PAT 464/564 (Winter 2026)

Generative AI for Music & Audio Creation

Lecture 9: RNNs & LSTMs

Instructor: Hao-Wen Dong

Representative Types of Deep Generative Models

- **Deep autoregressive models**
  - Recurrent neural network (RNN)
  - Long short-term memory (LSTM)

  Today's topic!

  - Transformer model

- **Deep latent variable models**
  - Generative adversarial network (GAN)
  - Variational autoencoder (VAE)
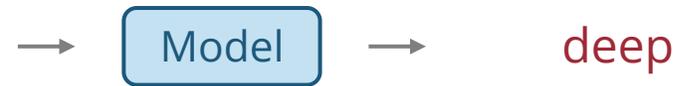 - Diffusion model
  - Flow-based model

- *And many others...*

Deep Autoregressive Models

3

Deep Autoregressive Models

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

A transformer is a $\longrightarrow$ [Model] $\longrightarrow$ deep

A transformer is a deep $\longrightarrow$ [Model] $\longrightarrow$ learning

A transformer is a deep learning $\longrightarrow$ [Model] $\longrightarrow$ model

A transformer is a deep learning model $\longrightarrow$ [Model] $\longrightarrow$ introduced

A transformer is a deep learning model introduced $\longrightarrow$ [Model] $\longrightarrow$ in

A transformer is a deep learning model introduced in $\longrightarrow$ [Model] $\longrightarrow$ 2017

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

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

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

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

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

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

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

Deep Autoregressive Models

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

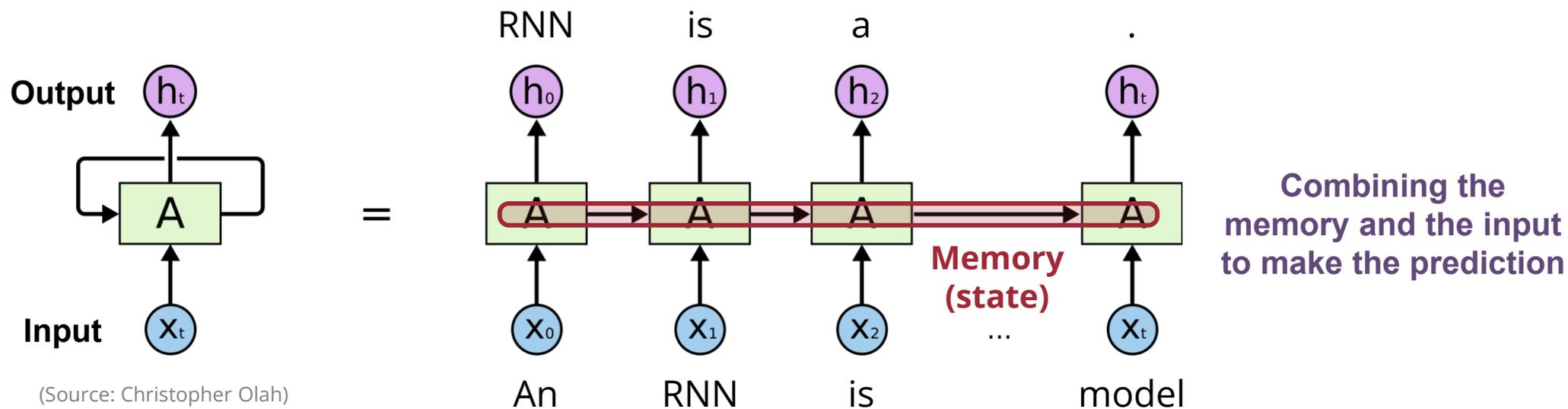$$P(\; x_i \mid x_1, x_2, \ldots, x_{i-1}\;)$$

Next word   Previous words

The whole sentence
$X = (x_0, x_1, \ldots, x_N)$

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

6

Deep Autoregressive Models

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

What we want the model to learn!

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

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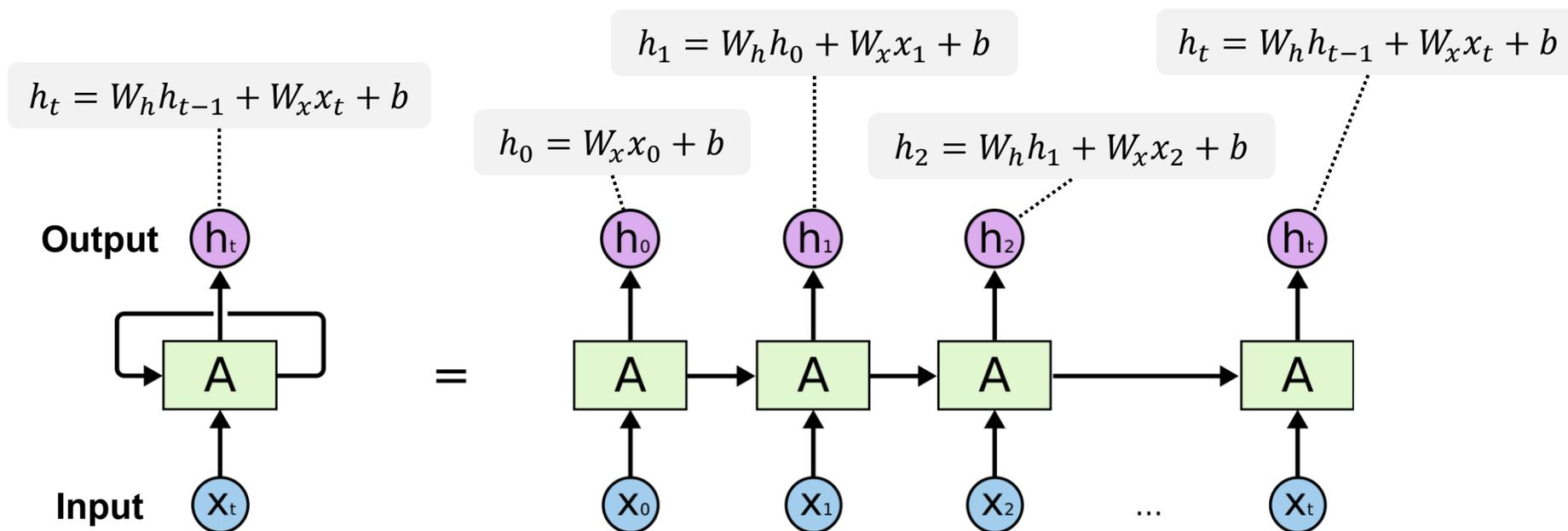
Recurrent Neural Networks (RNNs)

What is an RNN (Recurrent Neural Network)?

- A type of neural networks that have **loops**

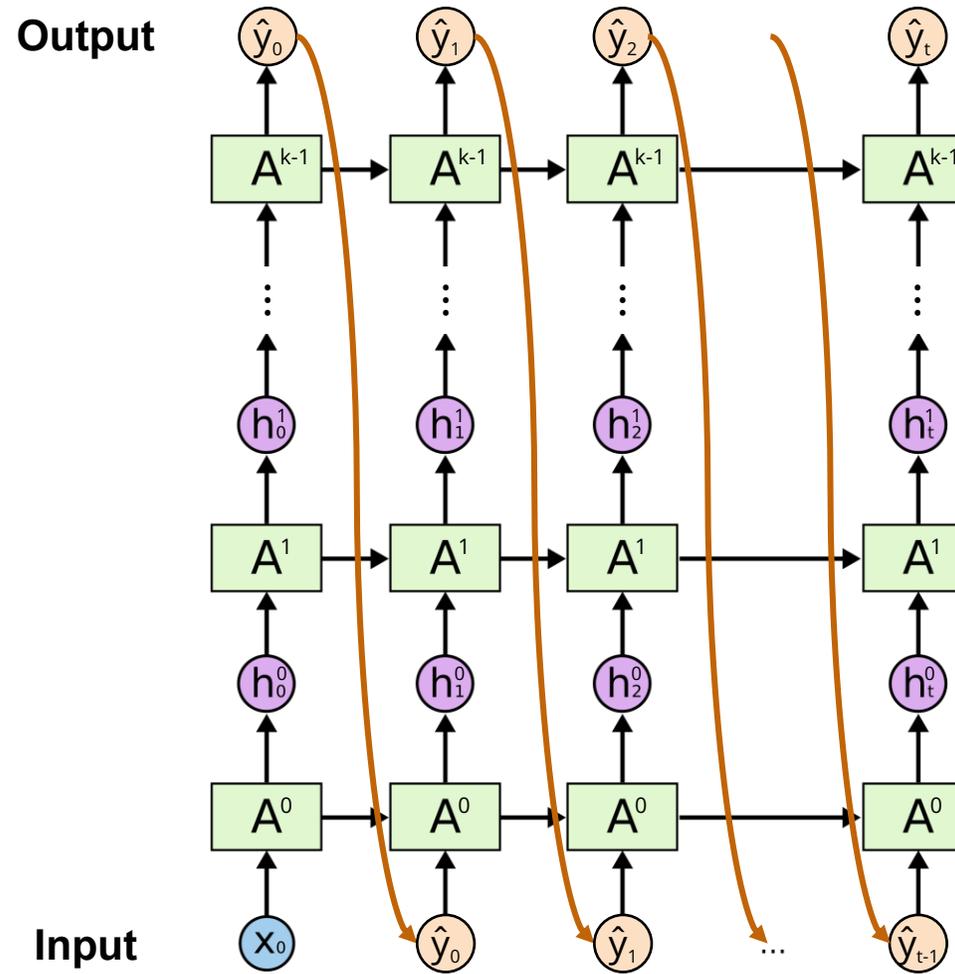- Widely used for **modeling sequences** (e.g., in natural language processing)

(Source: Christopher Olah)

