PAT 498/598 (Fall 2024)

Special Topics: Generative AI for Music and Audio Creation

Lecture 12: Language-based Music Generation

Instructor: Hao-Wen Dong

SCHOOL OF MUSIC, THEATRE & DANCE
PERFORMING ARTS TECHNOLOGY
UNIVERSITY OF MICHIGAN

Assignment 1: AI Song Contest

- **Q1**: Which is your favorite song? What did they do well? What can be improved?
- **Q2**: What is one dimension that most finalists didn't look into or didn't do well on?
- **Q3**: What tasks are easy for current AI? What are difficult?

Assignment 1 Discussions – Favorite Songs

- **1 vote** for "**Echoes of the Synthetic Forest**" by **KeRa**
- **1 vote** for "**One Mantra**" by **DJ Swami**
- **2 votes** for "**Genre Cannon**" by **Dadabots**
- **3 votes** for "**binary b1o0d**" by **HEL9000**

Assignment 1 Discussions – Limitations

- "... the artists **excused poor decision making by AIs** because of the novelty of the process."
- "... did not have **key modulations, tempo shifts, or very clearly demarcated distinct structures**, in either their lyrical or sonic content"
- "**Long-term musical development**, both in terms of the song structure and the evolution of musical ideas."
- "... **emotional depth**, particularly in the **vocal performances**. ... the music struggled to connect on a more human, emotional level."
- "... didn't explore deeply is the **generation of music form and cohesion** by AI. ... AI is not yet adept at generating long, cohesive musical forms or handling transitions in a way that feels natural over time."

Assignment 1 Discussions – What are easy?

- "… **instrumental timbre emulation**, **vocal emulation**, recreation of different **vocal mixing and production techniques** from different eras, **lyric creation**, **chord progression creation**"
- "… tasks that are highly repetitive or based on patterns that don't require complex reasoning. Generating **individual sound clips, synthesizing audio, and creating simple loops or short musical phrases** are relatively easy for AI."
- "**Generating and processing material** is easy for AI."
- "AI models thrive in terms of **idea creation**."
- "… quick, mass generation of **short musical snippets** specific to the musical genre and instrumentation style they are trained on."

Assignment 1 Discussions – What are difficult?

- "... AI struggles with **tasks that require abstract thinking** or the ability to generate **high-level structures**. ... AI's challenge lies in generating coherent, dynamic compositions that can **maintain interest over time without sounding repetitive**."
- "... the devil's in the details ... are **not refined enough** to the point that they can avoid those small flaws and fool the trained ears of experienced musicians or audio engineers."
- "**Refining material into a composition** is difficult for AI."
- "... **creating new harmonic and rhythmic patters different from the training data**, ..."
- "... **large-scale form, functional harmony** (when generating raw audio), **maintaining a tempo** in the way a human drummer would maintain a tempo, **sticking with a genre**, or **developing their compositional ideas** in ways familiar to human listeners"
- "**Lyric writing** remains challenging for AI."

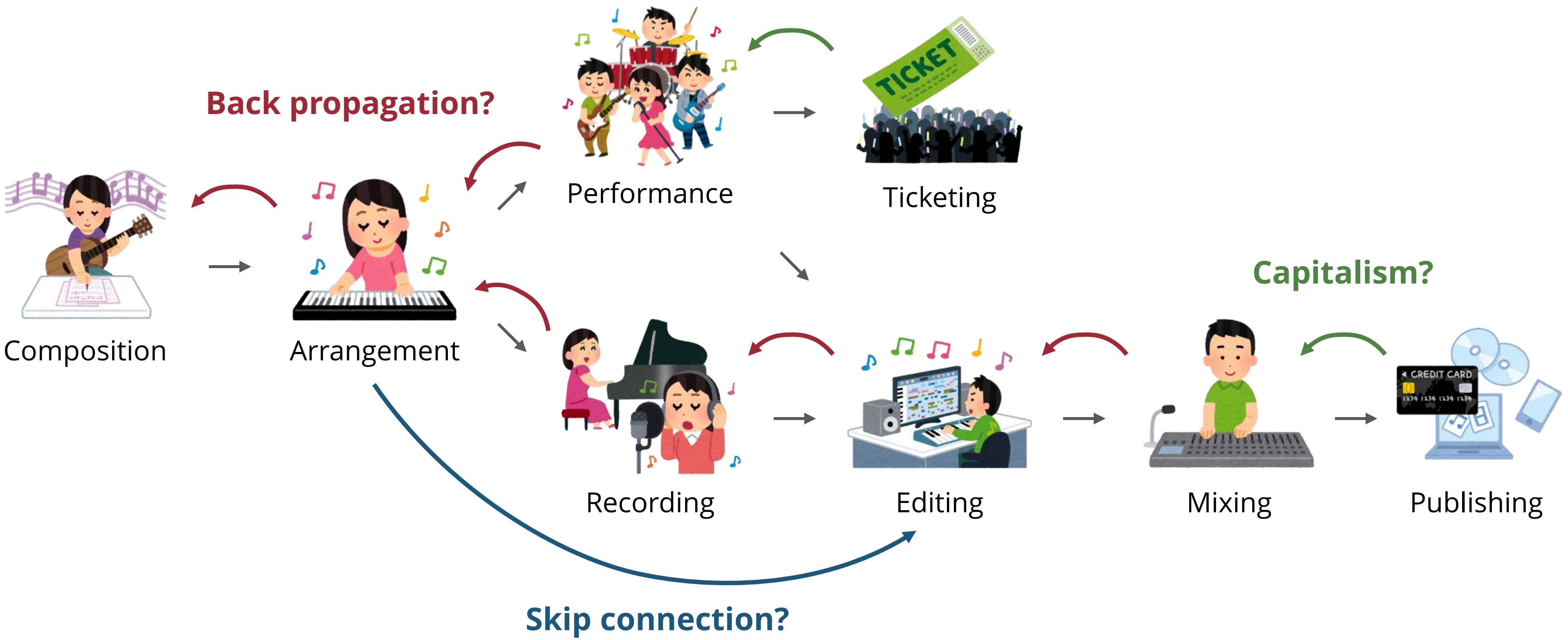
Discussions

- **To what extent of human involvements** can a song still be called AI music?
- **Shall we intervene** if AI-generated material doesn't sound polished?
- What is the **goal of AI music**?

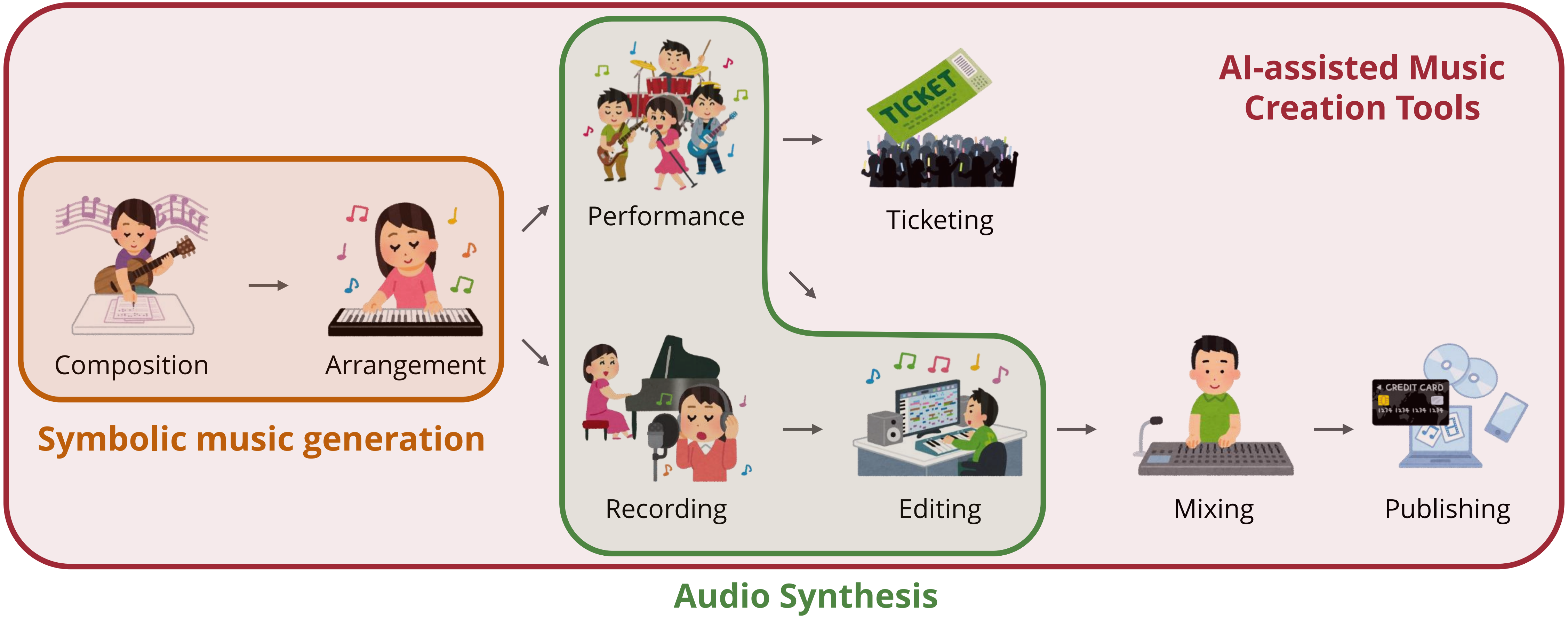
> "Whatever you now find weird, ugly, uncomfortable and nasty about a new medium will surely **become its signature**."
>
> – Brian Eno, 1996

The Landscape

A Simplified Music Production Workflow



A Simplified Music Production Workflow



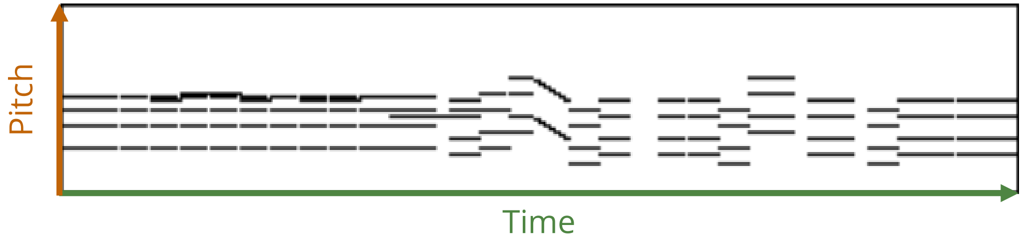
Music Generation – Symbolic vs Audio

Symbolic-domain

- MIDI

```
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, ...
```

- Piano-roll

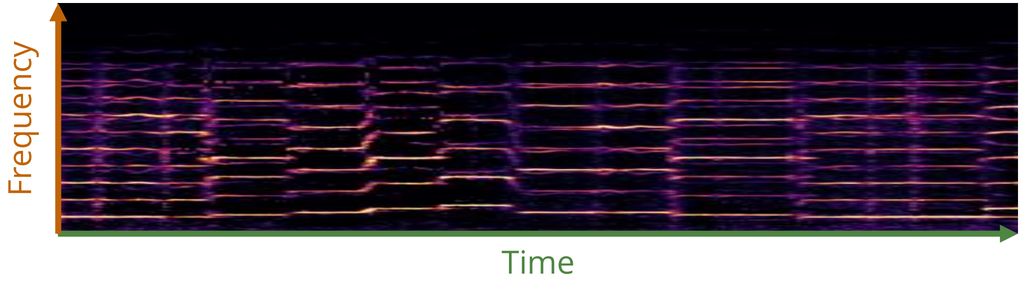


Today's topic!

Audio-domain

- Waveform

- Spectrogram



Symbolic Music Generation – Relevant Topics

Unconditional

Symbolic music generation
- ø → melody
- ø → lead sheet (Melody & chords)
- ø → sheet music

Today's topic!

Conditional

Automatic arrangement
- Melody → lead sheet
- Melody → multitrack
- Lead sheet → multitrack
- Solo → multitrack
- Multitrack → simple version

Performance rendering
- Sheet music → performance

Improvisation systems
- Performance → performance

Multimodal

X-to-music generation
- Text → sheet music
- Video → sheet music
- X → sheet music

Symbolic Music Generation – Two Main Approaches

Text-based

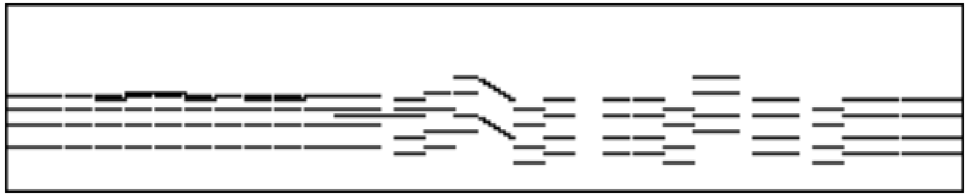
- Treat music like **text**
- Sharing models with **natural language processing (NLP)**
  - RNNs, LSTMs, Transformers, etc.

Today's topic!

```
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, ...
```

Image-based

- Treat music like **images**
- Sharing models with **computer vision (CV)**
  - GANs, VAEs, diffusion models, etc.



