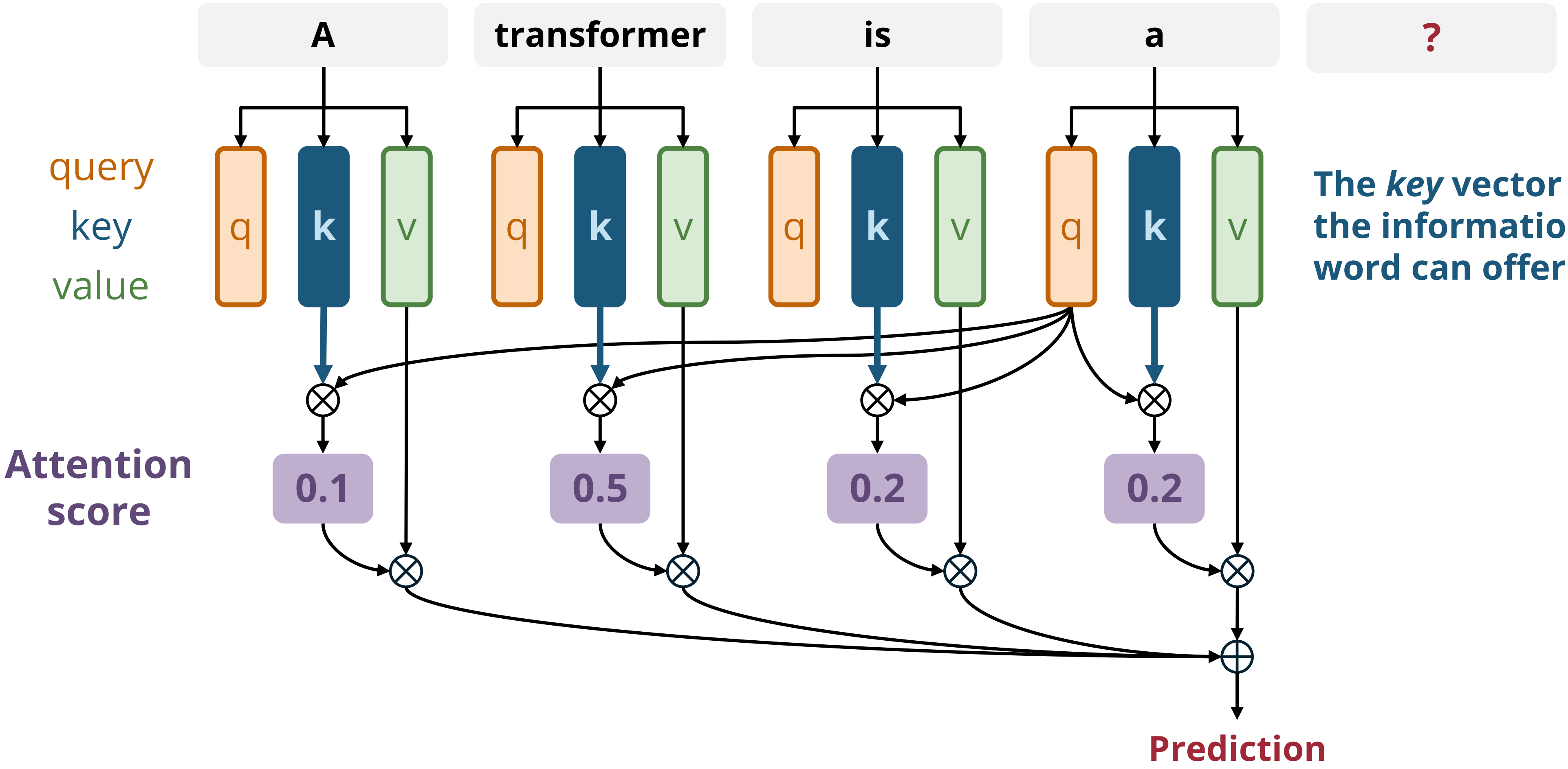
Demystifying Transformers (Vaswani et al., 2017)



The *query* vector captures the information needed to predict the next word.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," *NeurIPS*, 2017.