Vanilla RNNs

- The simplest form of RNNs

$$h_t = W_h h_{t-1} + W_x x_t + b$$

$$h_1 = W_h h_0 + W_x x_1 + b$$

$$h_t = W_h h_{t-1} + W_x x_t + b$$

$$h_0 = W_x x_0 + b$$

$$h_2 = W_h h_1 + W_x x_2 + b$$

Output h_t

h_0 h_1 h_2 h_t

A   =   A → A → A → A

Input x_t

x_0 x_1 x_2 ... x_t

(Source: Christopher Olah)
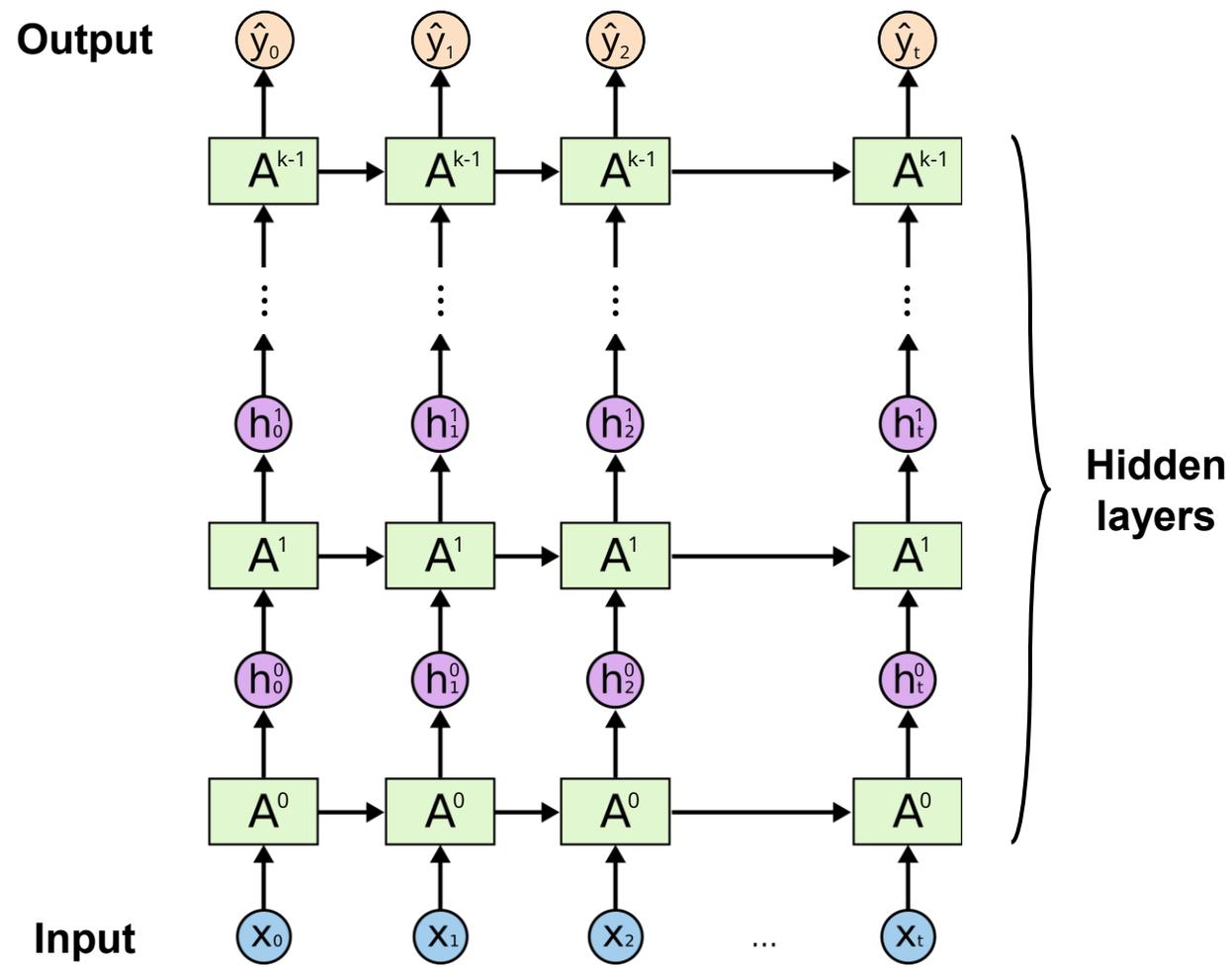
Vanilla RNNs

(Source: Christopher Olah)

Vanilla RNNs: Inference

(Source: Christopher Olah)

12

Vanilla RNNs: Training
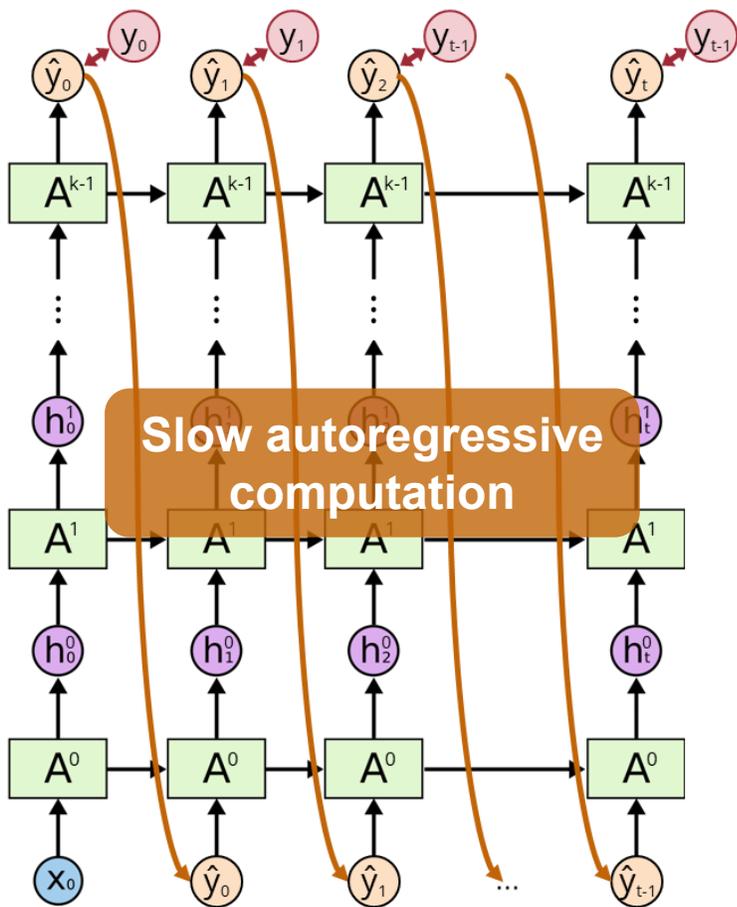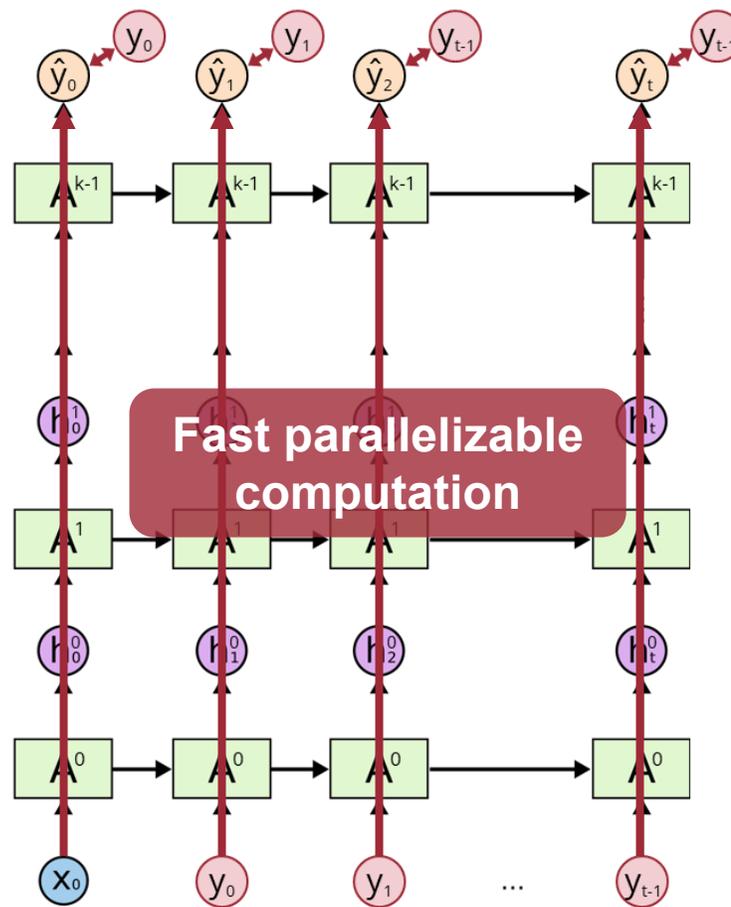

(Source: Christopher Olah)

13

Hao-Wen Dong, Generative AI for Music and Audio Creation (PAT 464/564), University of Michigan
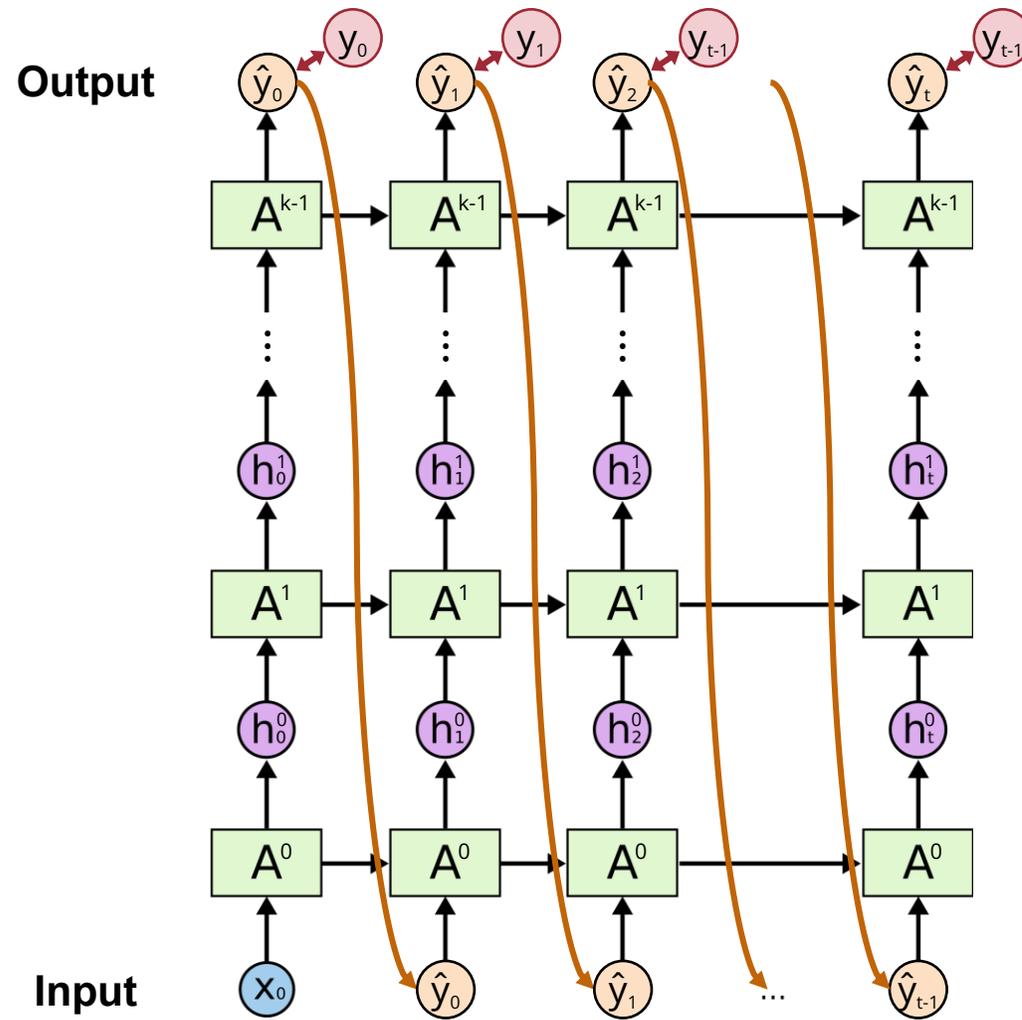
Vanilla RNNs: Training with Teacher Forcing

(Source: Christopher Olah)