Four Paradigms

Symbolic music generation

Audio-domain music generation

Text-based

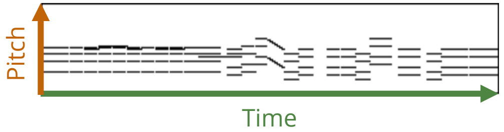
Image-based

Time series-based

Image-based

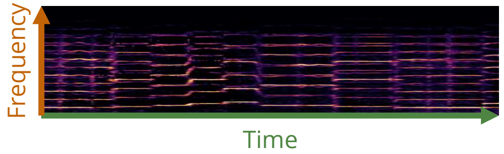
```
Program_change_0,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_76, Time_shift_2, Note_off_67,
Note_on_67, Time_shift_2, Note_off_67,
...
```

MIDI



Piano roll

Waveform



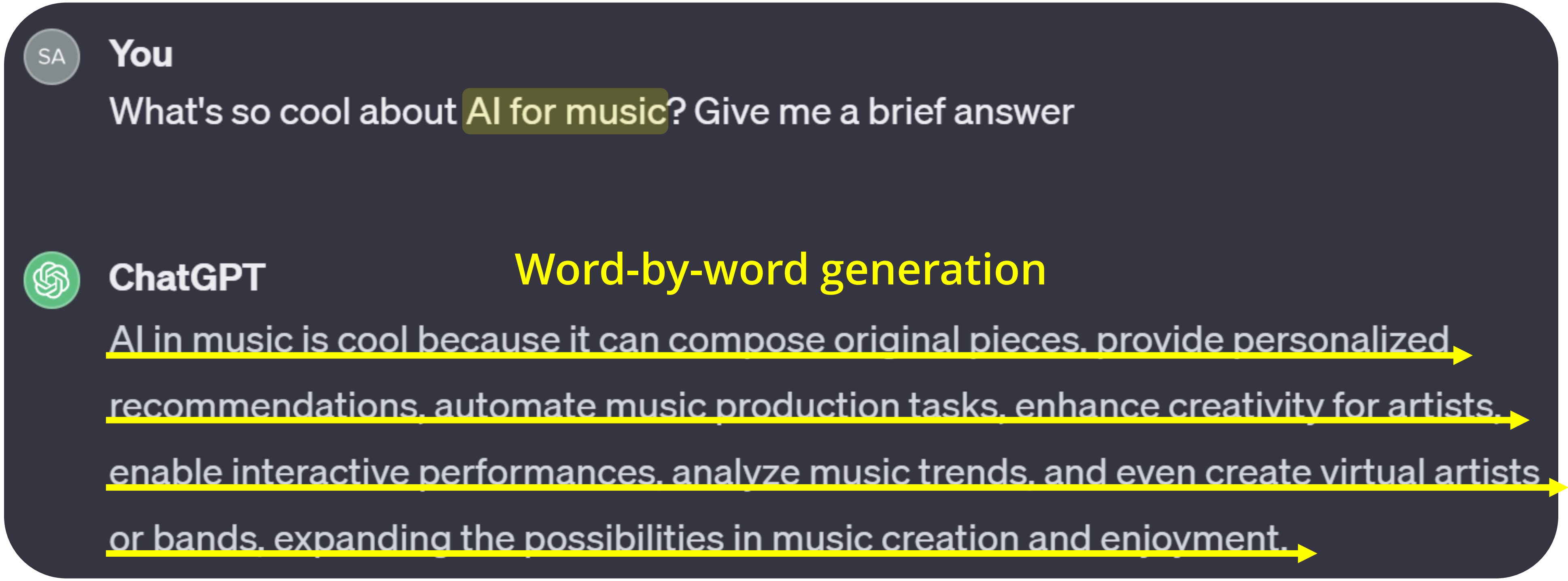
Spectrogram

Today, we also have many latent-space based systems!

Generating Music like Languages

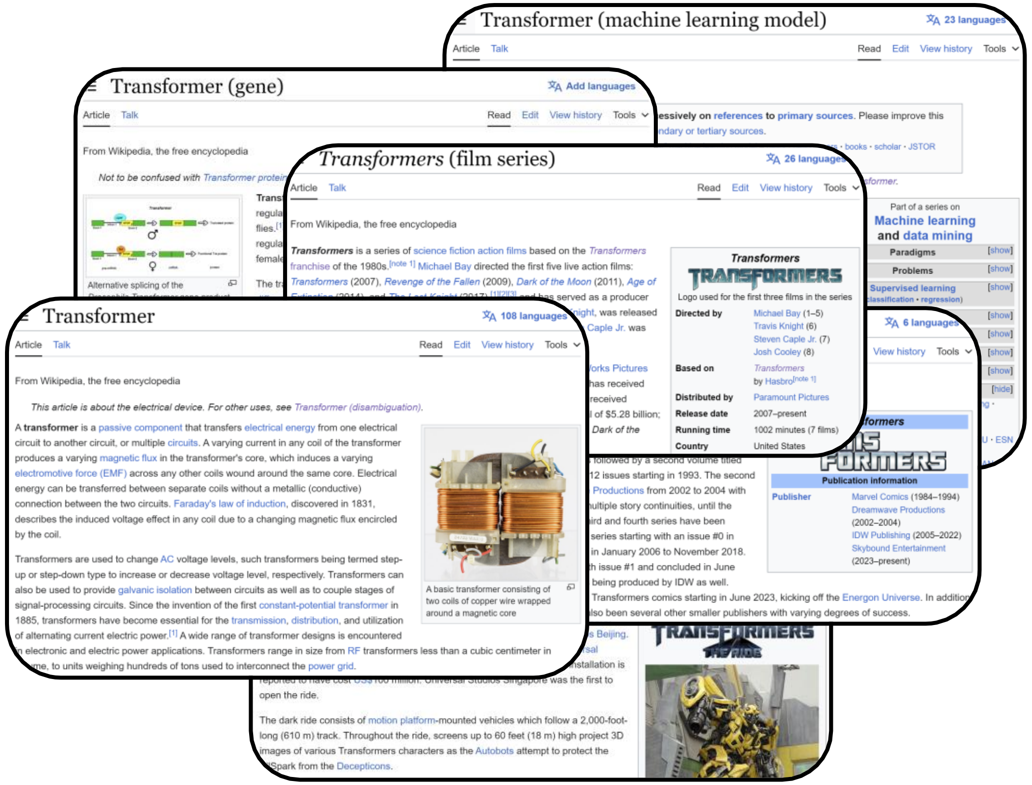
Large Language Models (LLMs)

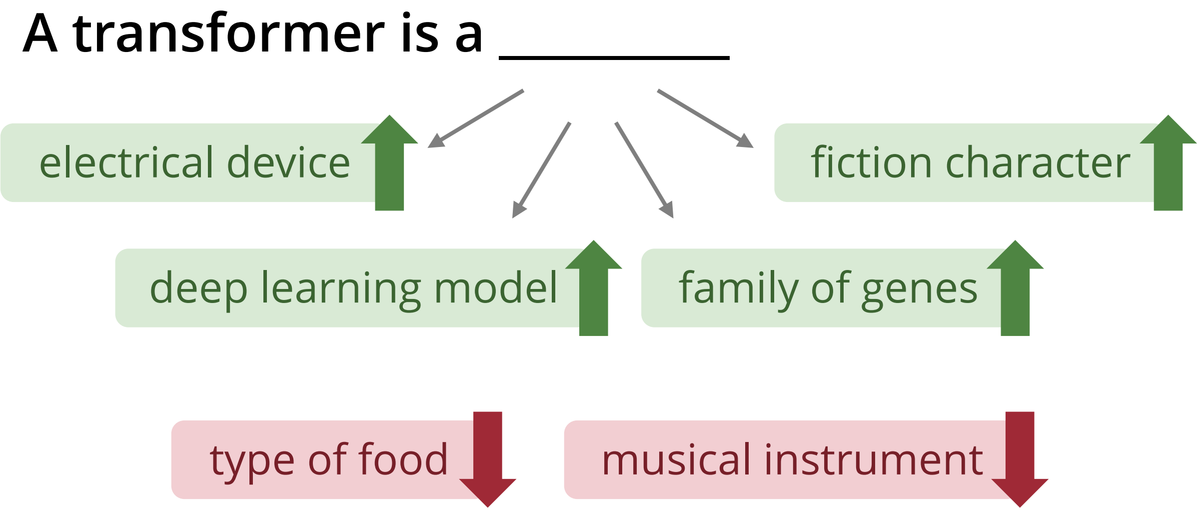
- The models behind ChatGPT!



Language Models

- Predicting the next word given the past sequence of words





Language Models (Mathematically)

- A class of machine learning models that learn the next word probability

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

$P(\text{electrical} \mid \text{A transformer is a})$ $\uparrow$

$P(\text{character} \mid \text{A transformer is a})$ $\uparrow$

$P(\text{gene} \mid \text{A transformer is a})$ $\uparrow$

$P(\text{model} \mid \text{A transformer is a})$ $\uparrow$

$P(\text{food} \mid \text{A transformer is a})$ $\downarrow$

$P(\text{musical} \mid \text{A transformer is a})$ $\downarrow$

Language Models – Generation

- How do we generate a new sentence using a trained language model?

Input				Output
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

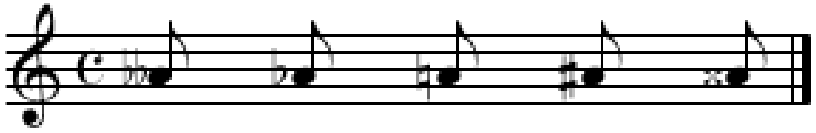
Designing a Machine-readable Music Language

- How can we "represent" music in a way that machines understand?
  - Musical representation is a key component of a music generation system
- Why not using sheet music "images" directly?
  - Machines still have a hard time reading sheet music
  - A challenging task known as "optical music recognition" (OMR)
- Examples:
  - ABC notation
  - MIDI

ABC Notation-based Representation

ABC Notation

- A simple text-based notation
- Use letters to denote pitches
  - Lower octave (A–G), higher octave (a–g)
- Use prefix to denote accidentals
  - Sharp (^), flat (_), natural (=)



__A _A =A ^A ^^A

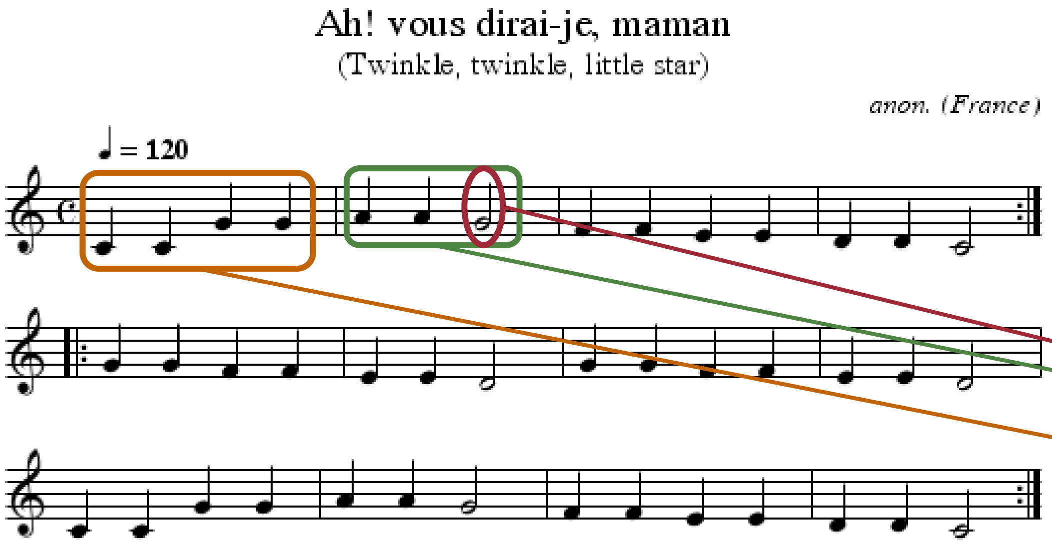
C, D, E, F,|G, A, B, C| D E F G | A B c d | e f g a | b c' d' e'| f' g' a' b'

What is this song in ABC notation?

```
CCGG|AAG2|FFEE|DDC2:|
 |:GGFF|EED2|GGFF|EED2|
CCGG|AAG2|FFEE|DDC2:|
```

Twinkle, twinkle, little star!

An Example of ABC Notation



Metadata

```
X:571
T:Ah! vous dirai-je, maman
T:(Twinkle, twinkle, little star)
C:anon.
O:France
R:Nursery song
```

M:C Meter

L:1/4 Unit note length (temporal resolution)

Q:120 Tempo

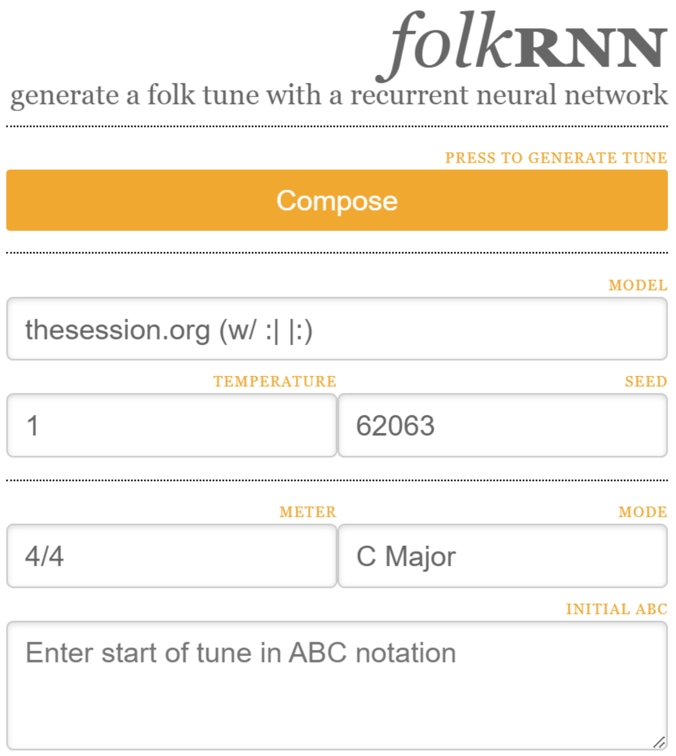
K:C Key

```
CCGG|AAG2|FFEE|DDC2:|
|:GGFF|EED2|GGFF|EED2|
CCGG|AAG2|FFEE|DDC2:|
```

abcnotation.com/tunePage?a=trillian.mit.edu/~jc/music/abc/mirror/musicaviva.com/france/ah-vous-c/ah-vous-c-1/0000

Example System: Folk RNN (Sturm et al., 2015)

- Data
  - Collections of folk tunes
- Representation
  - ABC notation without metadata
- Model
  - LSTM (long short-term memory)
  - Working on the character level

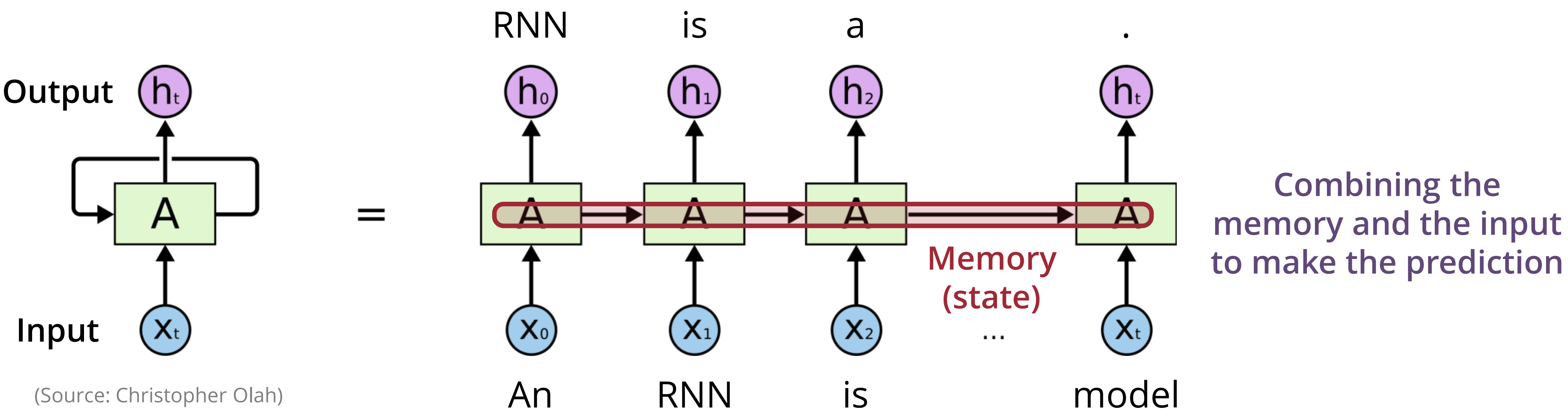


folkrnn.org

Bob L. Sturm, Joao Felipe Santos, and Iryna Korshunova, "Folk Music Style Modelling by Recurrent Neural Networks with Long Short Term Memory Units," *ISMIR Late-Breaking Demos*, 2015.