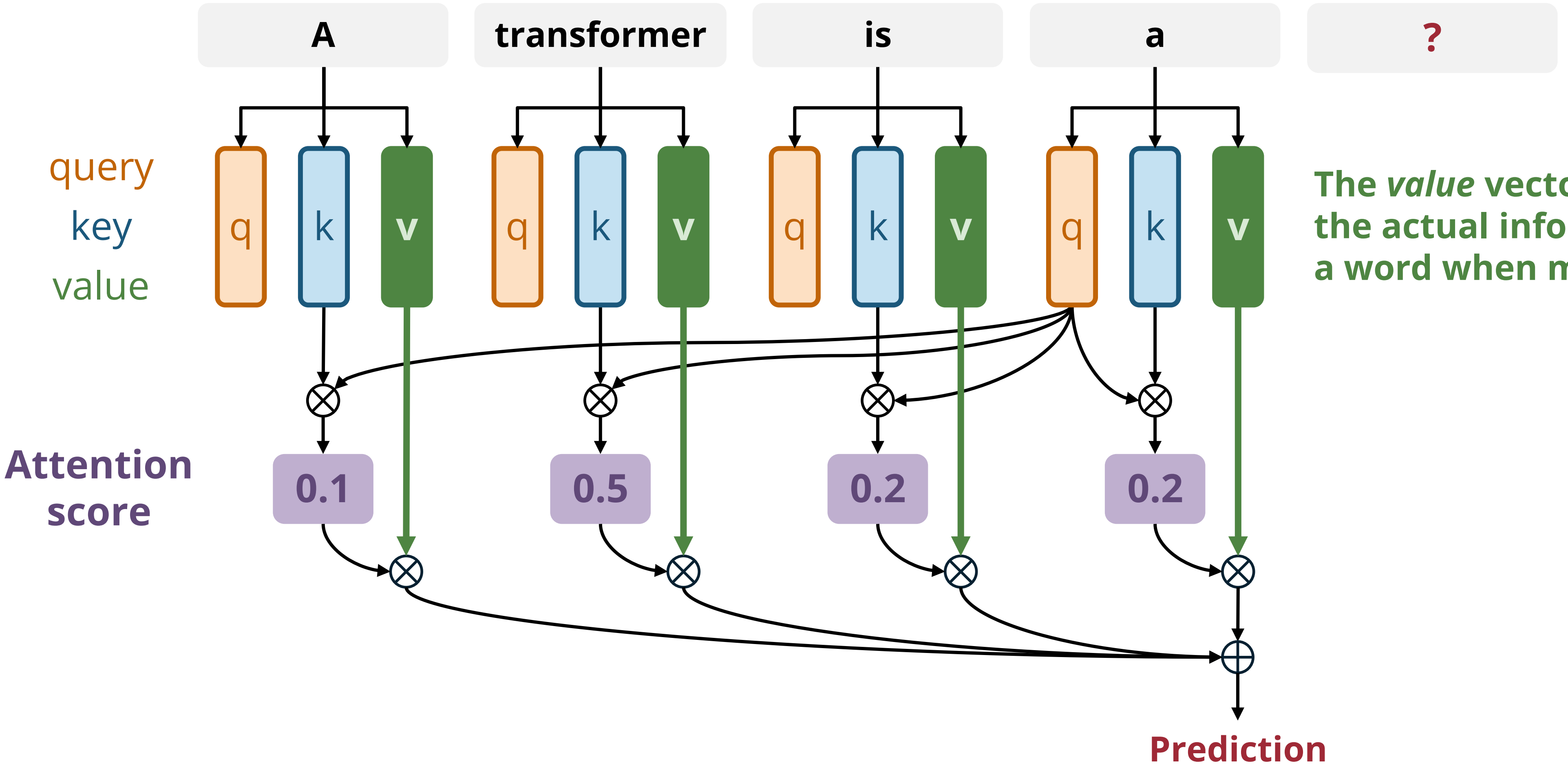
Demystifying Transformers (Vaswani et al., 2017)



The *key* vector captures the information that a word can offer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," *NeurIPS*, 2017.