Use ground truth instead
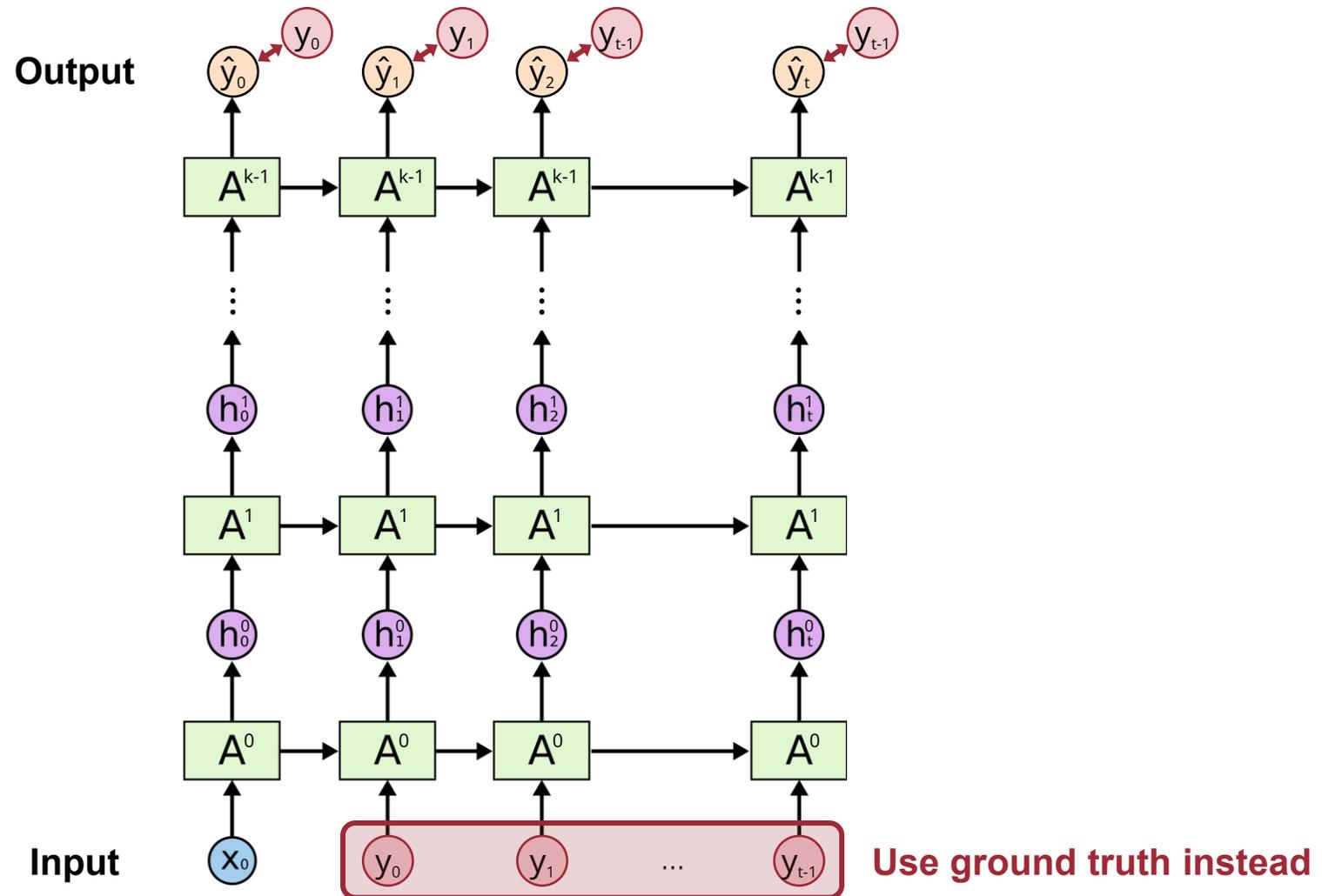
Vanilla RNNs: Training with vs. without Teacher Forcing

(Source: Christopher Olah)

(Source: Christopher Olah)

Vanilla RNNs: Training vs. Inference

Training

Inference

Fast parallelizable training

Slow autoregressive inference
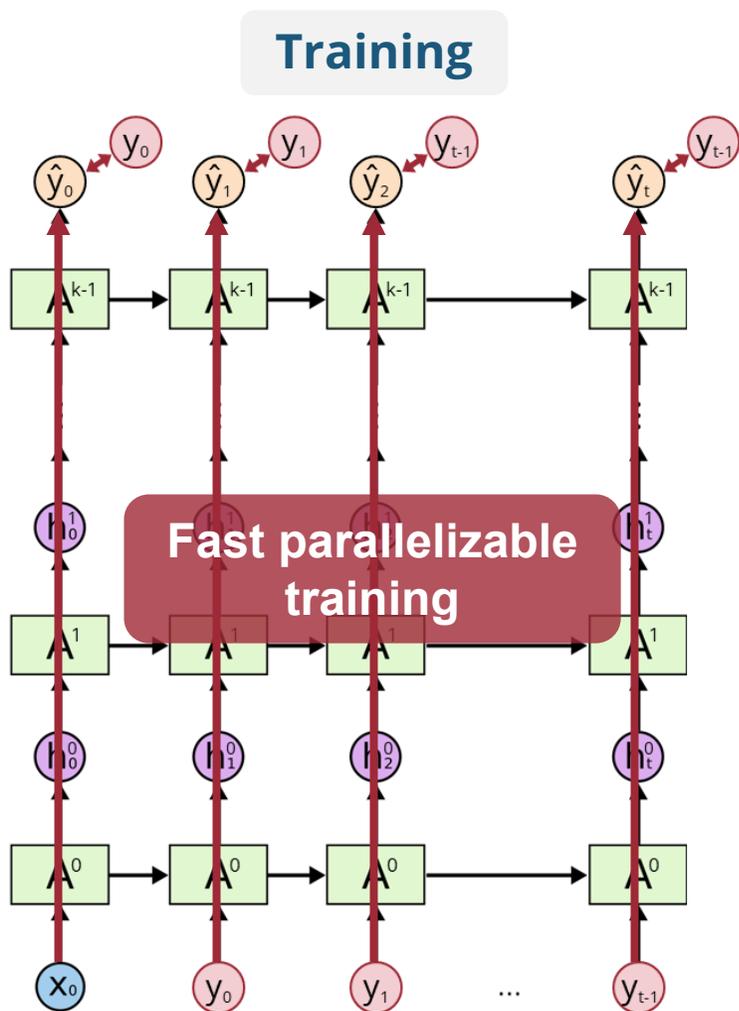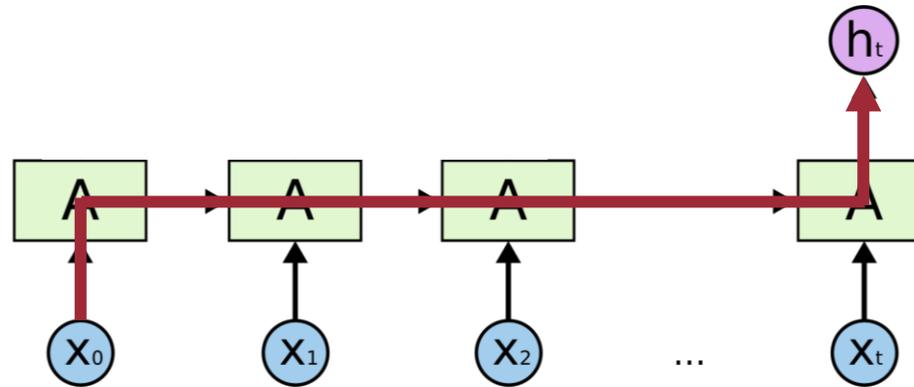
(Source: Christopher Olah)
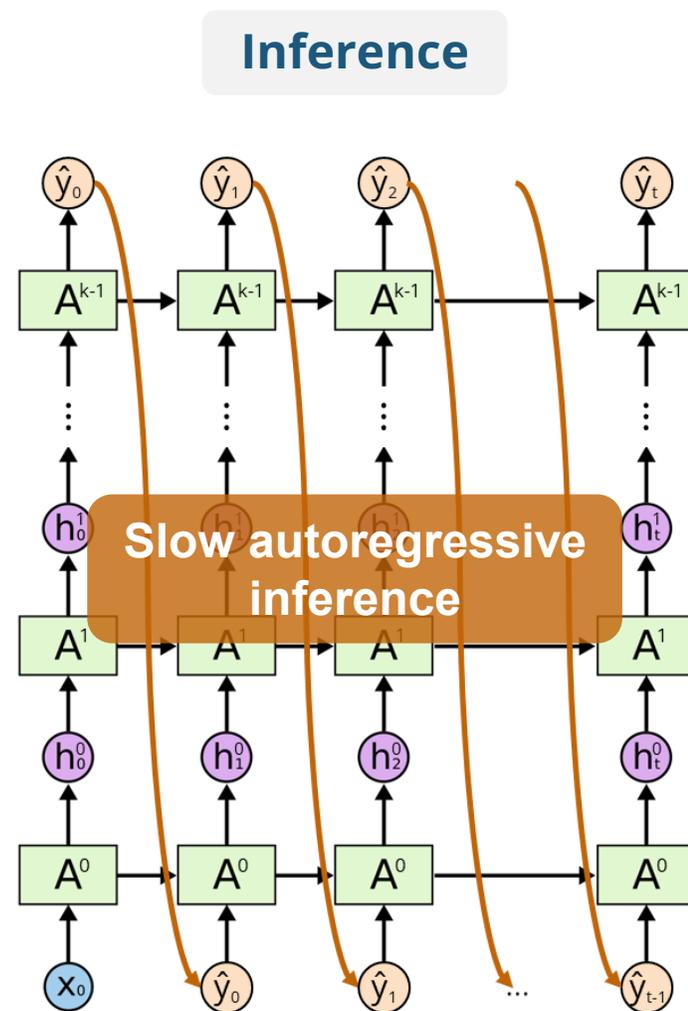
(Source: Christopher Olah)

Long Short-Term Memory (LSTMs)

Backpropagation Through Time

- An RNN is essentially a **very deep neural network**

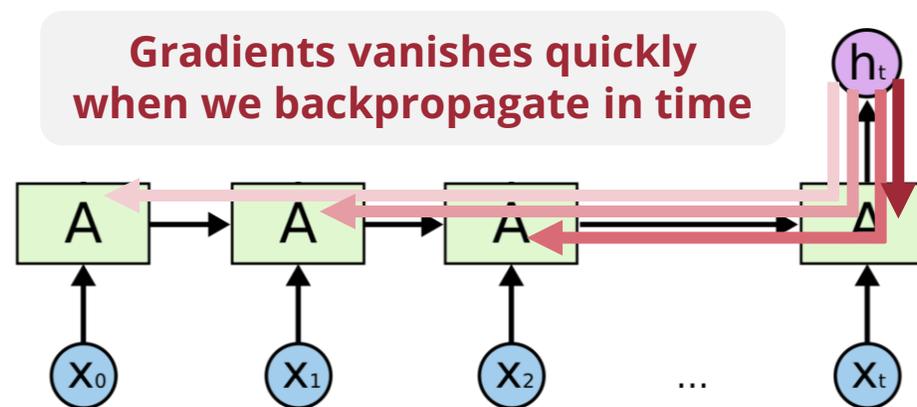(Source: Christopher Olah)

$$h_t = W_h h_{t-1} + W_x x_t + b$$

$$h_t = W_h(W_h h_{t-2} + W_x x_{t-1} + b) + W_x x_t + + b$$

$$\vdots$$

$$h_t = W_h(W_x x_{t-1} + W_h(\cdots W_h h_0 + W_x x_1 + b \cdots) + b) + W_x x_t + + b$$