What is an RNN (Recurrent Neural Network)?

- A type of neural networks that have **loops**
- Widely used for **modeling sequences** (e.g., in natural language processing)

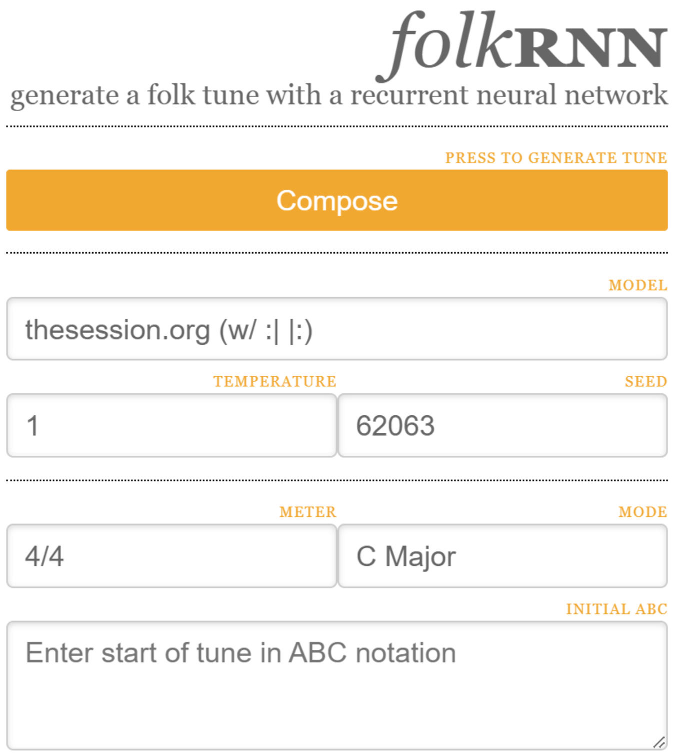


(Source: Christopher Olah)

colah.github.io/posts/2015-08-Understanding-LSTMs/

Example: Folk RNN (Sturm et al., 2015)

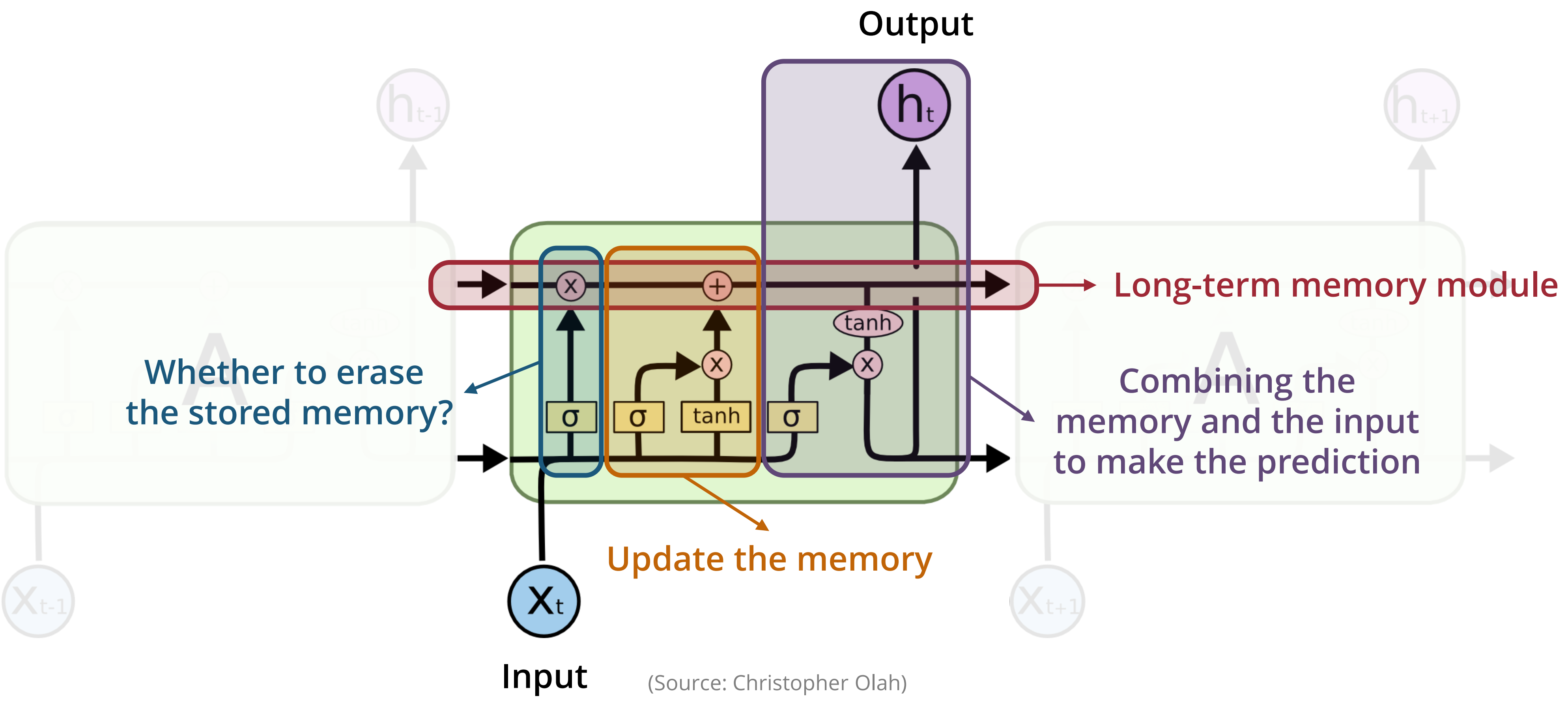
- Data
  - Collections of folk tunes
- Representation
  - ABC notation without metadata
- Model
  - LSTM (long short-term memory)
  - Working on the character level



folkrnn.org

Bob L. Sturm, Joao Felipe Santos, and Iryna Korshunova, "Folk Music Style Modelling by Recurrent Neural Networks with Long Short Term Memory Units," *ISMIR Late-Breaking Demos*, 2015.

(Recap) Demystifying LSTMs



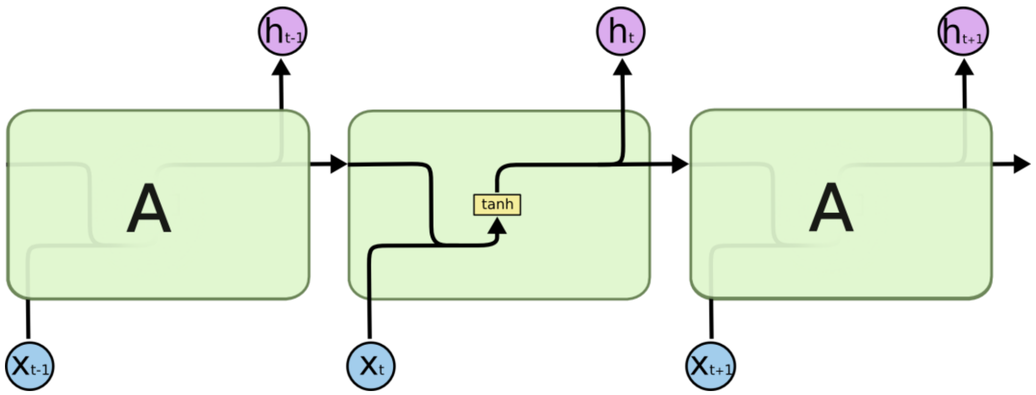
(Source: Christopher Olah)

colah.github.io/posts/2015-08-Understanding-LSTMs/

(Recap) Vanilla RNNs vs LSTMs

Vanilla RNN

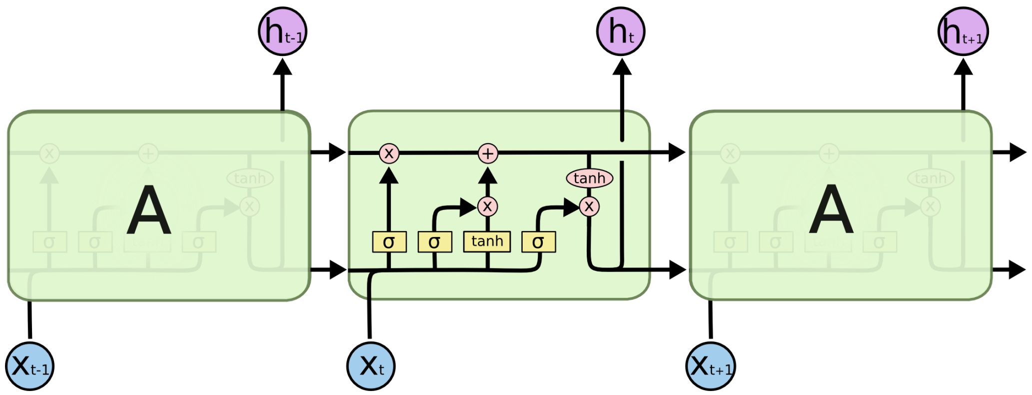
- Simplest form of RNNs
- Limited long-term memory



(Source: Christopher Olah)

LSTM

- Improved memory module
- Better long-term memory

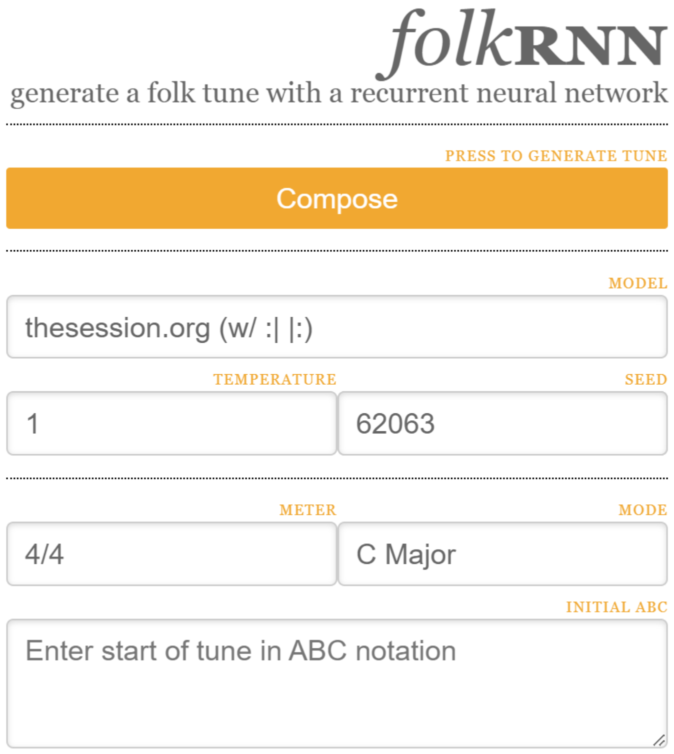


(Source: Christopher Olah)

colah.github.io/posts/2015-08-Understanding-LSTMs/

Example: Folk RNN (Sturm et al., 2015)

- Data
  - Collections of folk tunes
- Representation
  - ABC notation without metadata
- Model
  - LSTM (long short-term memory)
  - Working on the character level



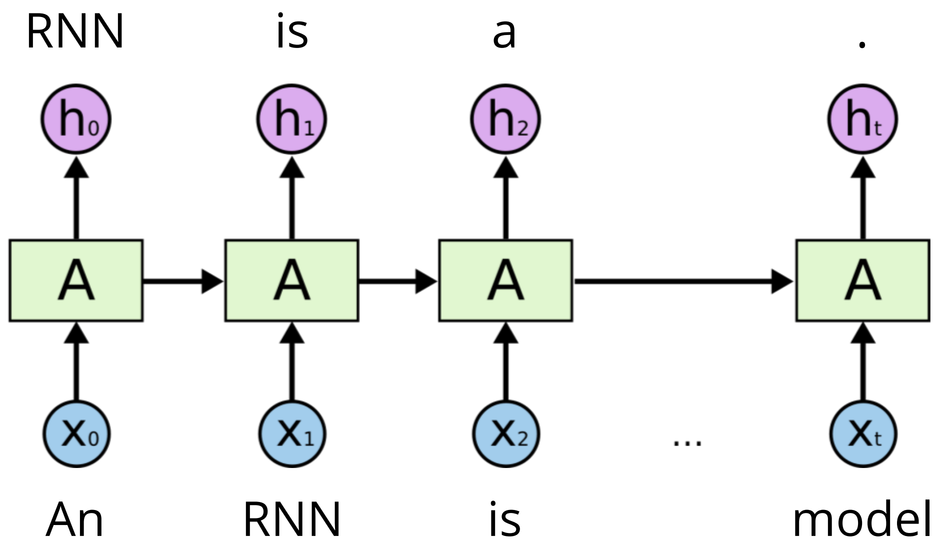
folkrnn.org

Bob L. Sturm, Joao Felipe Santos, and Iryna Korshunova, "Folk Music Style Modelling by Recurrent Neural Networks with Long Short Term Memory Units," *ISMIR Late-Breaking Demos*, 2015.