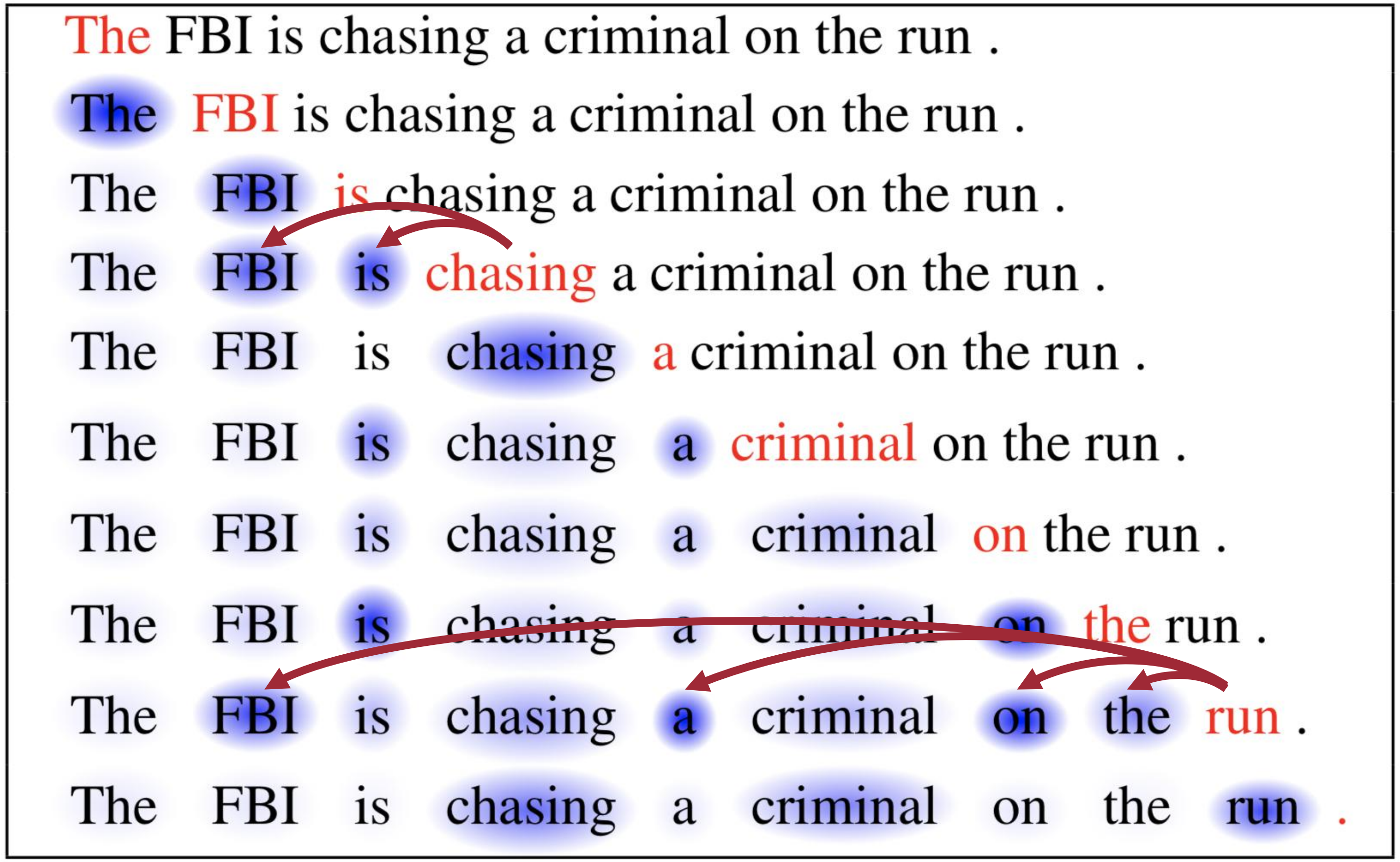
Demystifying Transformers (Vaswani et al., 2017)



The *value* vector captures the actual information of a word when matched.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," *NeurIPS*, 2017.

Why Self-Attention Mechanism?



The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

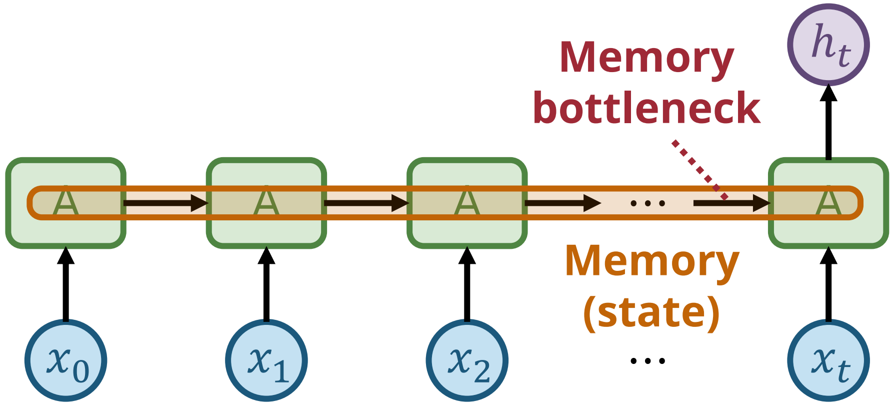
The FBI is chasing a criminal on the run .