18

Vanishing Gradients

- An RNN is essentially a **very deep neural network**

Gradients vanishes quickly when we backpropagate in time

(Source: Christopher Olah)

All the layers share the same weight matrix

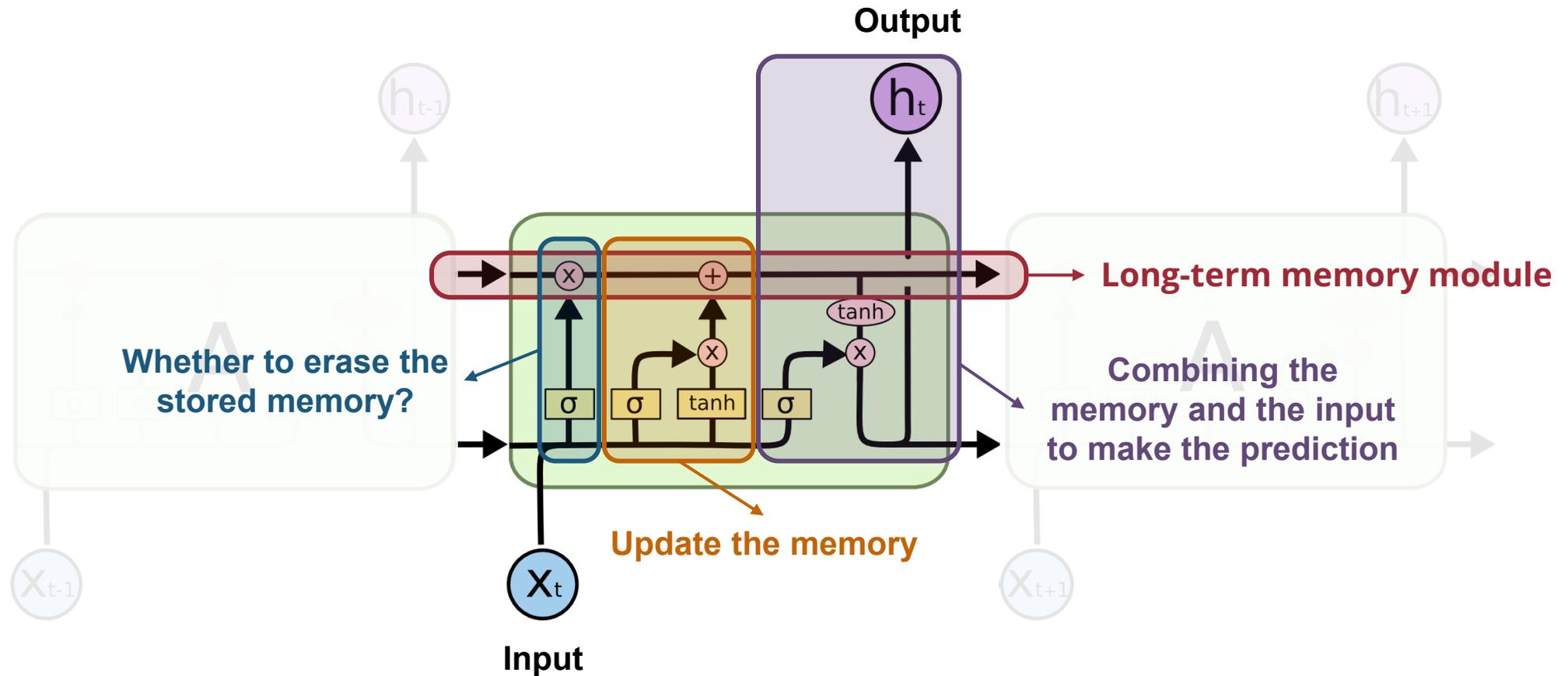**Can still train the model without deeper gradients**

Why bother?

$$h_t = W_h h_{t-1} + W_x x_t + b$$

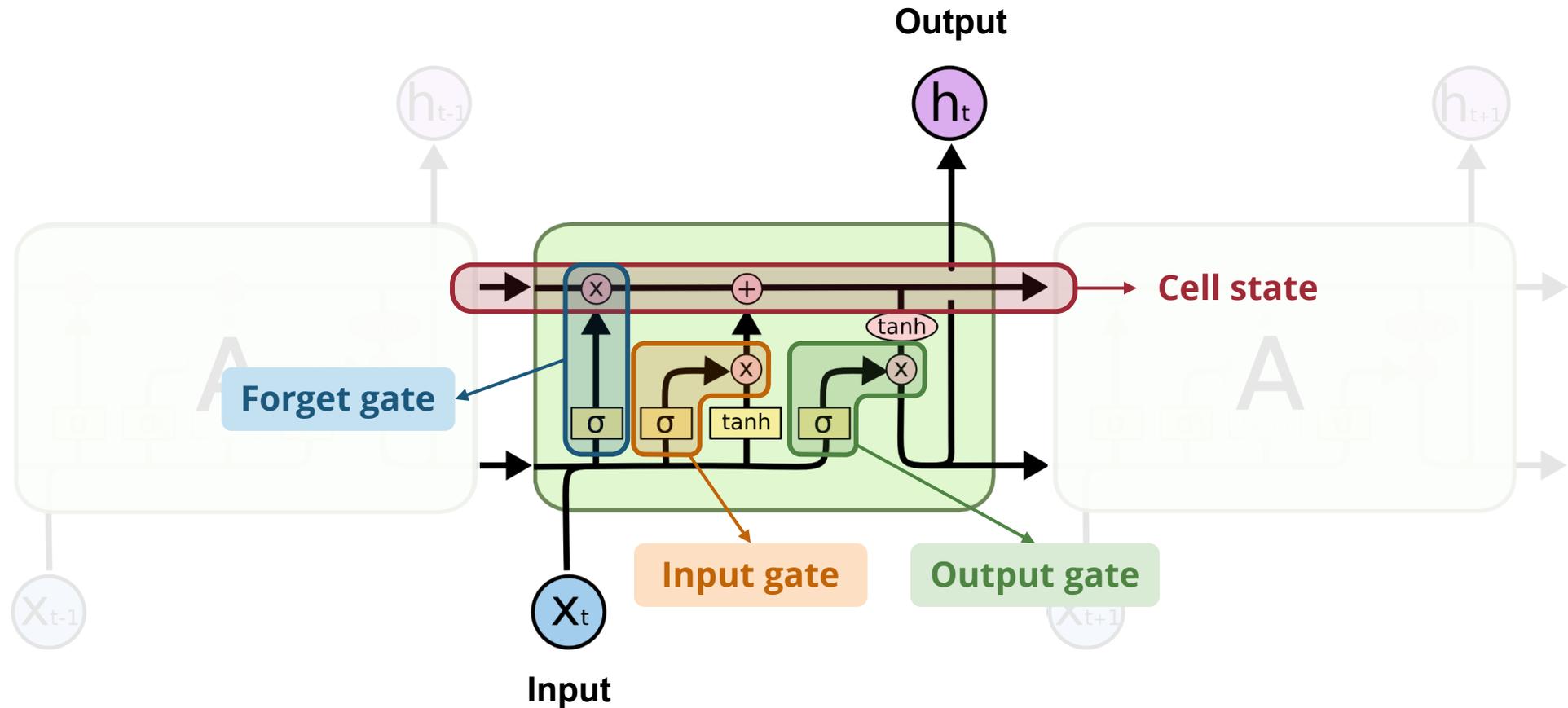$$h_t = W_h(W_h h_{t-2} + W_x x_{t-1} + b) + W_x x_t + +b$$

$$\vdots$$

$$h_t = W_h(W_x x_{t-1} + W_h(\cdots W_h h_0 + W_x x_1 + b \cdots) + b) + W_x x_t + +b$$

Demystifying LSTMs (Hochreiter & Schmidhuber, 1997)

Output

h_t

Long-term memory module

Whether to erase the stored memory?

Combining the memory and the input to make the prediction
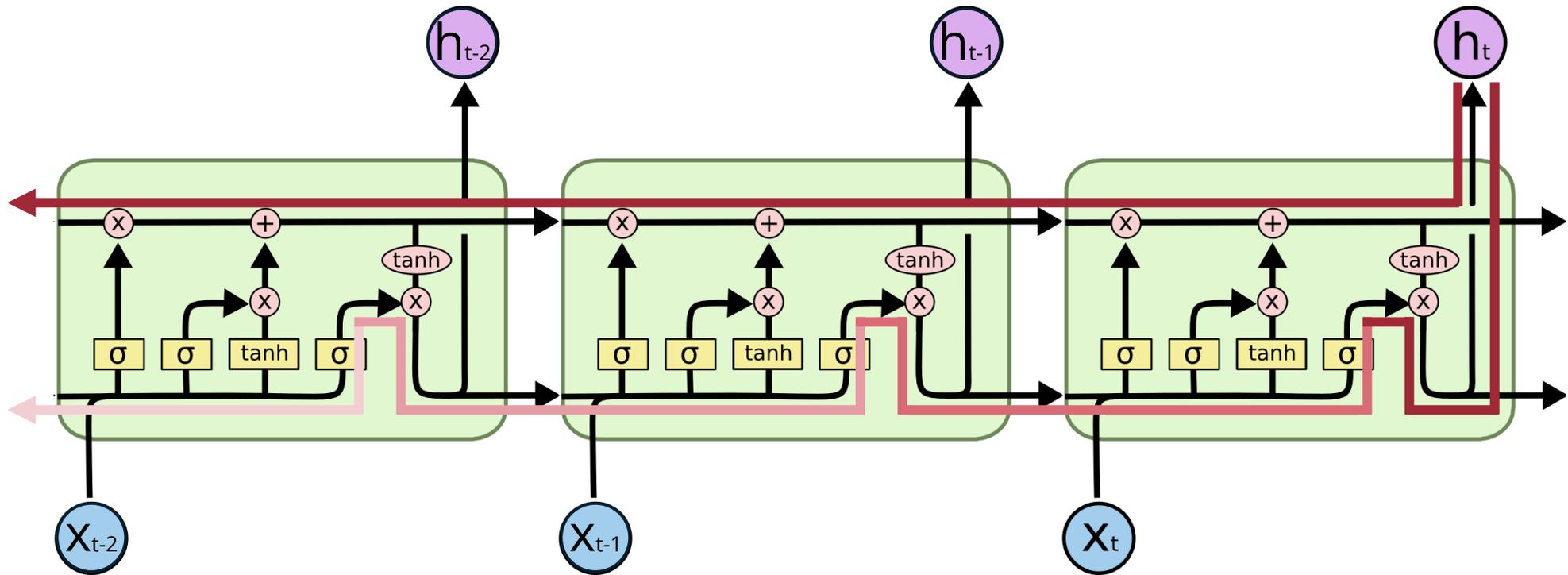
Update the memory

x_t

Input

(Source: Christopher Olah)

20

Demystifying LSTMs (Hochreiter & Schmidhuber, 1997)

(Source: Christopher Olah)
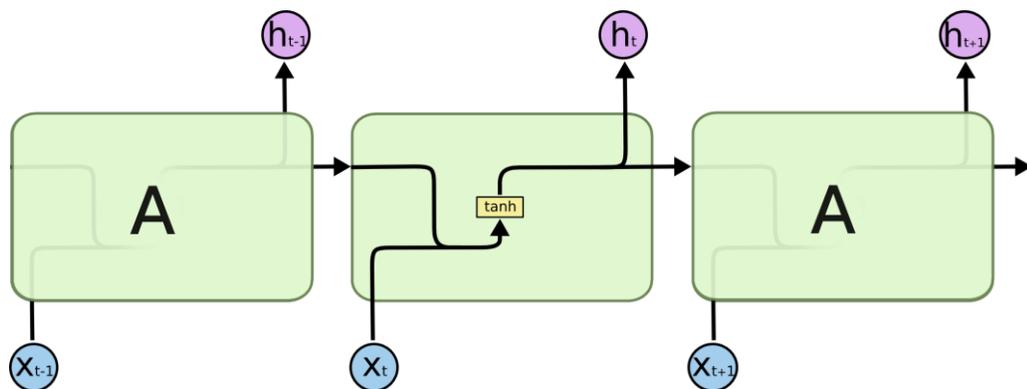
How can LSTMs Help Alleviate Vanishing Gradients?