Word-level vs Character-level RNNs

Word-level RNNs

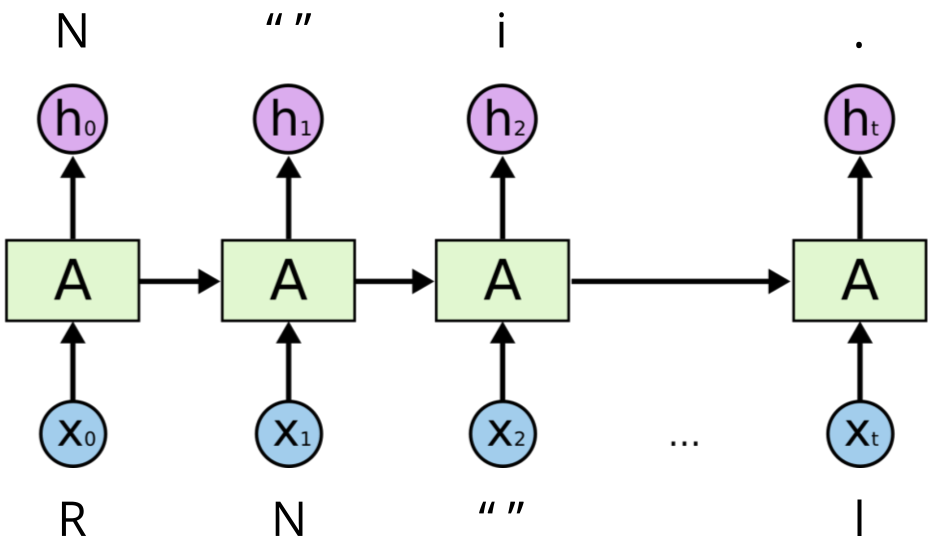
- Predicting word by word
- Most common



(Source: Christopher Olah)

Character-level RNNs

- Predicting character by character
- Useful when there is no natural “spaces”



(Source: Christopher Olah)

colah.github.io/posts/2015-08-Understanding-LSTMs/

ChatGPT can also Compose Songs in ABC Notation!

You
Write a cheerful folk song using ABC notation

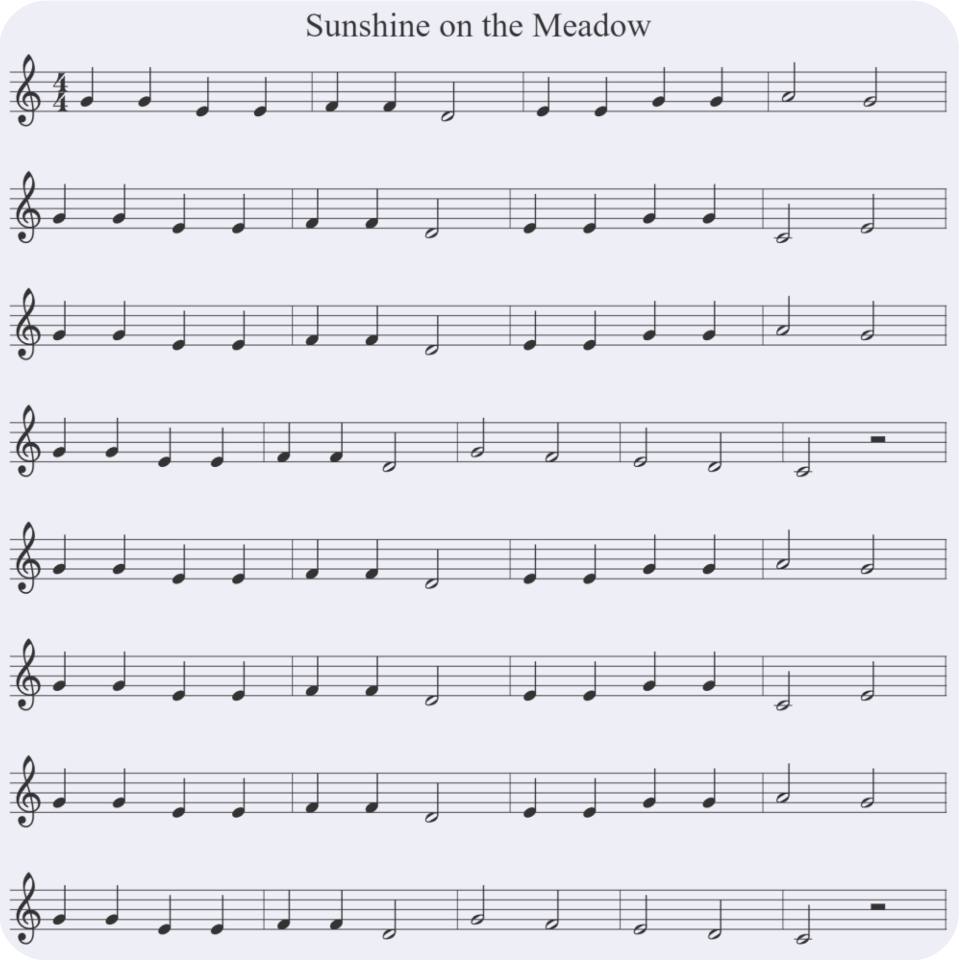
ChatGPT
Certainly! Here's a cheerful folk song in ABC notation:

vbnet   Copy code

```
X:1
T:Sunshine on the Meadow
M:4/4
K:C
G2 G2 E2 E2 | F2 F2 D4 | E2 E2 G2 G2 | A4 G4 |
G2 G2 E2 E2 | F2 F2 D4 | E2 E2 G2 G2 | C4 E4 |
G2 G2 E2 E2 | F2 F2 D4 | E2 E2 G2 G2 | A4 G4 |
G2 G2 E2 E2 | F2 F2 D4 | G4 F4 | E4 D4 | C4 z4 |
G2 G2 E2 E2 | F2 F2 D4 | E2 E2 G2 G2 | A4 G4 |
G2 G2 E2 E2 | F2 F2 D4 | E2 E2 G2 G2 | C4 E4 |
G2 G2 E2 E2 | F2 F2 D4 | E2 E2 G2 G2 | A4 G4 |
G2 G2 E2 E2 | F2 F2 D4 | G4 F4 | E4 D4 | C4 z4 |
```

Feel free to play around with the tempo and instrumentation to give it your own unique style!

chat.openai.com



The sheet music is rendered with abc.rectanglered.com.

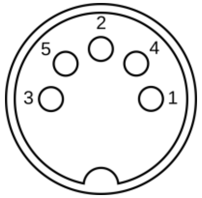
Limitations of ABC Notations

- Limited expressiveness
- Monophonic tunes only

MIDI-like Representation

MIDI (Musical Instrument Digital Interface)

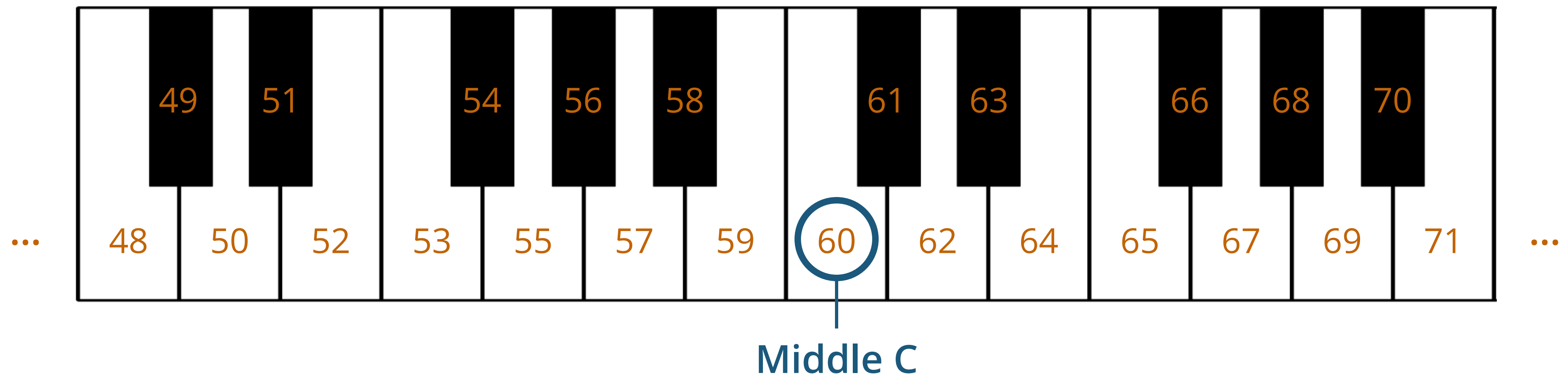
- A communication protocol between devices
- MIDI Messages
  - Note on
  - Note off
  - Delta time
  - Program change
  - Control change
  - Pitch bend change



MIDI I/O

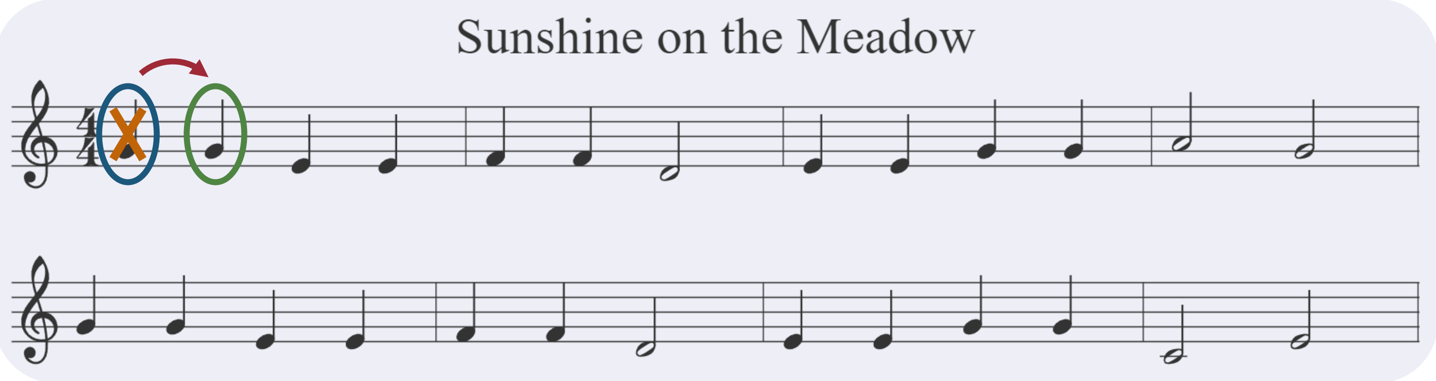
MIDI Note Numbers

- Ranging from 0 to 127
  - Middle C is 60
  - Wider than standard piano's pitch range
- Widely used in various software, keyboards and algorithms



Representing Music using MIDI Messages

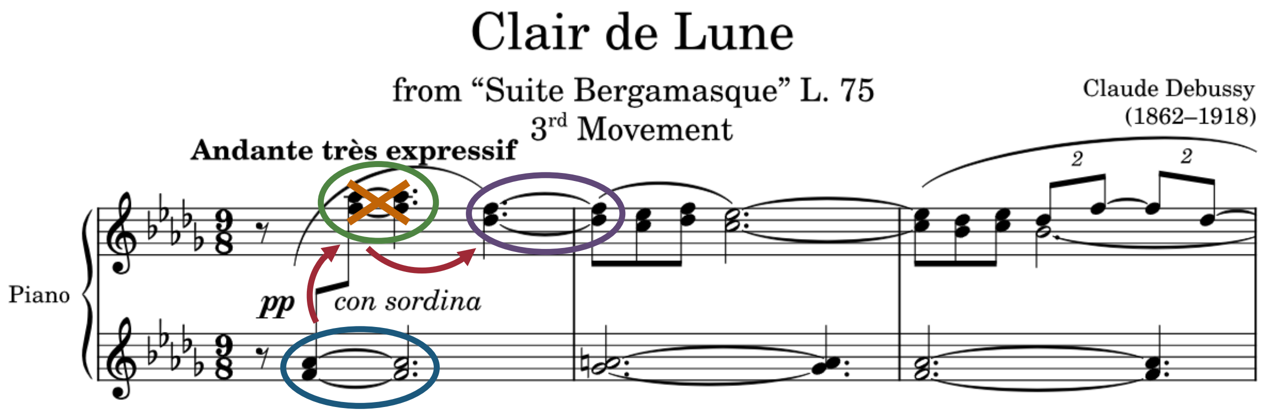
- Three main MIDI messages
  - Note on
  - Note off
  - Time Shift




```
Note_on_67, Time_shift_quarter_note, Note_off_67,
Note_on_67, Time_shift_quarter_note, Note_off_67,
Note_on_64, Time_shift_quarter_note, Note_off_64,
Note_on_64, Time_shift_quarter_note, Note_off_64,
...
```