(Source: Cheng et al., 2016)

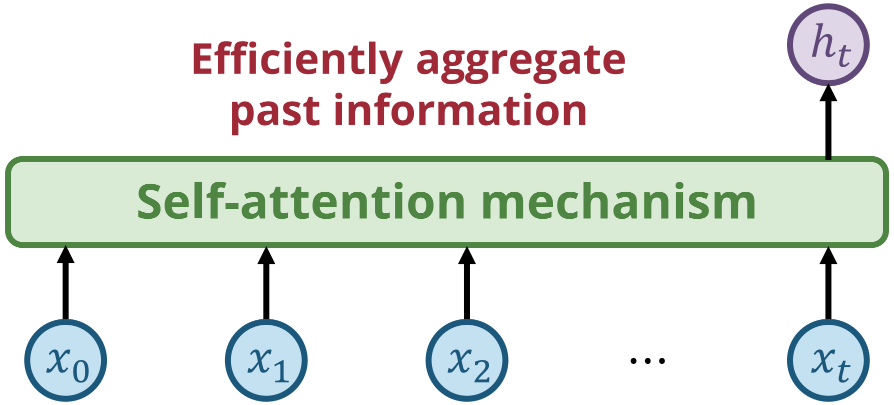
Jianpeng Cheng, Li Dong, and Mirella Lapata, "Long Short-Term Memory-Networks for Machine Reading," *EMNLP*, 2016.