(Source: Christopher Olah)

LSTMs does not completely solve vanishing gradients
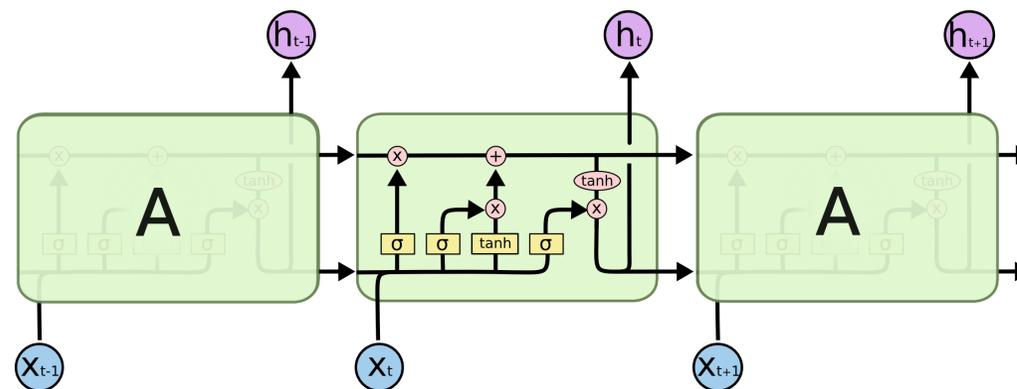
Vanilla RNNs vs. LSTMs

Vanilla RNN

- Simplest form of RNNs

- Limited long-term memory

- Harder to train (due to gradient vanishing)

(Source: Christopher Olah)
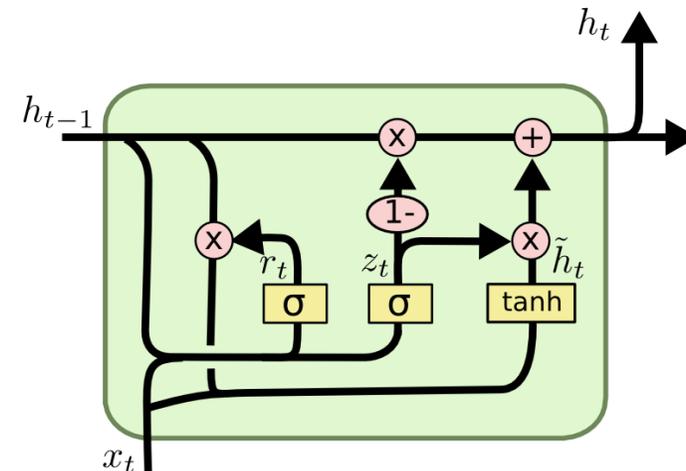
LSTM

- Improved memory module

- Better long-term memory

- Easier to train

(Source: Christopher Olah)

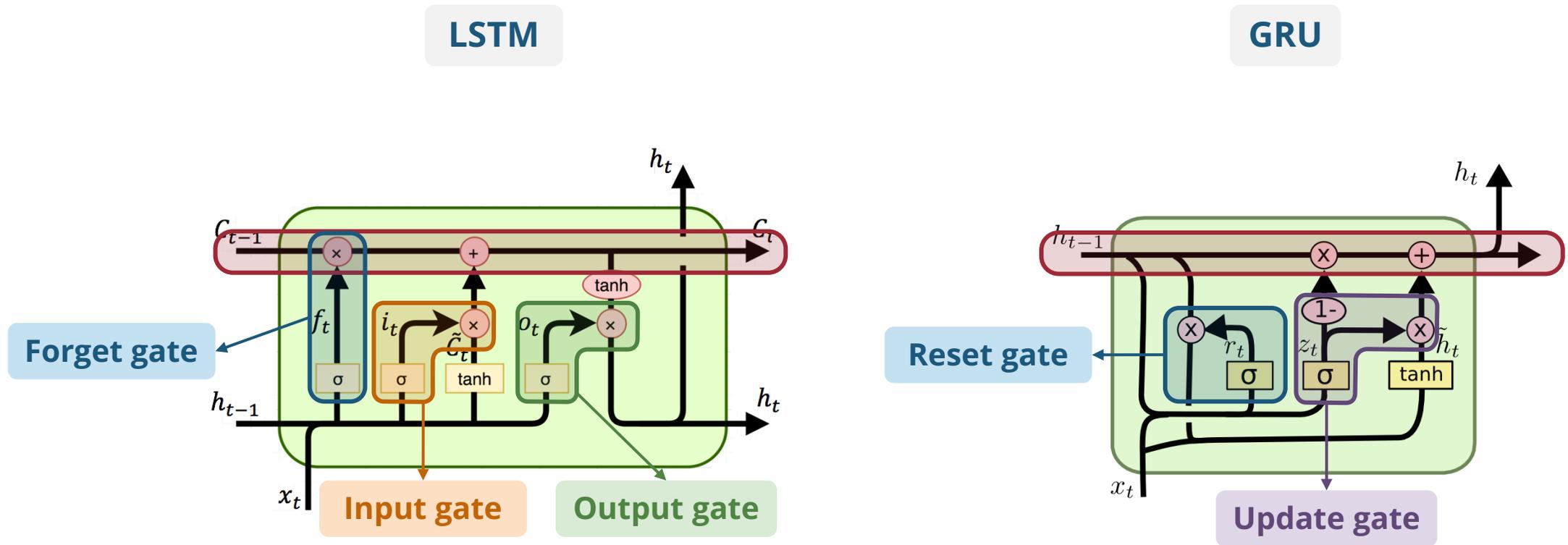colah.github.io/posts/2015-08-Understanding-LSTMs

Gated Recurrent Units (GRUs)

- A **simplified** version of LSTM

- An LSTM consists of
  - **Forget** gate
  - **Input** gate
  - **Output** gate

- An GRU consists of
  - **Reset** gate
  - **Update** gate

(Source: dprogrammer.org)

LSTMs vs. GRUs

(Source: dprogrammer.org)

Generating Music like Languages

Large Language Models (LLMs)

- The models behind ChatGPT!

Word-by-word generation

Language Models

- Predicting the next word given the past sequence of words

A transformer is a _____

electrical device ⬆

fiction character ⬆

deep learning model ⬆

family of genes ⬆

type of food ⬇

musical instrument ⬇

Language Models (Mathematically)

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

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

Next word   Previous words

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

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

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

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

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

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

29

Music Language Models (Mathematically)

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

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

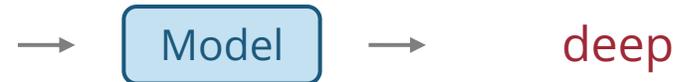Next note   Previous notes

$$P(\ G\ |\ C\ C\ G\ G\ A\ A\ )$$

$$P(\ A\ |\ C\ C\ G\ G\ A\ A\ )$$

$$P(\ C\ |\ C\ C\ G\ G\ A\ A\ )$$

$$P(\ F\ |\ C\ C\ G\ G\ A\ A\ )$$

$$P(\ Ab\ |\ C\ C\ G\ G\ A\ A\ )$$

$$P(\ A\sharp\ |\ C\ C\ G\ G\ A\ A\ )$$

30

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

Some randomness involved!

Designing a Machine-readable Music Language

- How can we "represent" music in a way that machines understand?

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'