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, ...
```

Example: Performance RNN (Oore et al., 2020)

- Data
  - Yamaha e-Piano Competition dataset (MAESTRO)
- Representation
  - 128 Note-On events
  - 128 Note-Off events
  - 125 Time-Shift events (8ms–1s)
  - 32 Set-Velocity events

Handle dynamics

- Model
  - LSTM

Examples of generated music

Ian Simon and Sageev Oore, "Performance RNN: Generating Music with Expressive Timing and Dynamics," *Magenta Blog*, June 29, 2017.
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.

Example: A.I. Duet (Mann et al., 2016)

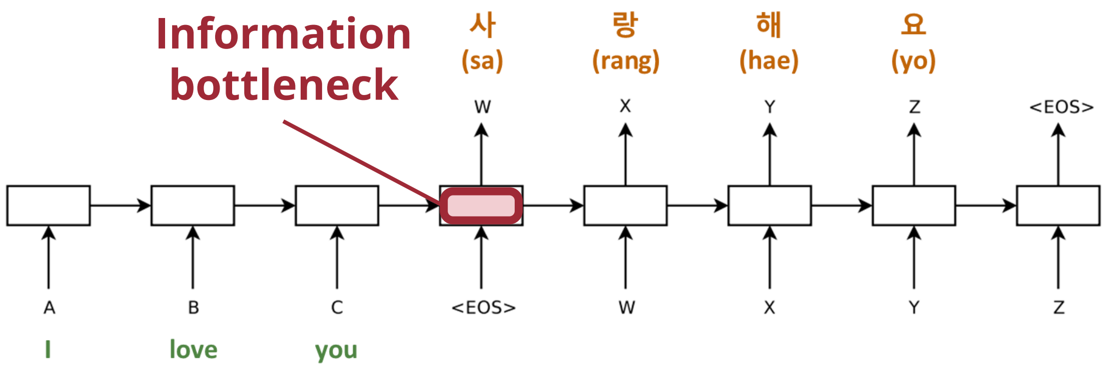
youtu.be/0ZE1bfPtvZo

experiments.withgoogle.com/ai/ai-duet/view/

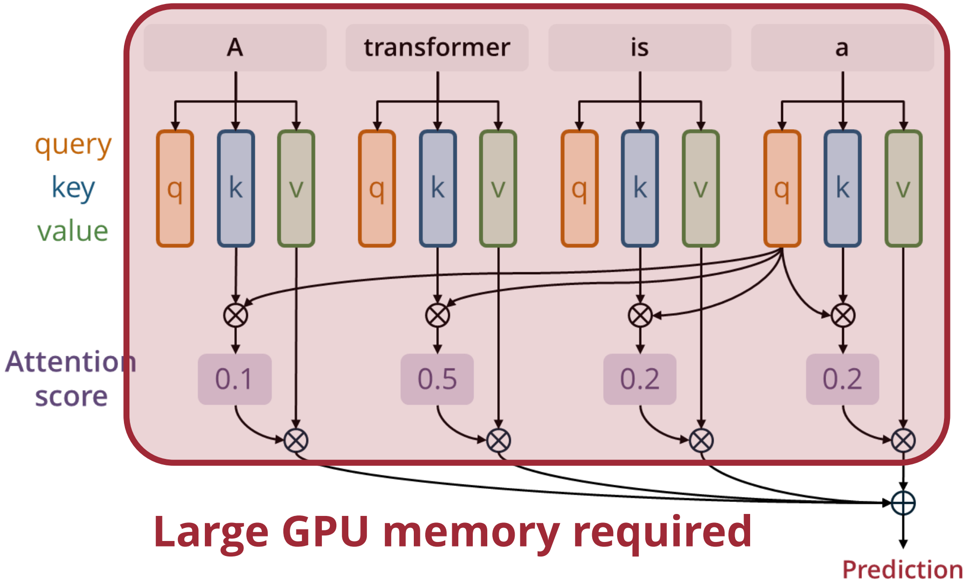
github.com/googlecreativelab/aiexperiments-ai-duet

(Recap) Seq2seq vs Transformers

Seq2seq



Transformers



Example: 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

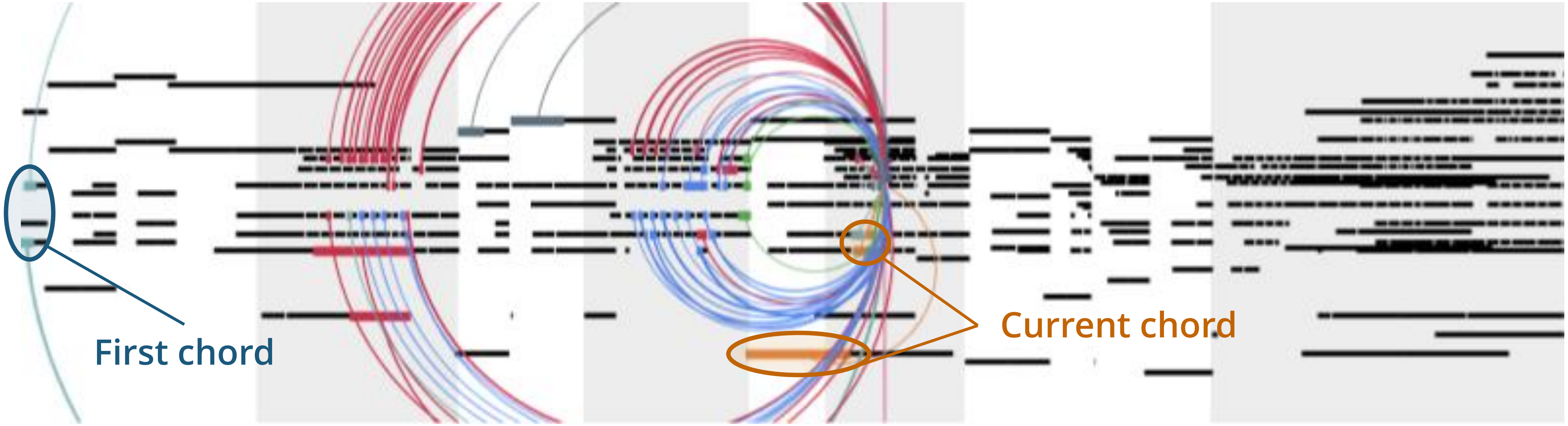
Handle dynamics

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.
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

(Each color represents an attention head)

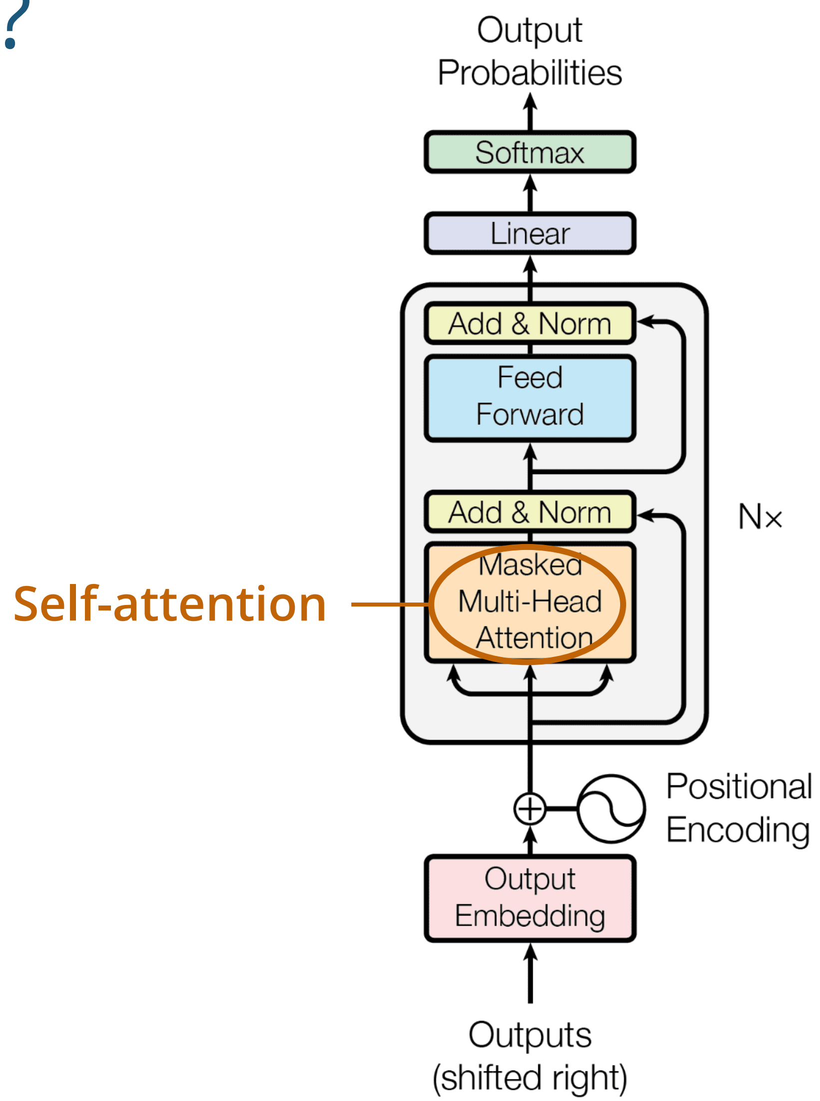


(Source: Huang et al., 2018)

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," *Magenta Blog*, December 13, 2018.

(Recap) What is a Transformer?

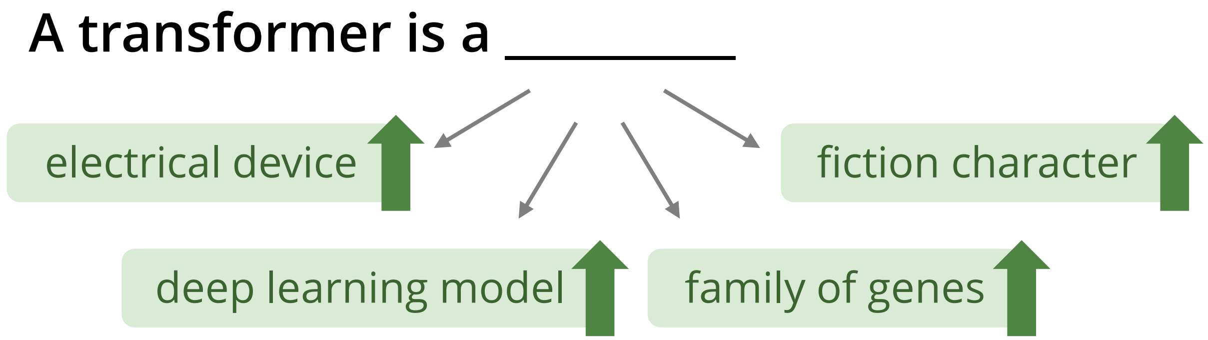
- A type of neural network that use the **self-attention mechanism**

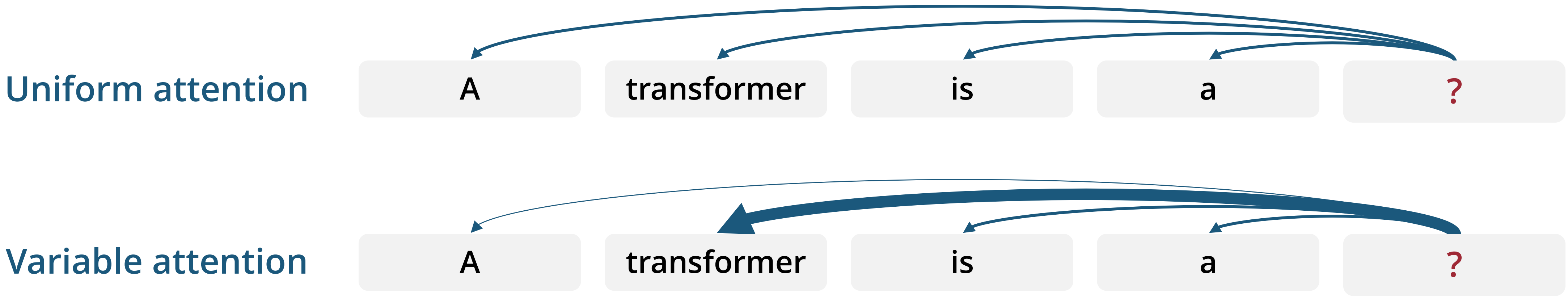


(Source: Vaswani et al., 2017; adapted)

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.

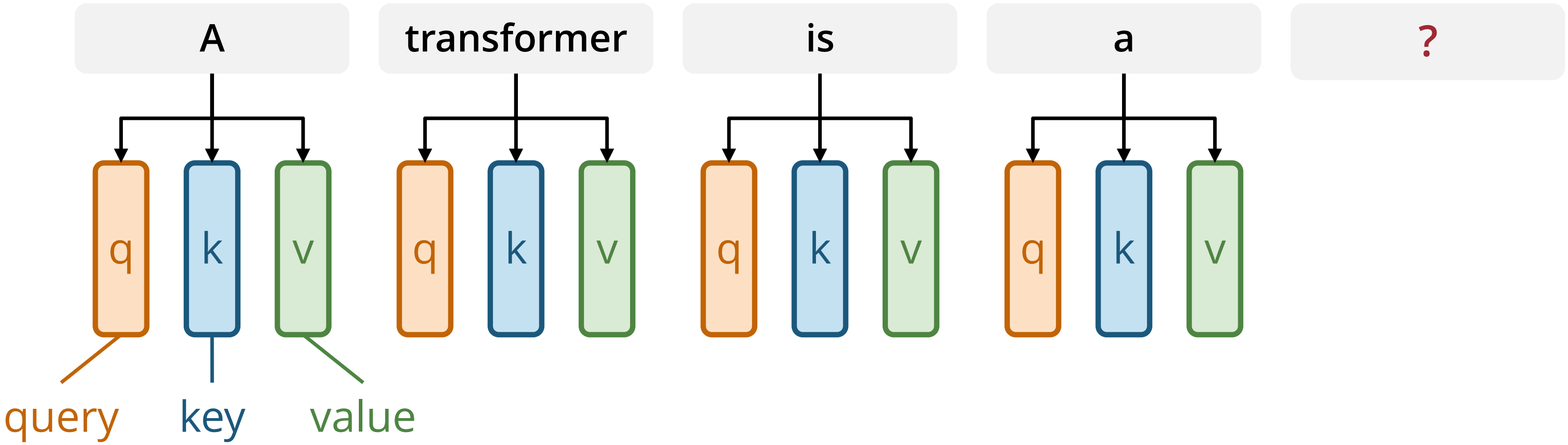
(Recap) Self-attention Mechanism



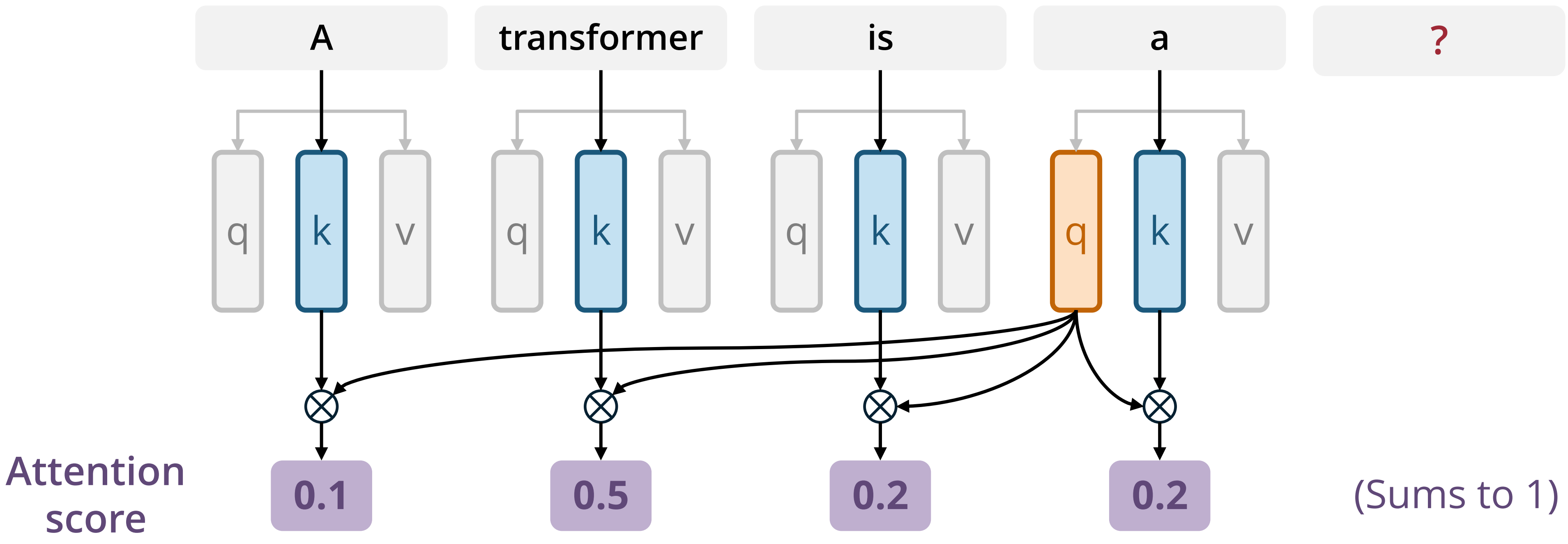


Transformers learn what to attend to from big data!

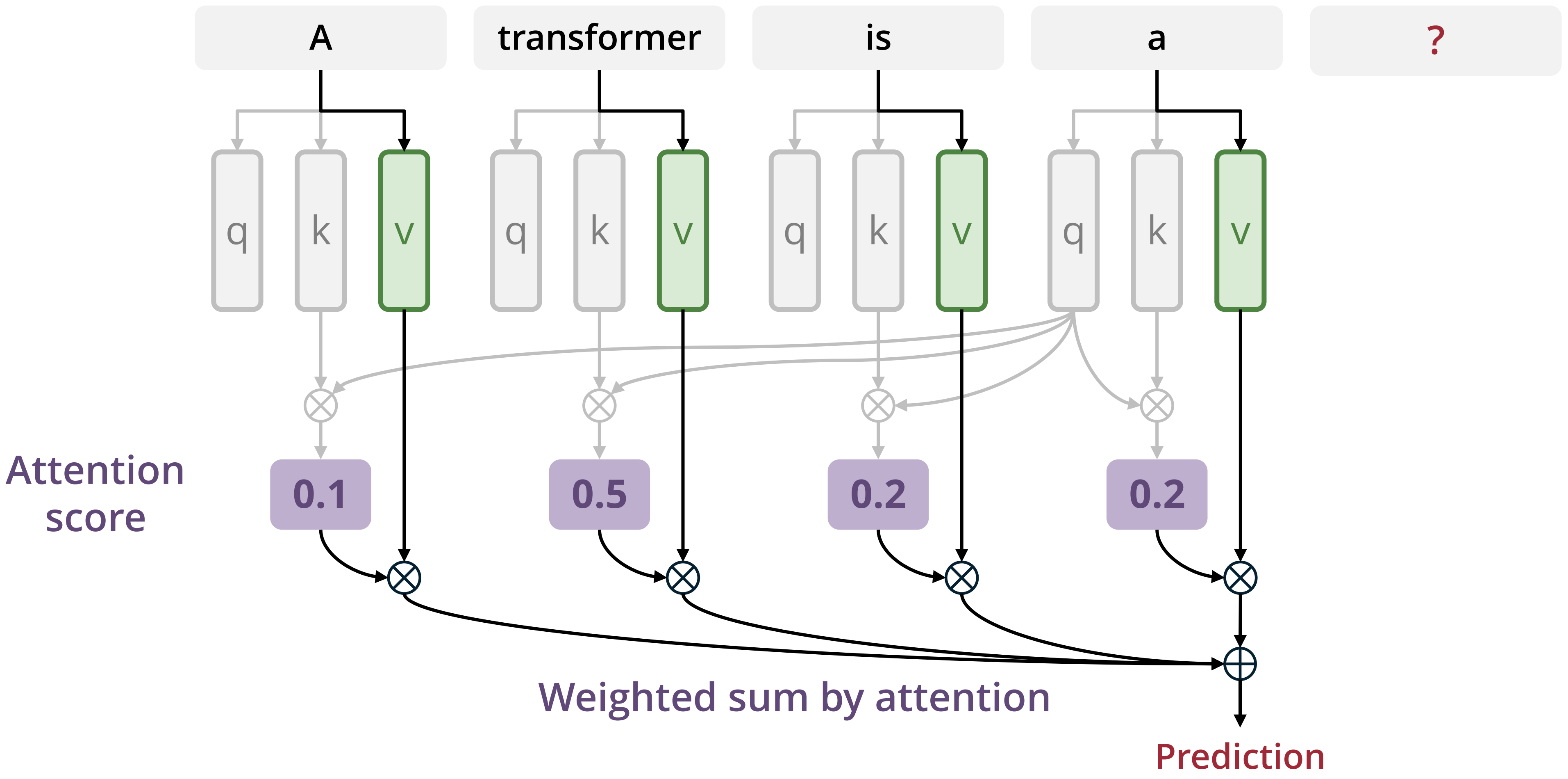
(Recap) Demystifying Transformers



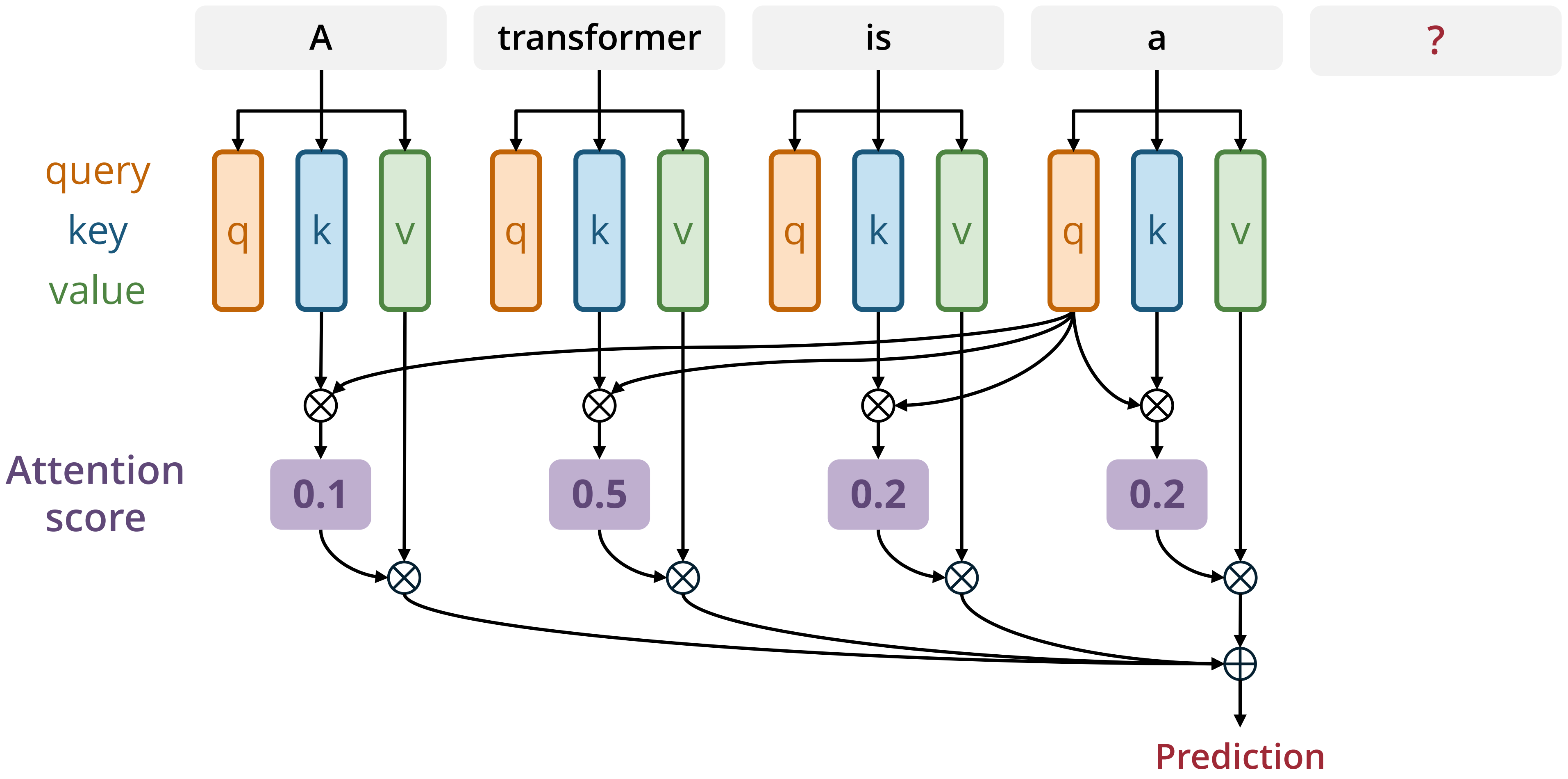
(Recap) Demystifying Transformers



(Recap) Demystifying Transformers

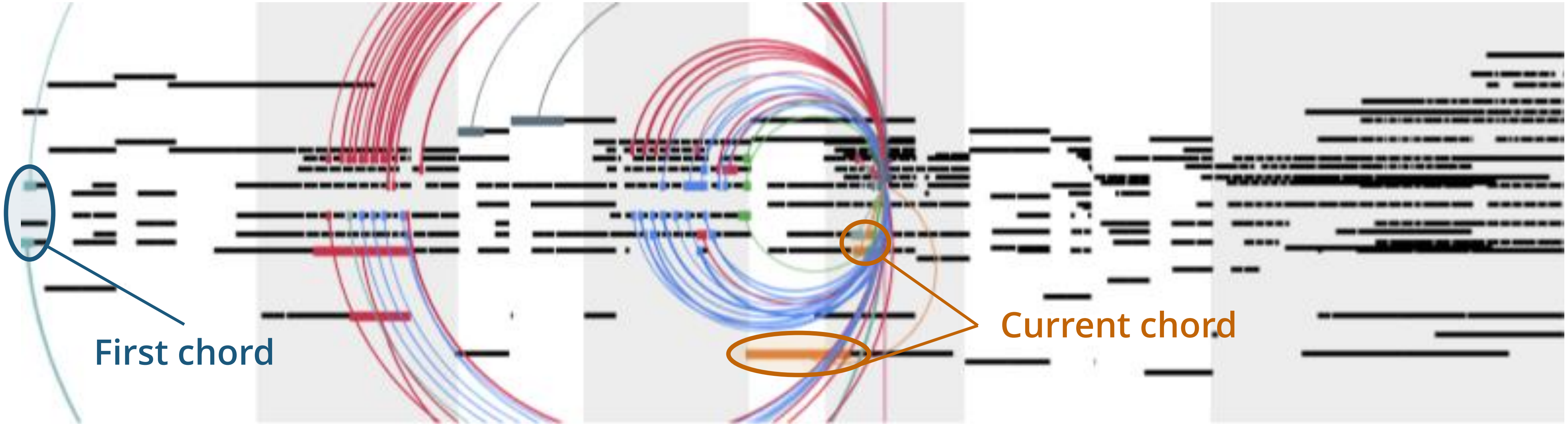


(Recap) Demystifying Transformers



Visualizing Musical Self-attention

(Each color represents an attention head)

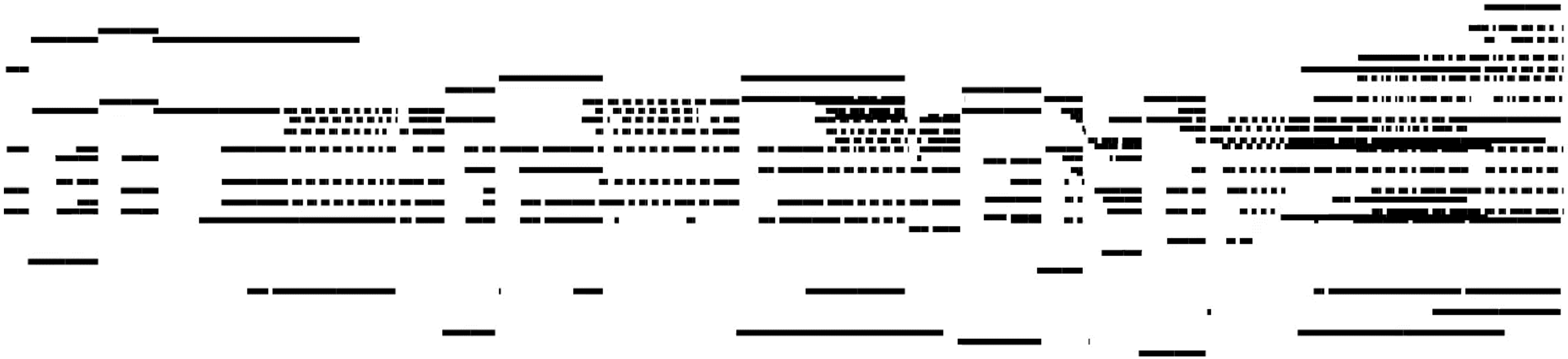


(Source: Huang et al., 2018)

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," *Magenta Blog*, December 13, 2018.

Visualizing Musical Self-attention

(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," *Magenta Blog*, December 13, 2018.

Beyond Solo Music

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

Example: 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

Example: Multitrack Music Machine (Ens & Pasquier, 2020)

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

youtu.be/NdeMZ3y-84Q

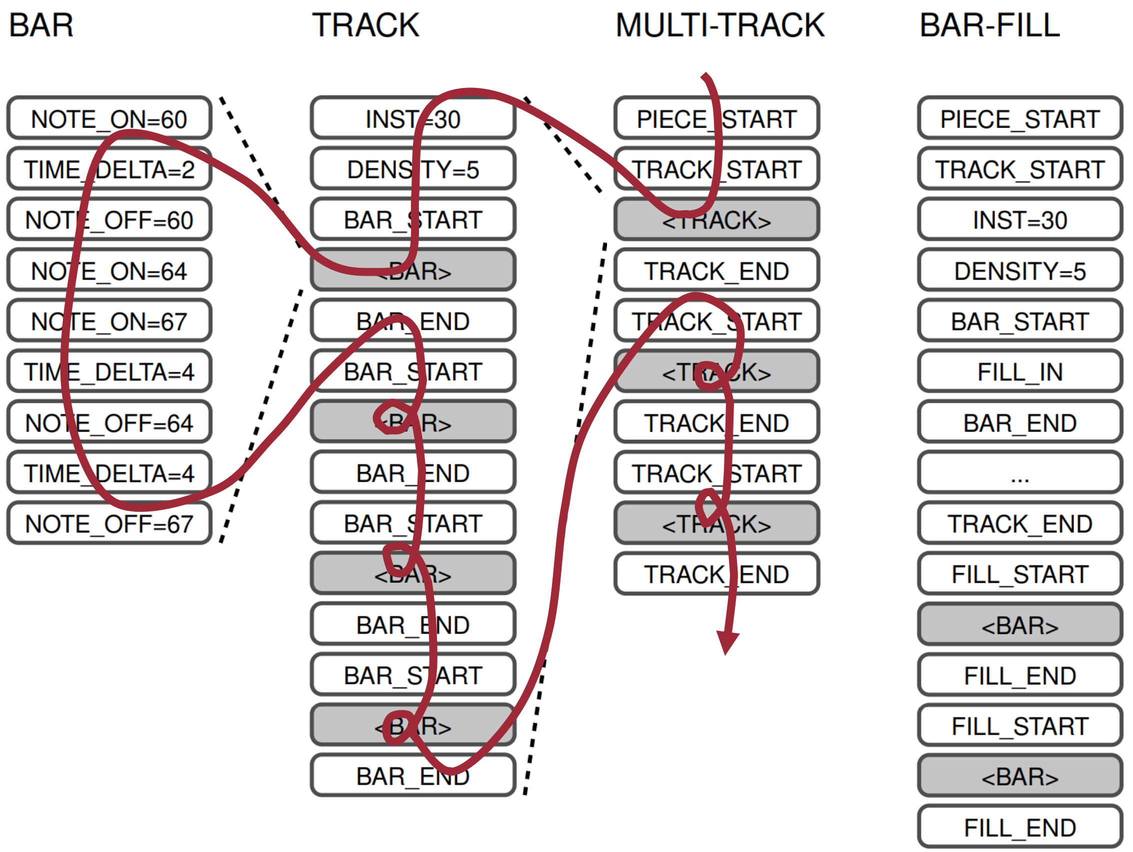


Fig. 1. The MultiTrack and BarFill representations are shown. The <bar> tokens correspond to complete bars, and the <track> tokens correspond to complete tracks.

(Ens & Pasquier, 2020)

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

Example: Multitrack Music Transformer (Dong et al., 2023)

- **Data**: Symbolic Orchestral Database (SOD)
- **Representation**: "(**beat**, **position**, **pitch**, **duration**, **instrument**)"
  - No time shift events **Why?**
- **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
```