RNN vs. Transformer

RNN



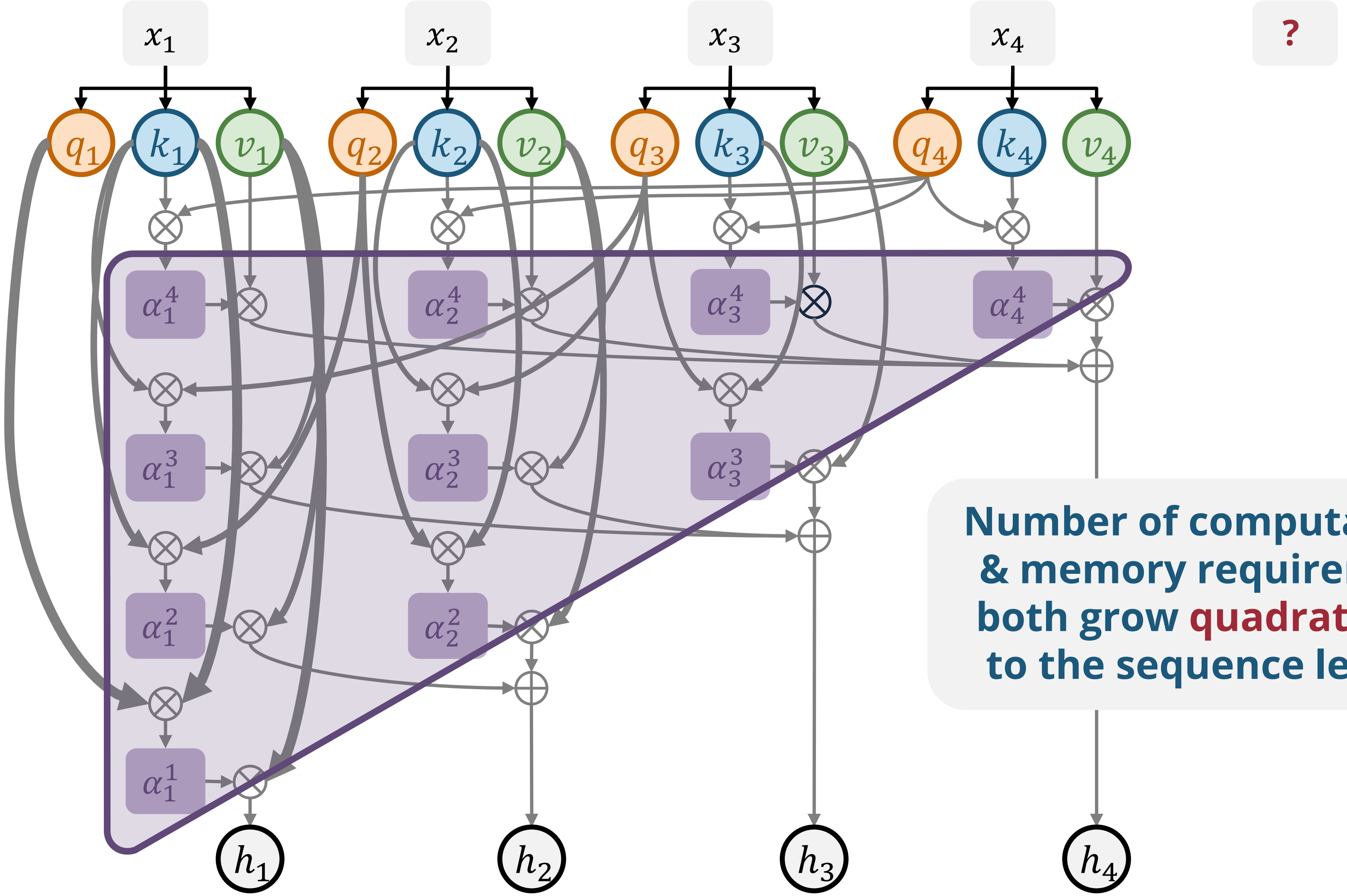
Transformer



Pros: Requires less GPU memory
Cons: Memory bottleneck

Pros: Alleviate memory bottleneck constraints
Cons: Requires more GPU memory

Demystifying Transformers (Vaswani et al., 2017)

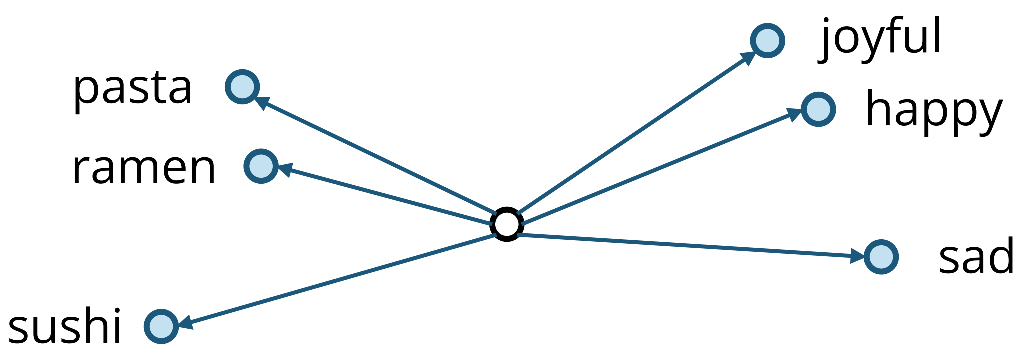


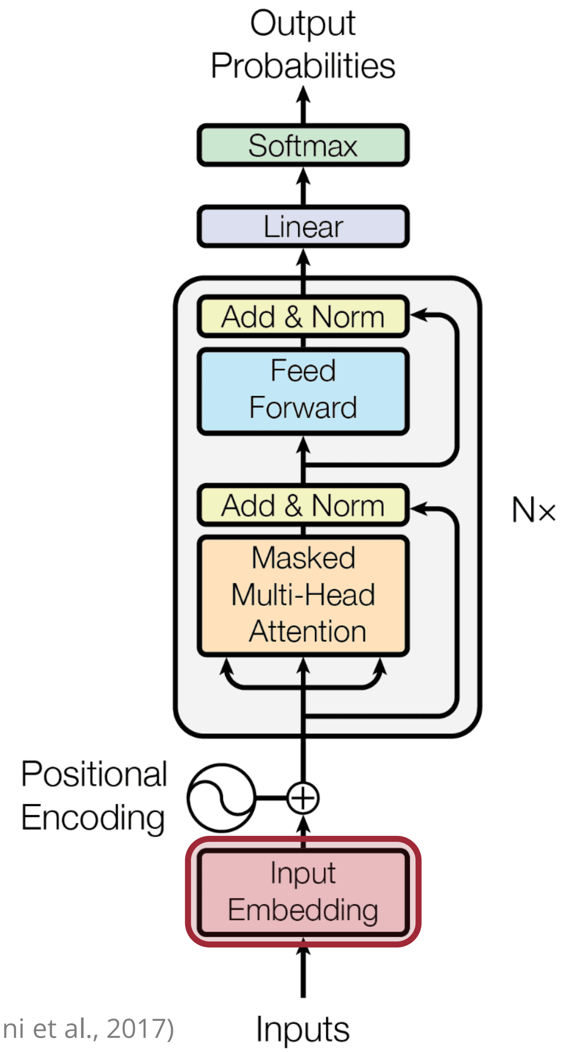
Number of computations & memory requirement both grow quadratically to the sequence length

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," *NeurIPS*, 2017.