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:|
```

Folk RNN (Sturm et al., 2015)

- **Data**
  - 23,958 folk tunes

- **Representation**
  - ABC notation without metadata

- **Model**
  - LSTM (long short-term memory)
  - Working on the character level

folkRNN

generate a folk tune with a recurrent neural network

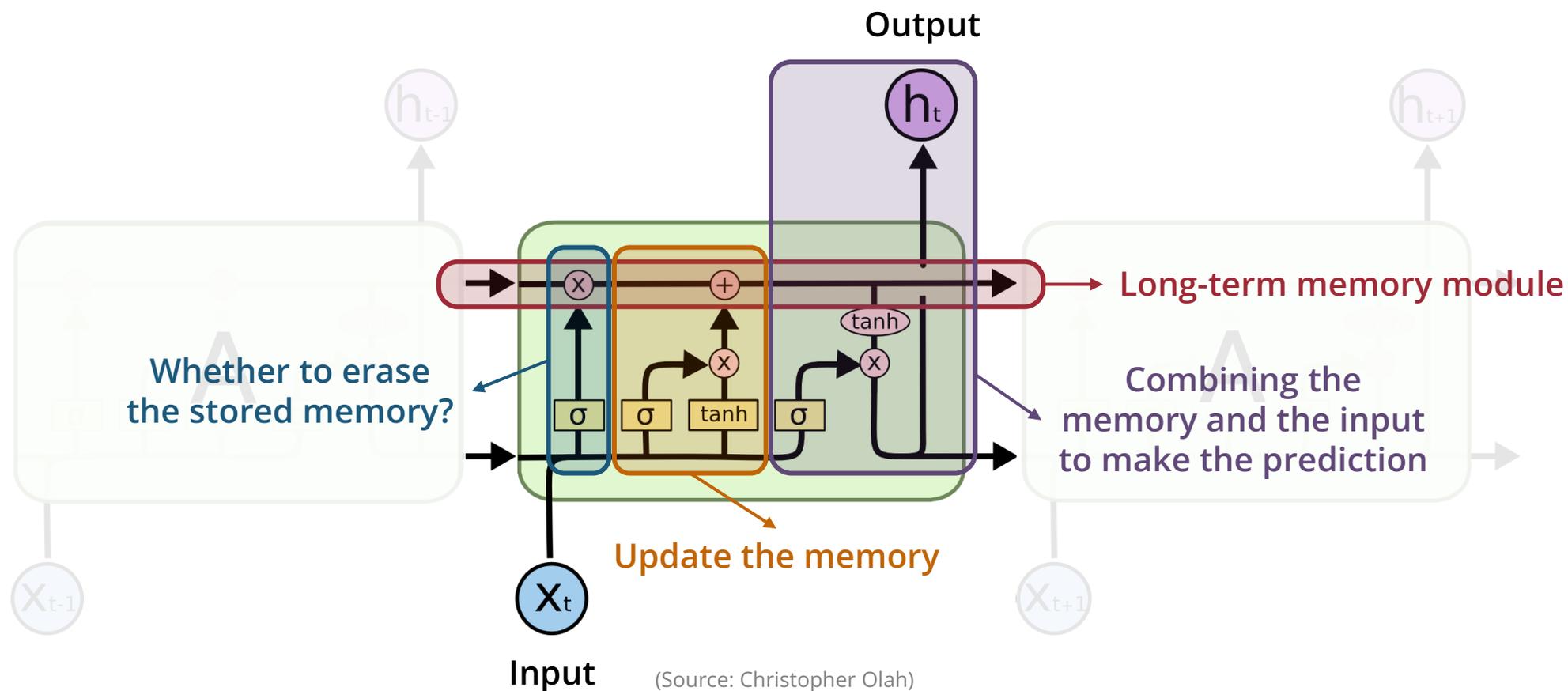PRESS TO GENERATE TUNE

Compose

MODEL

thesession.org (w/ :| |:)

TEMPERATURE   SEED

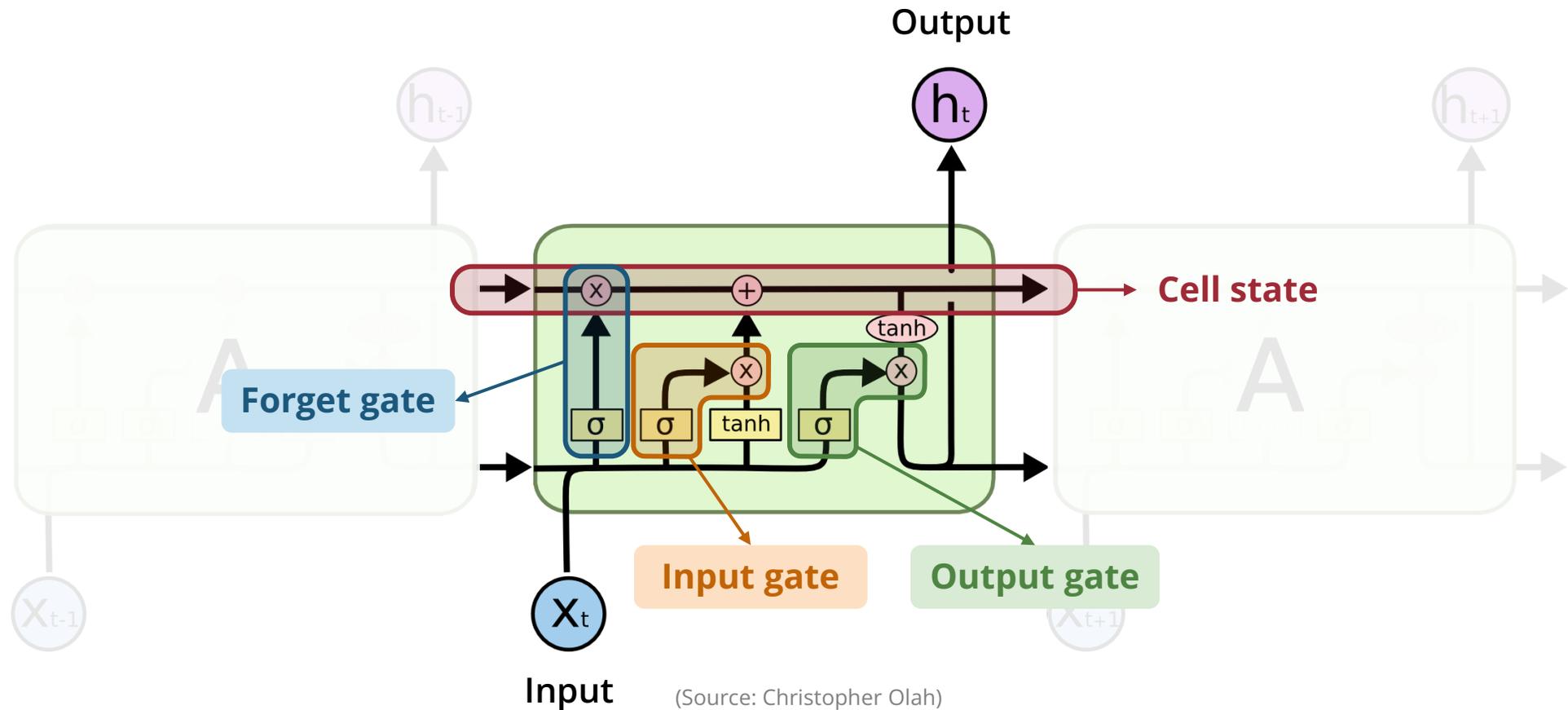1   62063

METER   MODE

4/4   C Major

INITIAL ABC

Enter start of tune in ABC notation

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.

36

Demystifying LSTMs (Hochreiter & Schmidhuber, 1997)

Output

h_t

Long-term memory module

Whether to erase the stored memory?

Update the memory

Combining the memory and the input to make the prediction

x_t

Input

(Source: Christopher Olah)

37

Demystifying LSTMs (Hochreiter & Schmidhuber, 1997)

(Source: Christopher Olah)

colah.github.io/posts/2015-08-Understanding-LSTMs
Sepp Hochreiter and Jürgen Schmidhuber, "Long Short-Term Memory," *Neural Computation*, 9(8):1735-1780, 1997.

38

Folk RNN (Sturm et al., 2015)

- **Data**
  - 23,958 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
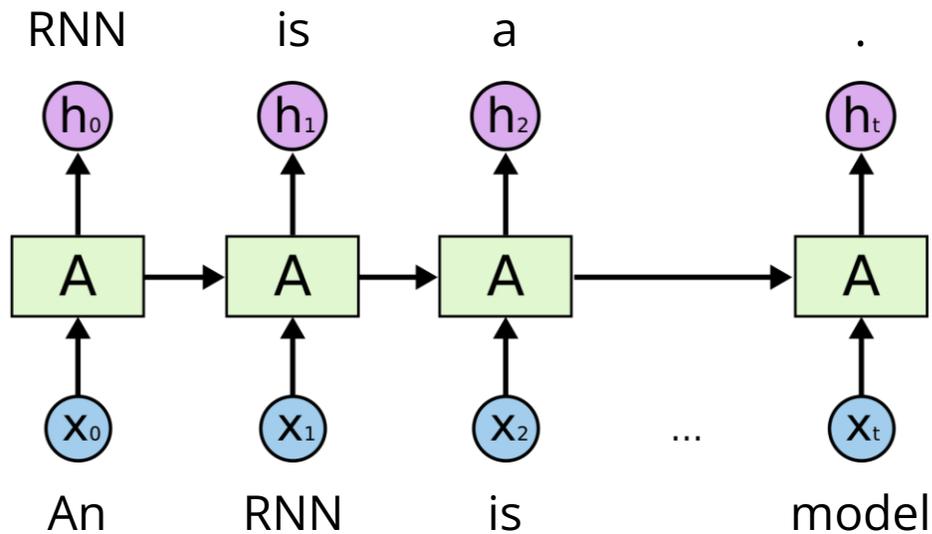
Character-level RNNs

- Predicting character by character
- Useful when there is no natural "spaces"

(Source: Christopher Olah)
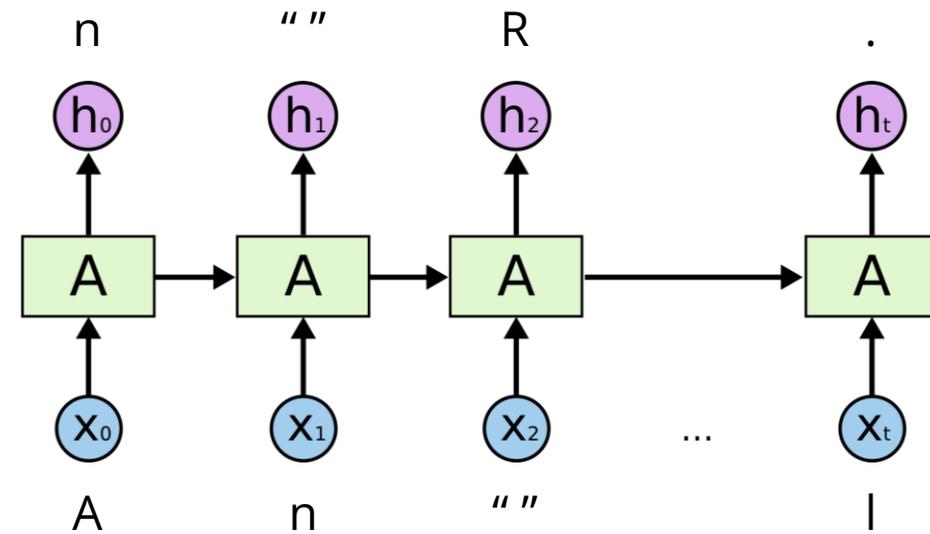
(Source: Christopher Olah)

Limitations of ABC Notations

- Limited expressiveness

- Monophonic tunes only

MIDI-like Representation
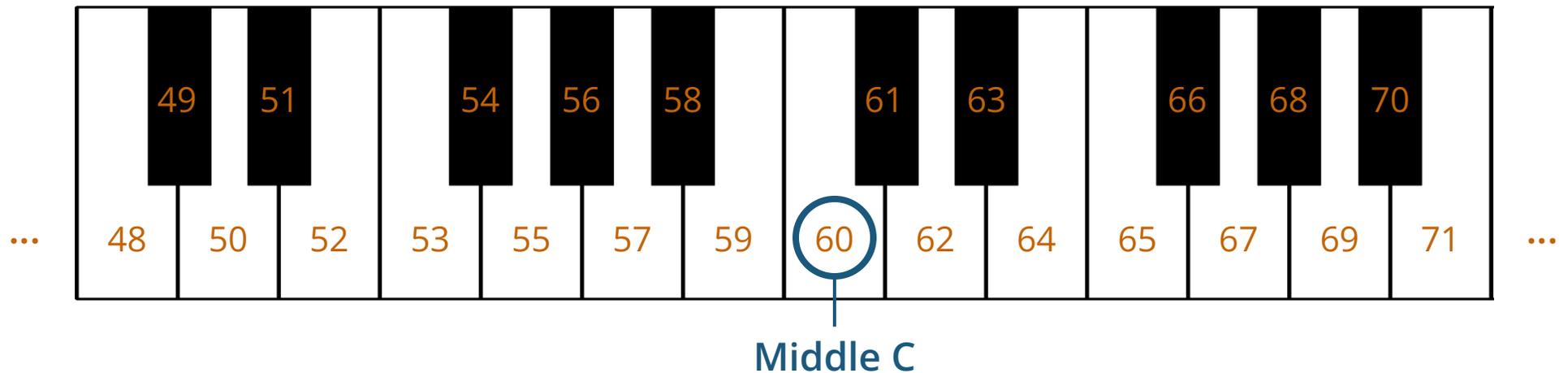
MIDI (Musical Instrument Digital Interface)

- A communication protocol between devices

- MIDI Messages
  - Note on
  - Note off
  - Delta time
  - Program change
  - Control change
  - Pitch bend change
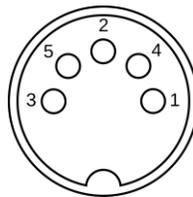
MIDI I/O

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

Middle C

Representing Music using MIDI Messages

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

Sunshine on the Meadow

```
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, ...

Expressive Timing & Dynamics

- MIDI-like representations allow **expressive timing** and **dynamics**

Quantized MIDI

Recorded MIDI
(performed by Sageev Oore)

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.