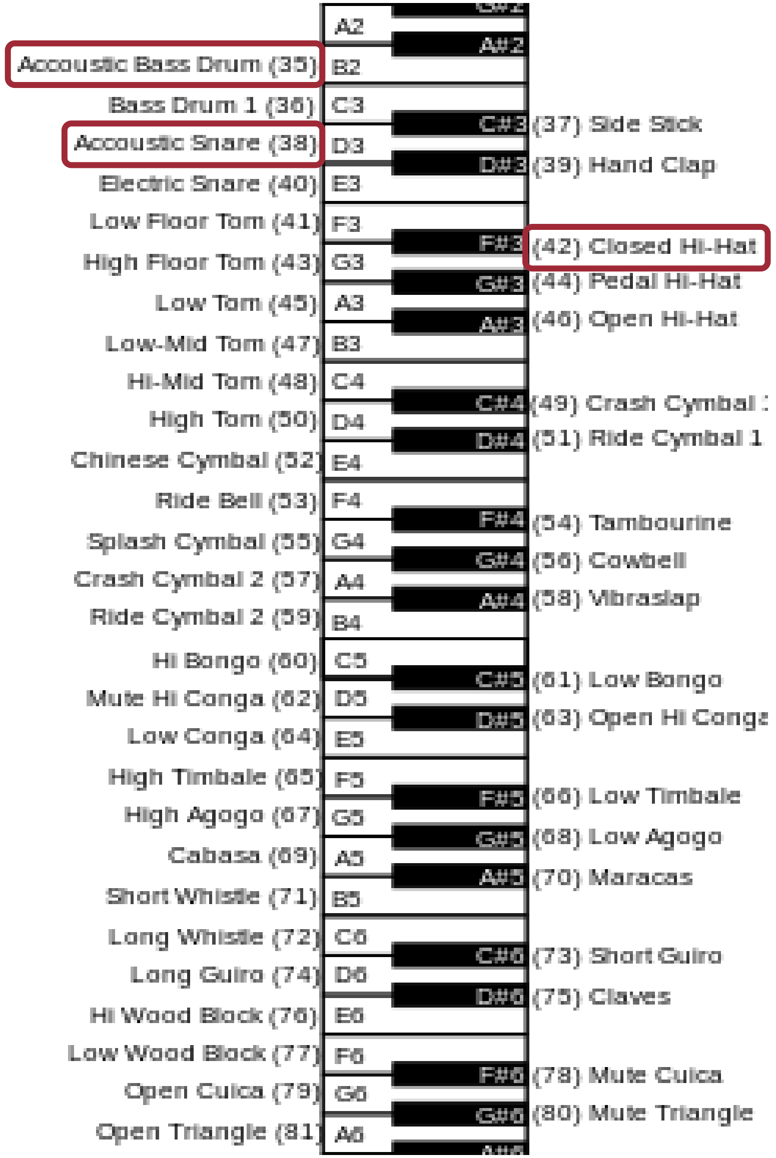
(Source: Dong et al., 2023)