Word Embedding

- **Goal**: Learn to **represent words as vectors**
- **Intuition**: Synonyms should have close embeddings
- Should antonyms be far apart?
  - Not quite, antonyms usually fall in the same "topic"
  - For example, "happy" & "sad" are both emotions



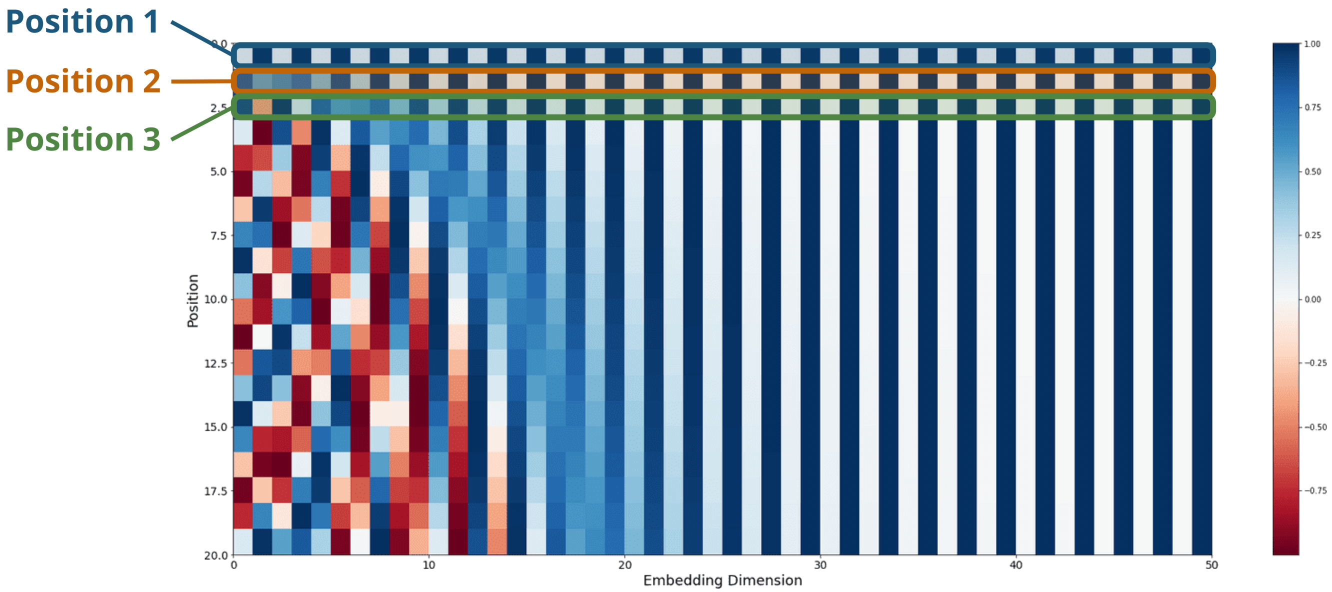


(Source: Vaswani et al., 2017)

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," *NeurIPS*, 2017.

Positional Encoding

- **Intuition**: A word could have **different meanings at different positions**
- Positional encoding provides positional information of the words to the model

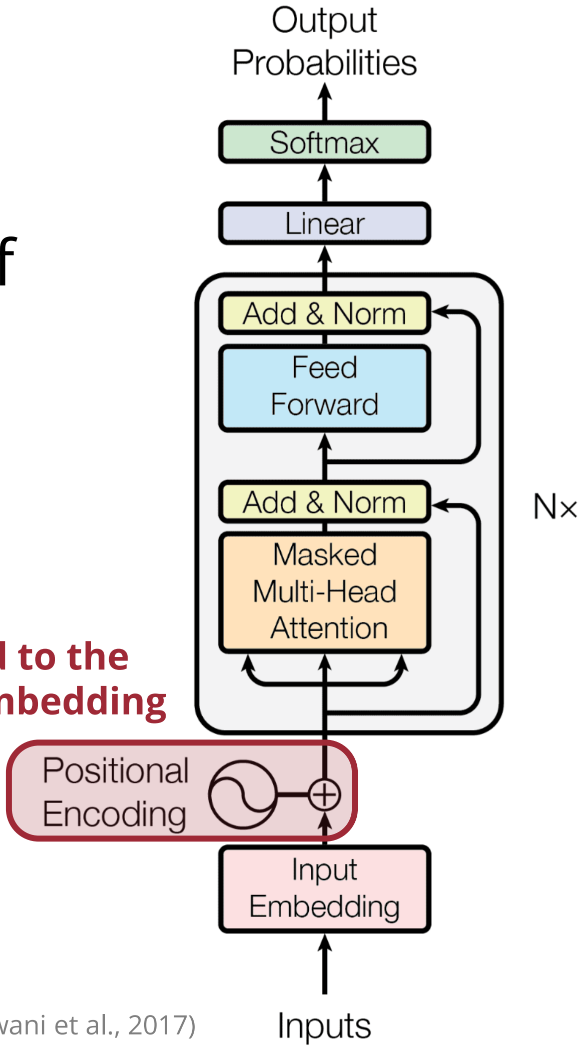


Position 1

Position 2

Position 3

(Source: erdem.pl)