47

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) — Expressive timing
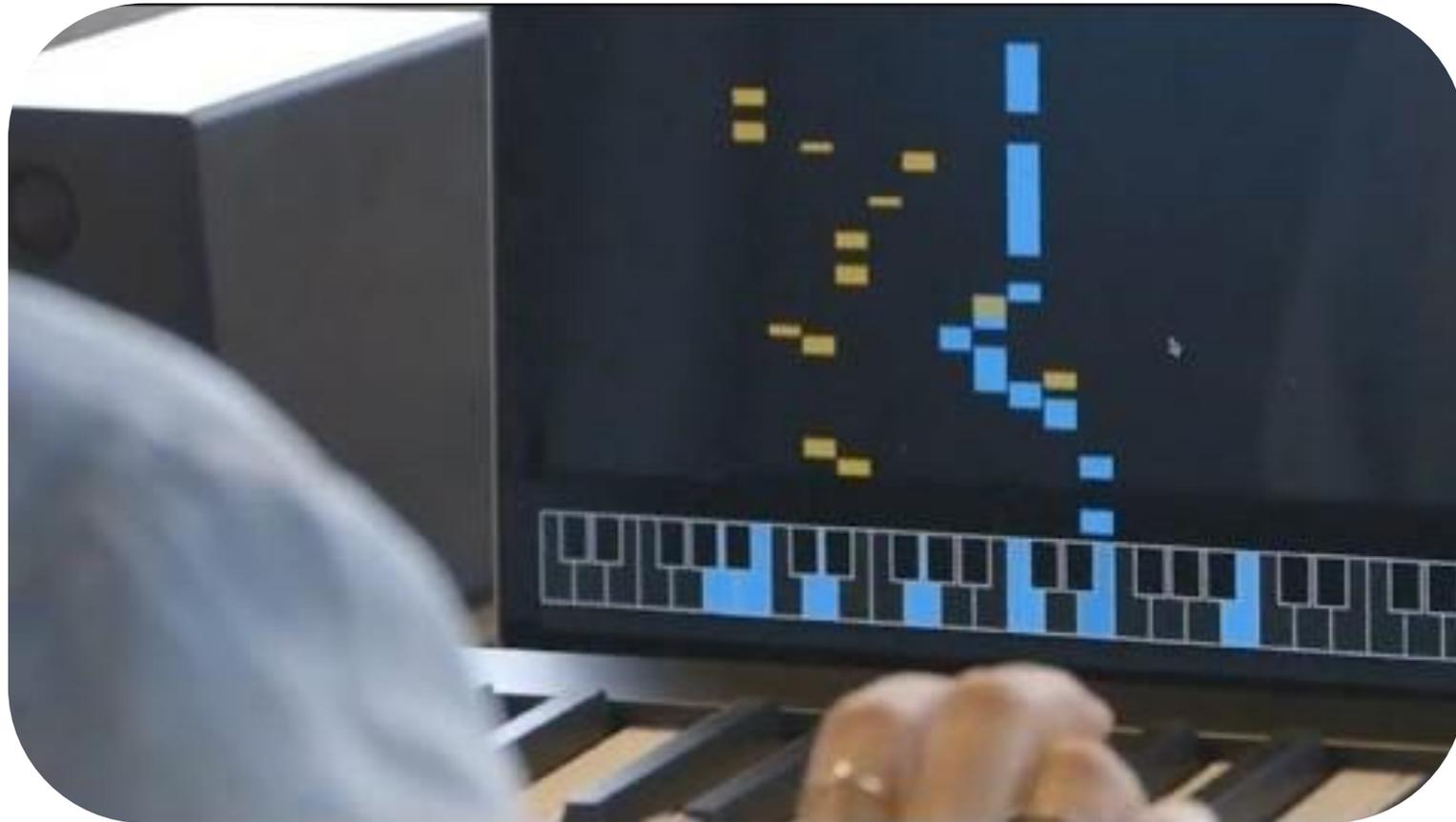 - 32 Set-Velocity events — Expressive dynamics

- **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.

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

youtu.be/0ZE1bfPtvZo
experiments.withgoogle.com/ai/ai-duet/view

Variants of RNNs

Language Models (Mathematically)

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

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

Next word   Previous words

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

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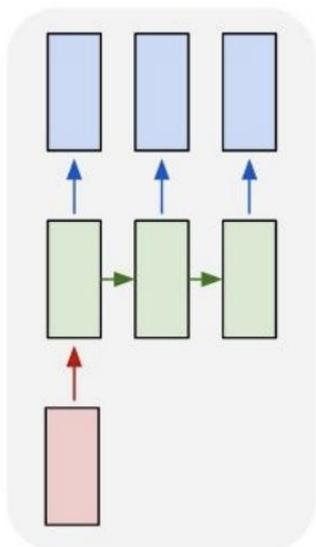
$P(\text{ gene} \mid \text{A transformer is a })$ ⬆

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

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

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

51

Types of RNNs

one to many

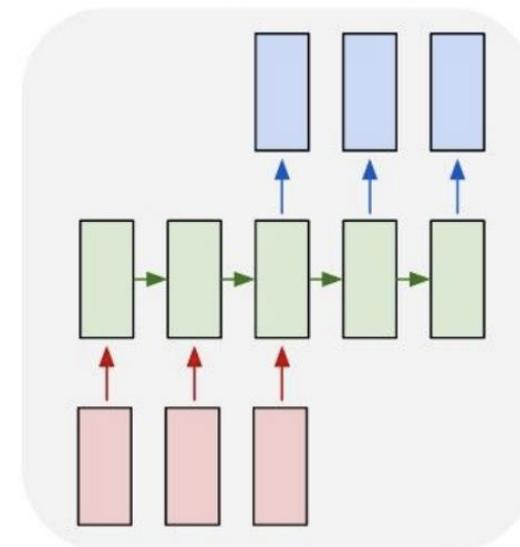**Text generation**
Music generation

many to one
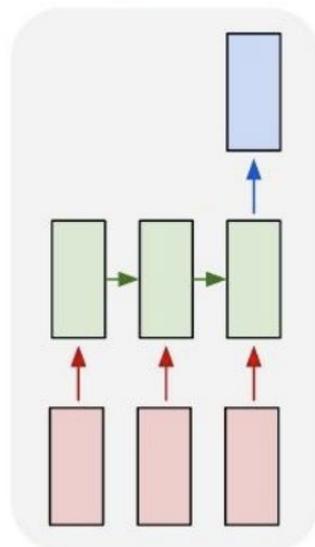
Sentiment classification
Genre classification

many to many
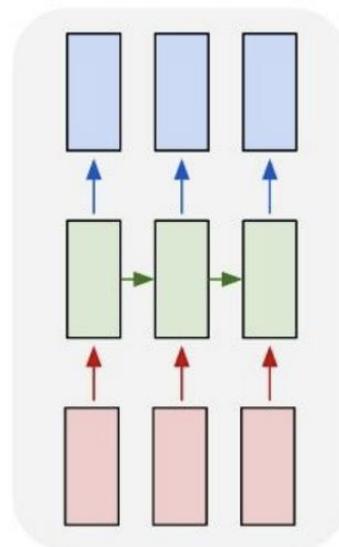
Name entity recognition
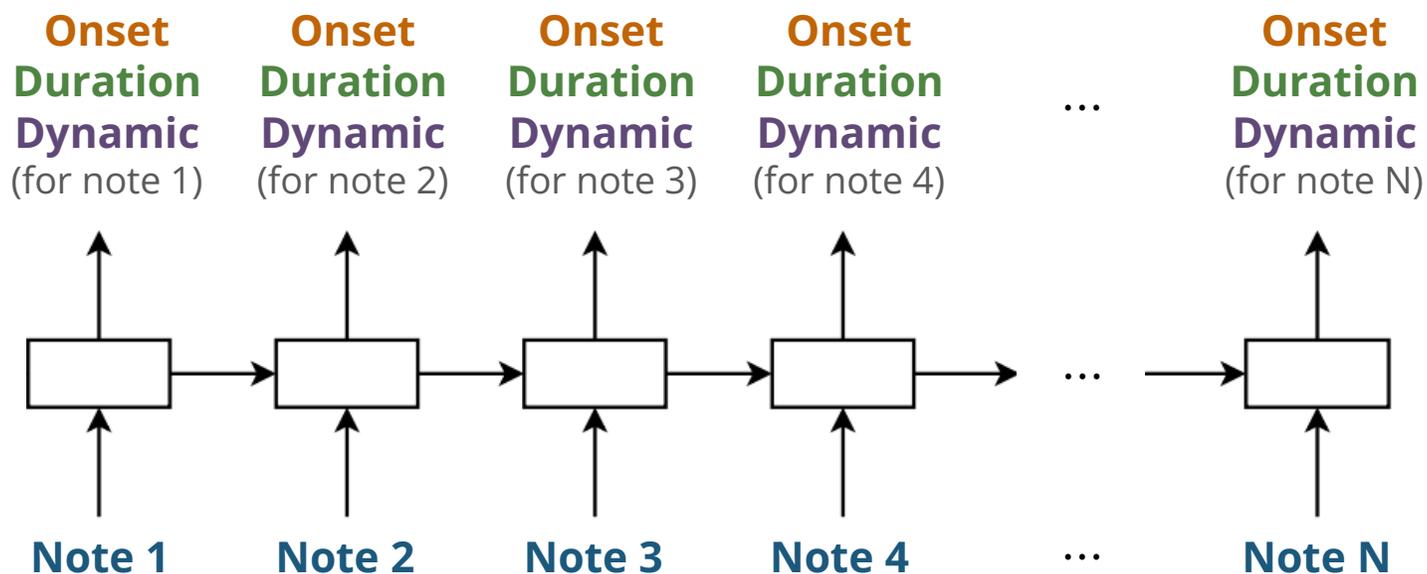**Performance rendering**

many to many

Machine translation
Music accompaniment
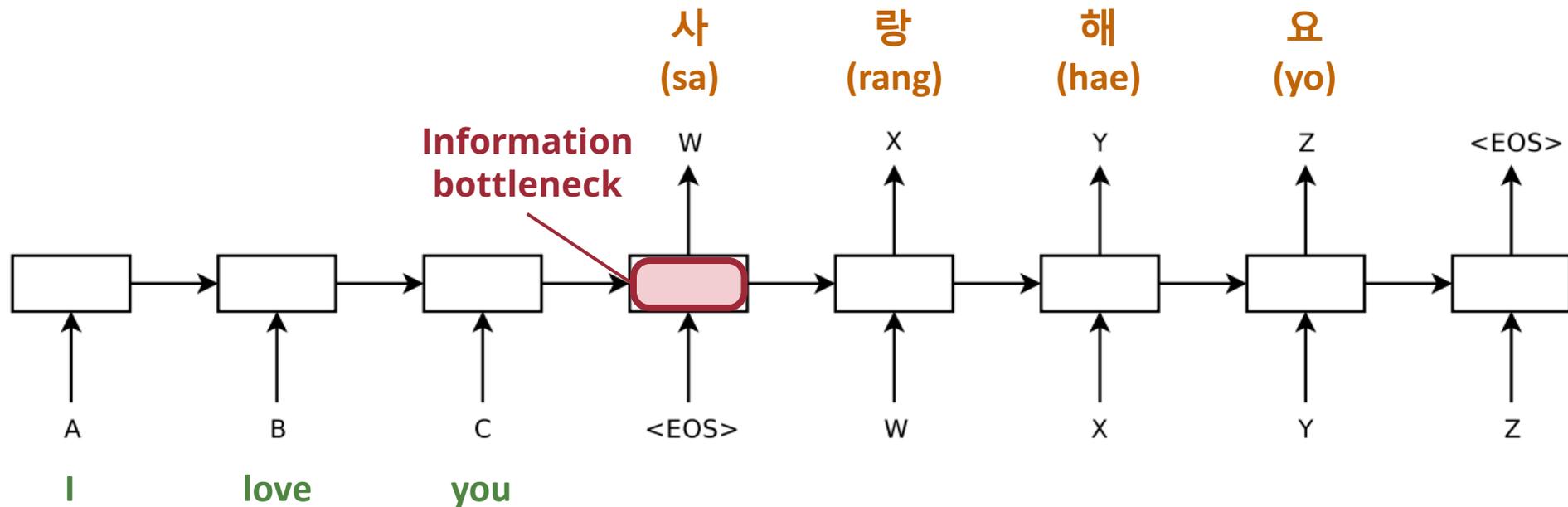**Style Transfer**