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.

Decoding Strategies

(Recap) Language Models (Mathematically)

- A class of machine learning models that learn the next word probability

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

Next word

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 $)$ ⬇

(Recap) Language Models – Generation

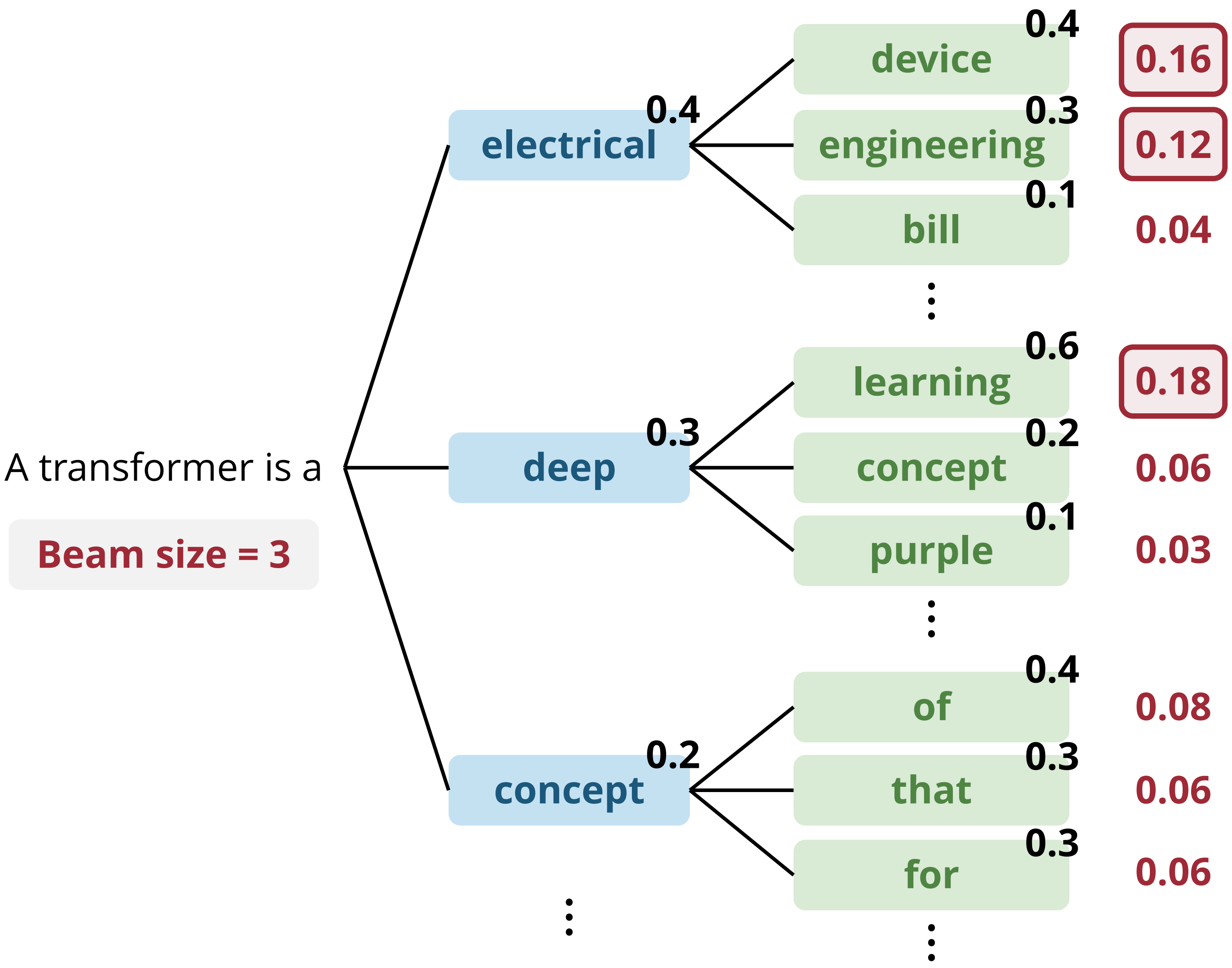
- How do we generate a new sentence using a trained language model?

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

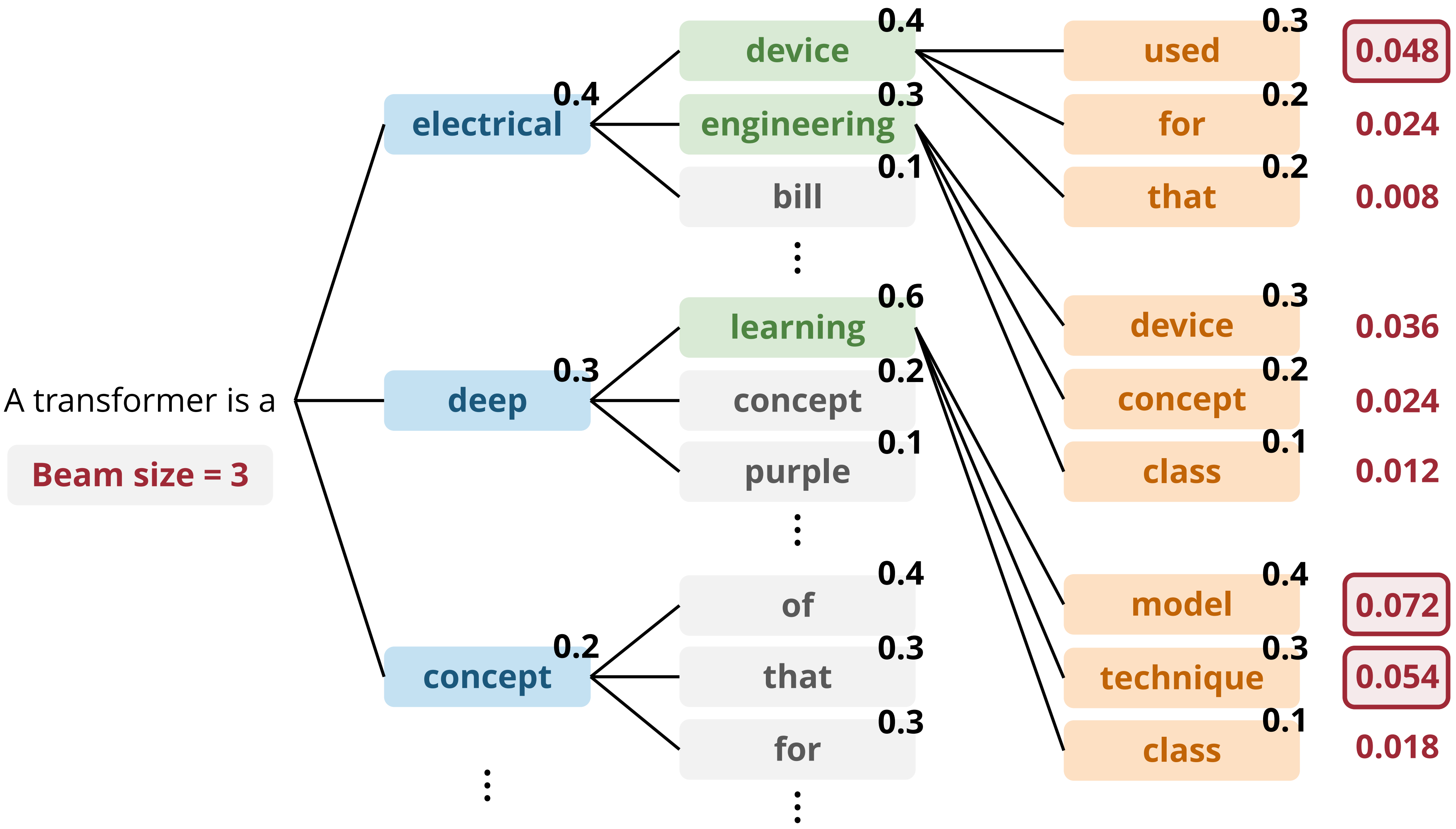
Decoding Strategies

- **Search-based**
  - **Greedy search**: Pick **the most probable word** at each step
  - **Beam search**: Pick **the most probably word sequences** jointly considering the top-K probable sequences so far
- **Sampling-based**
  - **Random sampling**: Sample randomly from the **raw output distribution**
  - **Top-K sampling**: Sample randomly from the output distribution considering **only the top-K probable words**
  - **Top-p sampling**: Sample randomly from the output distribution considering **only the top several probable words that make up the top p% of the distribution**