Added to the word embedding

(Source: Vaswani et al., 2017)

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," *NeurIPS*, 2017.
erdem.pl/2021/05/understanding-positional-encoding-in-transformers

Music Transformer (Huang et al., 2019)

- **Data**
  - Yamaha e-Piano Competition dataset (MAESTRO)
- **Representation**
  - 128 Note-On events
  - 128 Note-Off events
  - 100 Time-Shift events (10ms–1s)
  - 32 Set-Velocity events
- **Model**
  - Transformer

Almost the same representation as PerformanceRNN

Expressive timing

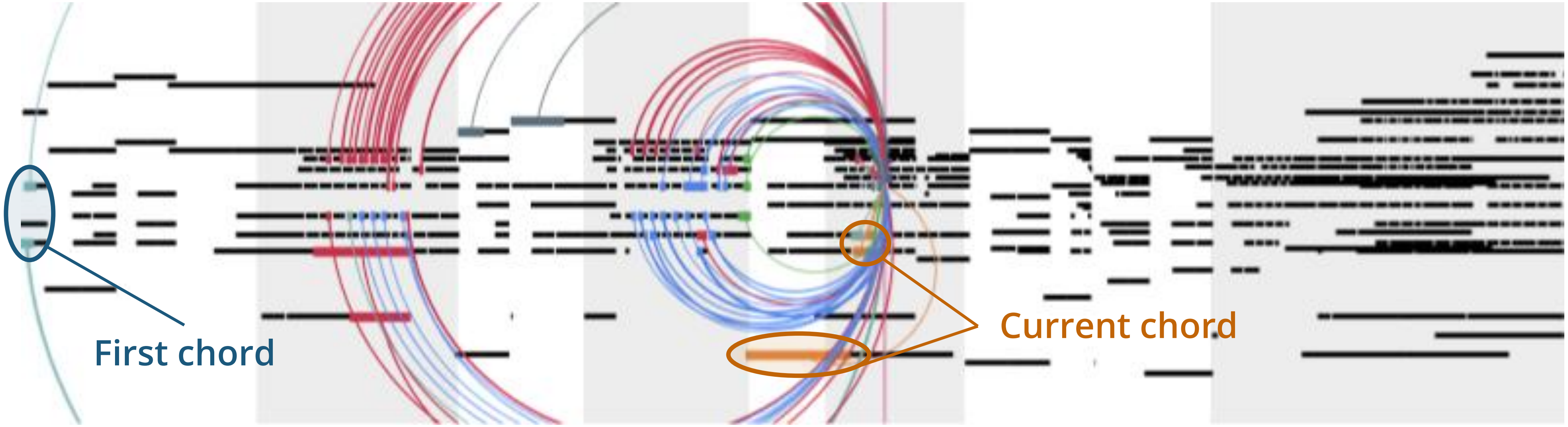
Expressive dynamics

Examples of generated music

Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck, "Music Transformer: Generating Music with Long-Term Structure," *ICLR*, 2019.

Visualizing Musical Self-attention (Huang et al., 2018)

(Each color represents an attention head)



First chord

Current chord

(Source: Huang et al., 2018)

Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck, "Music Transformer: Generating Music with Long-Term Structure," *ICLR*, 2019.
Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck, "Music Transformer: Generating Music with Long-Term Structure," *Magenta Blog*, December 13, 2018.

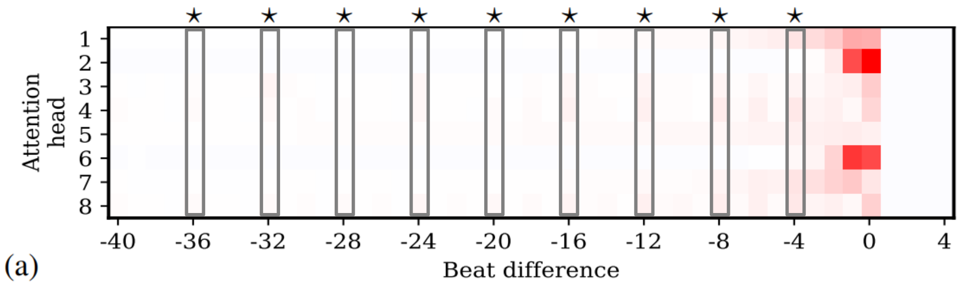
Analyzing Musical Self-attention (Dong et al., 2023)