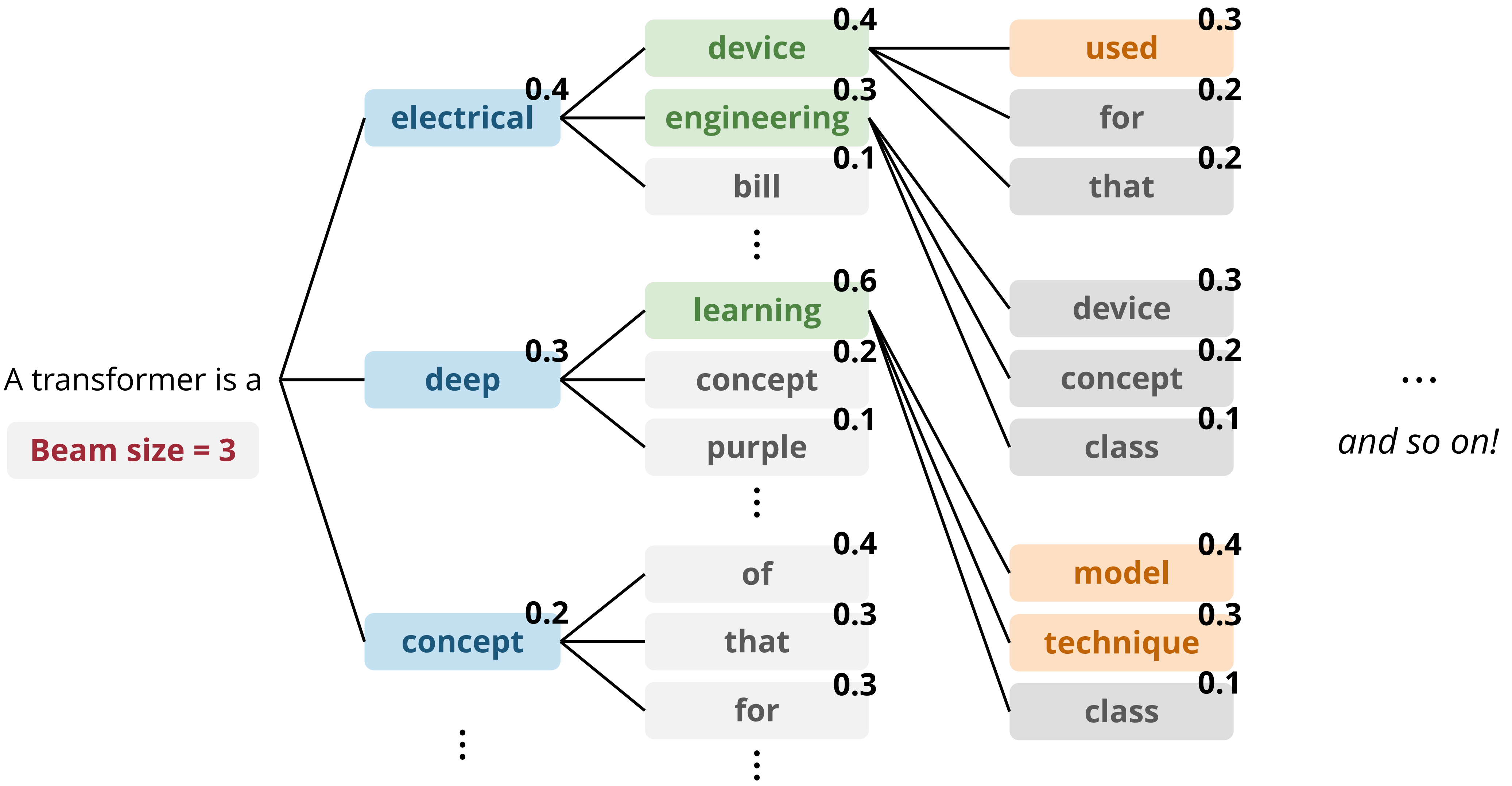
Decoding Strategies – Beam Search



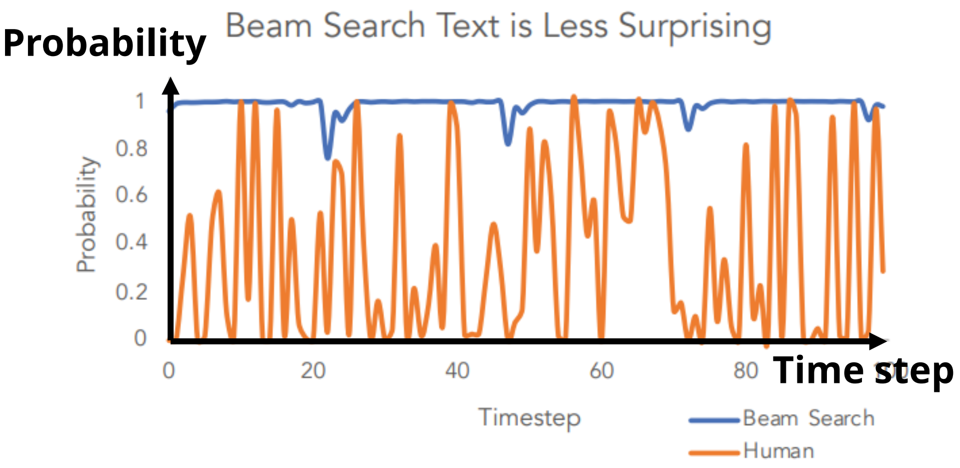
Decoding Strategies – Beam Search



Decoding Strategies – Beam Search



Is the Most Probably Sequence What We Want?



(Source: Holtzman et a., 2020)

Beam Search

...to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and...

Human

...which grant increased life span and three years warranty. The Antec HCG series consists of five models with capacities spanning from 400W to 900W. Here we should note that we have already tested the HCG-620 in a previous review and were quite satisfied With its performance. In today's review we will rigorously test the Antec HCG-520, which as its model number implies, has 520W capacity and contrary to Antec's strong beliefs in multi-rail PSUs is equipped...

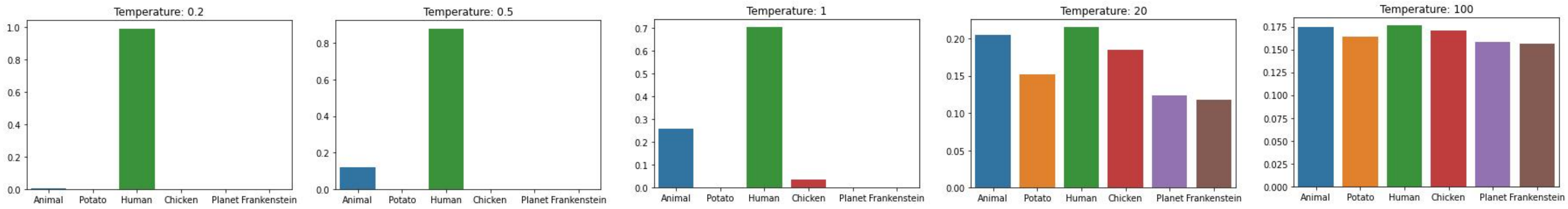
Decoding Strategies – Temperature

Softmax $$\hat{y}_i = \frac{e^{\tilde{y}_i}}{\sum_{j=1}^{n} e^{\tilde{y}_j}} \longrightarrow \hat{y}_i = \frac{e^{\tilde{y}_i/\tau}}{\sum_{j=1}^{n} e^{\tilde{y}_j/\tau}}$$

Temperature

Original

$\tau = 0.2$ $\tau = 0.5$ $\tau = 1$ $\tau = 20$ $\tau = 100$



(Source: Mehta et a., 2020)

Low temperature ⟷ **High temperature**

Temperature adjusts the "contrast" of the distribution!

Shivam Mehta, "The need for sampling temperature and differences between whisper, GPT-3, and probabilistic model's temperature," *Shivam Mehta's Blog*, 2023

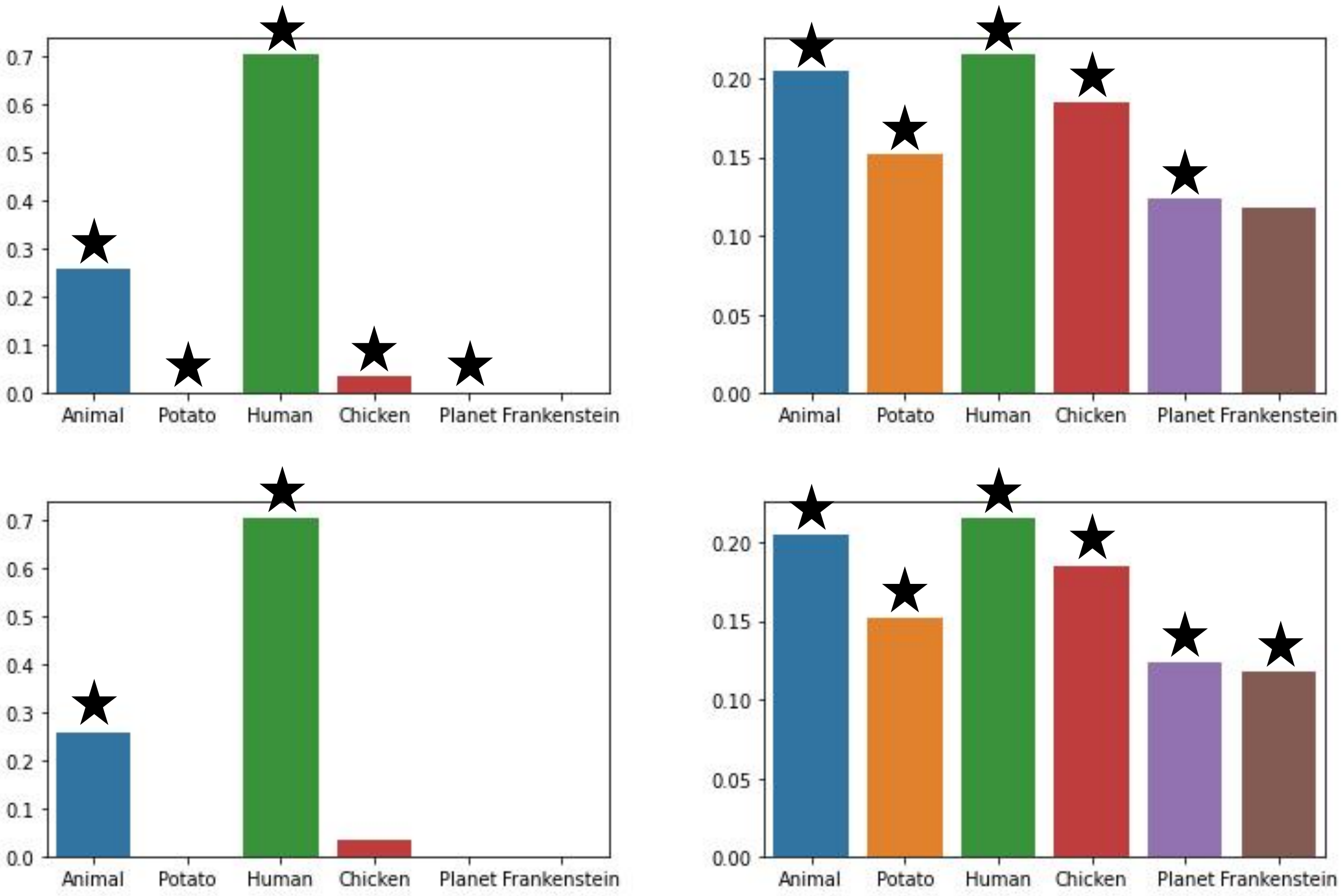
Decoding Strategies – Top-K vs Top-p Sampling

Top-K sampling

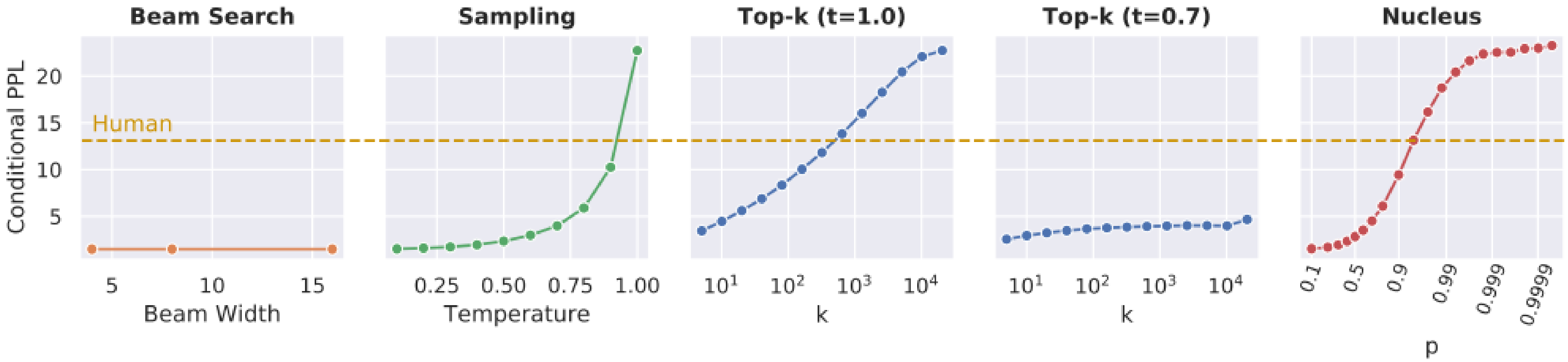
$K = 5$

Top-p sampling

$p = 0.95$



Striking a Balance between Coherence & Interestingness



(Source: Holtzman et a., 2020)

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi, "The Curious Case of Neural Text Degeneration," *ICLR*, 2020.

(Recap) Decoding Strategies

Search-based	**Greedy search**	Pick **the most probable word** at each step
	Beam search	Pick **the most probably word sequences** jointly considering the top-K probable sequences so far
Sampling-based	**Random sampling**	Sample randomly from the **raw output distribution**
	Top-K sampling	Sample randomly from the output distribution considering **only the top-K probable words**
	Top-p sampling	Sample randomly from the output distribution considering **only the top several probable words that make up the top p% of the distribution**