- Measuring **mean relative attention**

$$\gamma_k^{(d)} = \frac{\sum_{\mathbf{x}\in\mathcal{D}} \sum_{s>t} a_{s,t}(\mathbf{x})\, \mathbb{1}_{x_t^{(d)} - x_s^{(d)} = k}}{\sum_{\mathbf{x}\in\mathcal{D}} \sum_{s>t} a_{s,t}(\mathbf{x})} \qquad \tilde{\gamma}_k^{(d)} = \gamma_k^{(d)} - \frac{\sum_{\mathbf{x}\in\mathcal{D}} \sum_{s>t} \mathbb{1}_{x_t^{(d)} - x_s^{(d)} = k}}{\sum_{\mathbf{x}\in\mathcal{D}} \sum_{s>t} 1}$$

- The MMT model attends more to notes

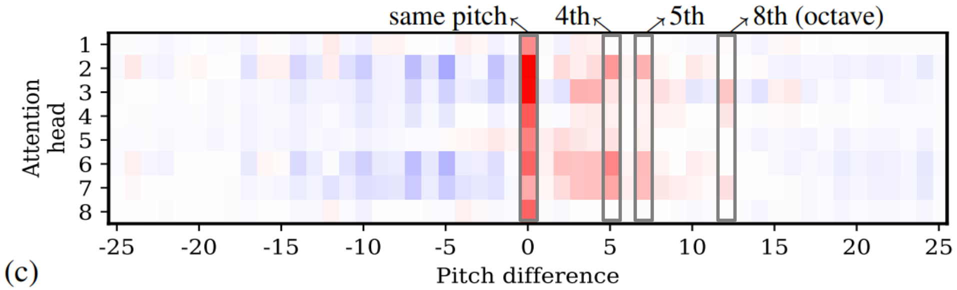
that are ***4N* beats away** in the past



Positive/negative gain

(Source: Dong et al., 2023)

that has a pitch in an octave above which **forms a consonant interval**



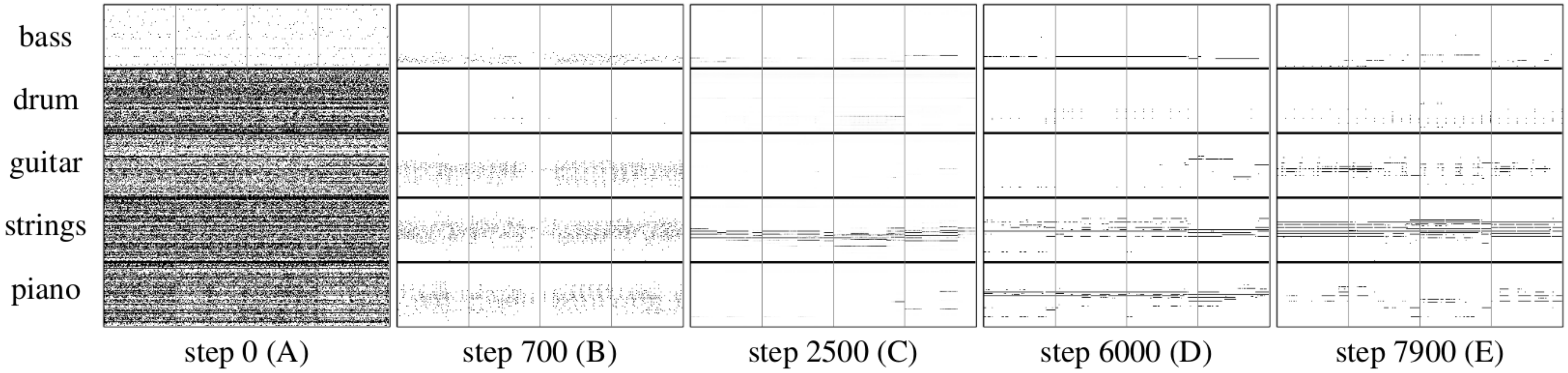
Positive/negative gain

(Source: Dong et al., 2023)

Hao-Wen Dong, Ke Chen, Shlomo Dubnov, Julian McAuley, and Taylor Berg-Kirkpatrick, "Multitrack Music Transformer," *ICASSP*, 2023.

Next Lecture

VAEs & GANs



(Source: Dong et al., 2018)