(Source: CS231n)

Many-to-Many RNNs for Performance Rendering
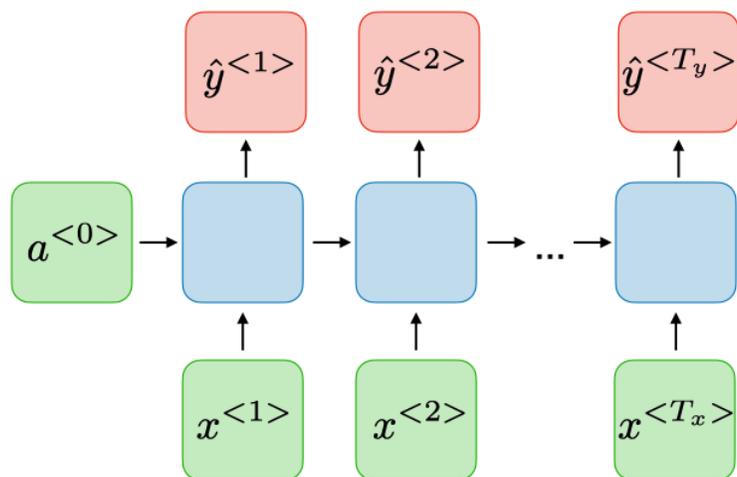
- Inputs and outputs are **aligned sequences**

Sequence-to-Sequence Model (Seq2seq)

- Widely used for **machine translation**
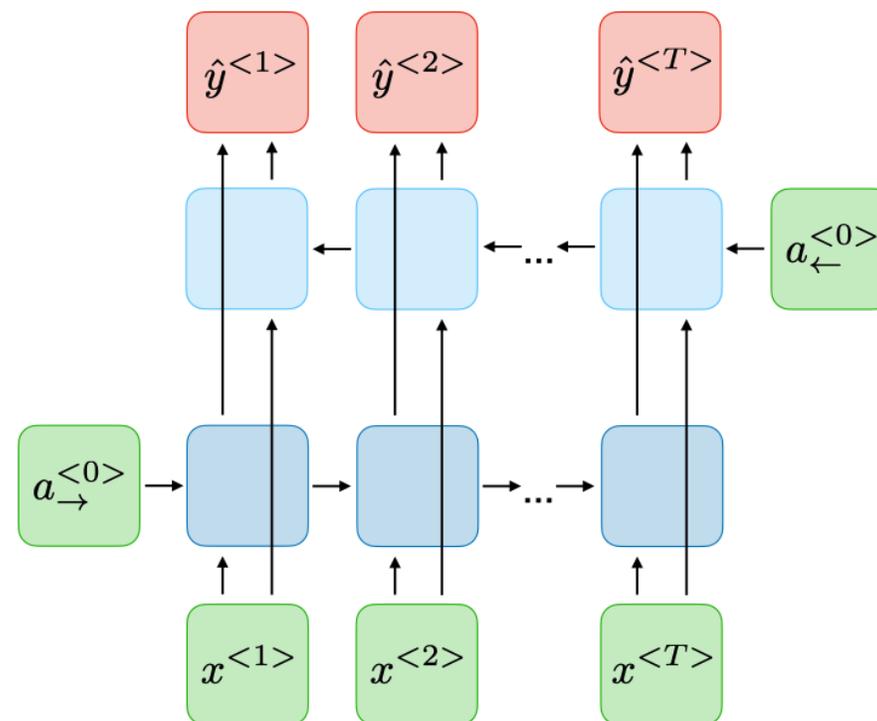
- Inputs and outputs are **unaligned sequences**

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le, "Sequence to Sequence Learning with Neural Networks," *NeurIPS*, 2014.

Unidirectional vs Bidirectional RNNs

Unidirectional RNNs
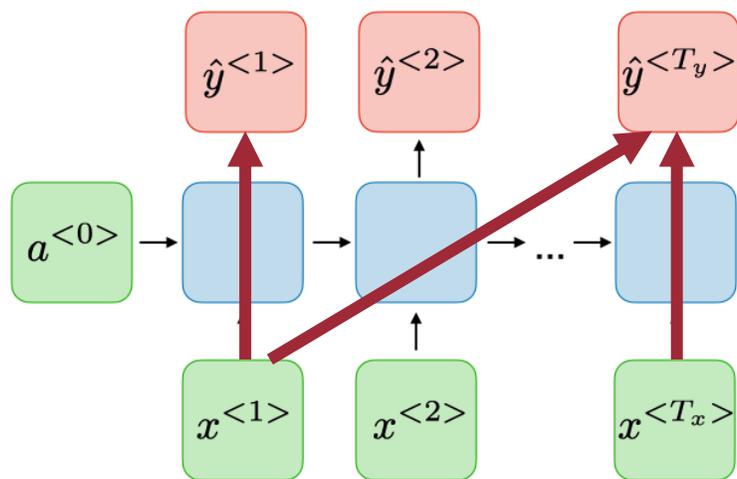
Bidirectional RNNs

Access to only past information

Access to past & future information

(Source: Amidi & Amidi, 2019)

Unidirectional vs Bidirectional RNNs

Unidirectional RNNs

Bidirectional RNNs

Access to only past information
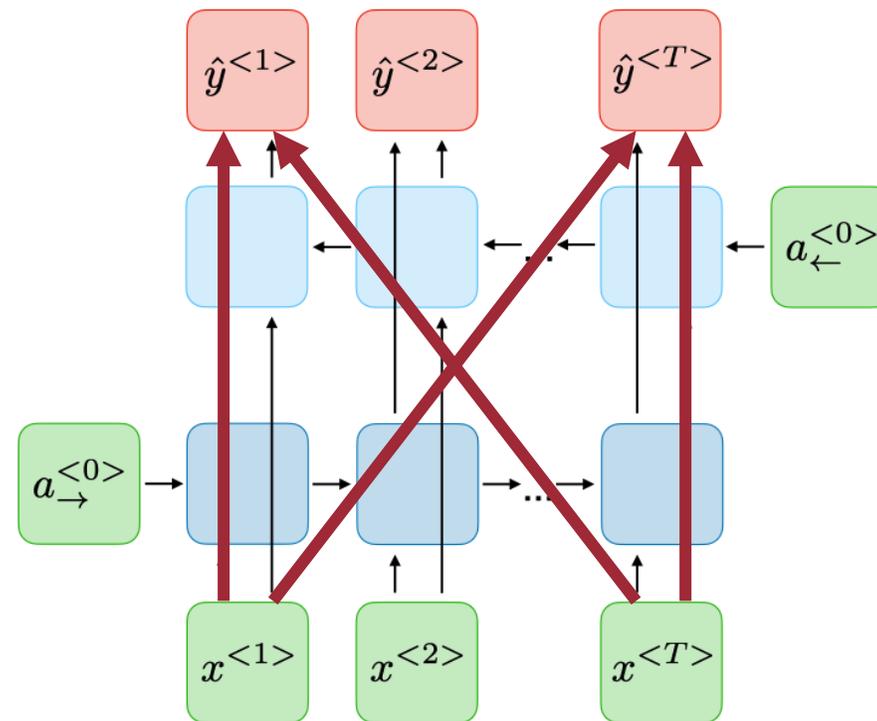
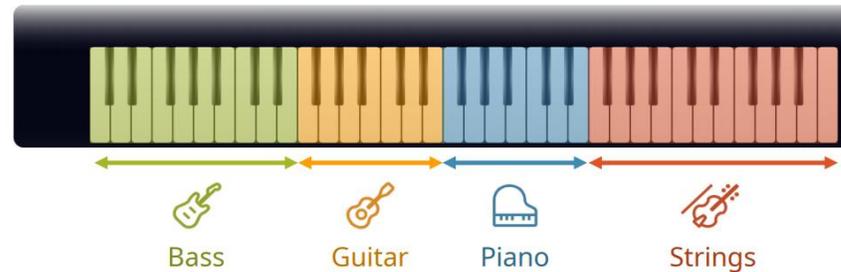**Access to past & future information**

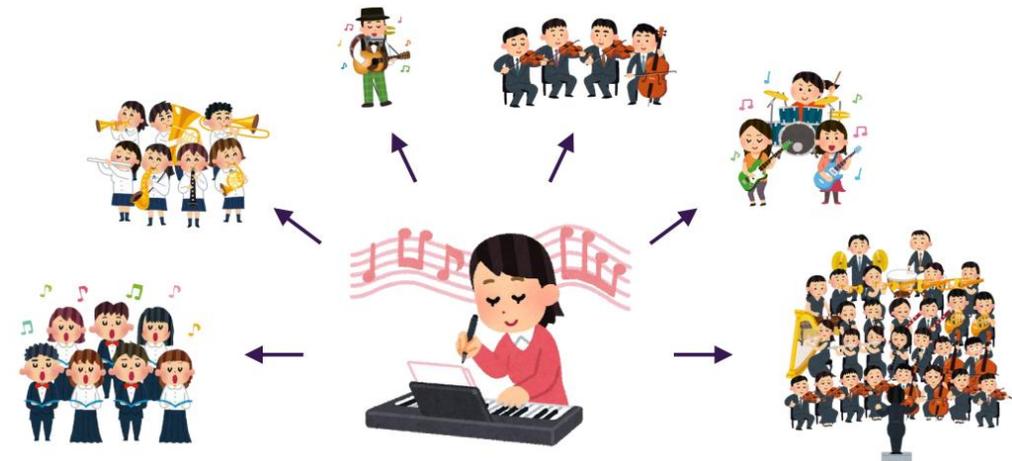(Source: Amidi & Amidi, 2019)

Automatic Instrumentation

Automatic Instrumentation

- **Goal**: Dynamically assign instruments to notes in solo music

Intelligent Musical Instruments

Bass   Guitar   Piano   Strings

Assistive Composing Tools

How can we acquire paired data?

Hao-Wen Dong, Chris Donahue, Taylor Berg-Kirkpatrick, and Julian McAuley, "Towards Automatic Instrumentation by Learning to Separate Parts in Symbolic Multitrack Music," *ISMIR*, 2021.

Arranger (Dong et al., 2021)

Hao-Wen Dong, Chris Donahue, Taylor Berg-Kirkpatrick, and Julian McAuley, "Towards Automatic Instrumentation by Learning to Separate Parts in Symbolic Multitrack Music," *ISMIR*, 2021.

59
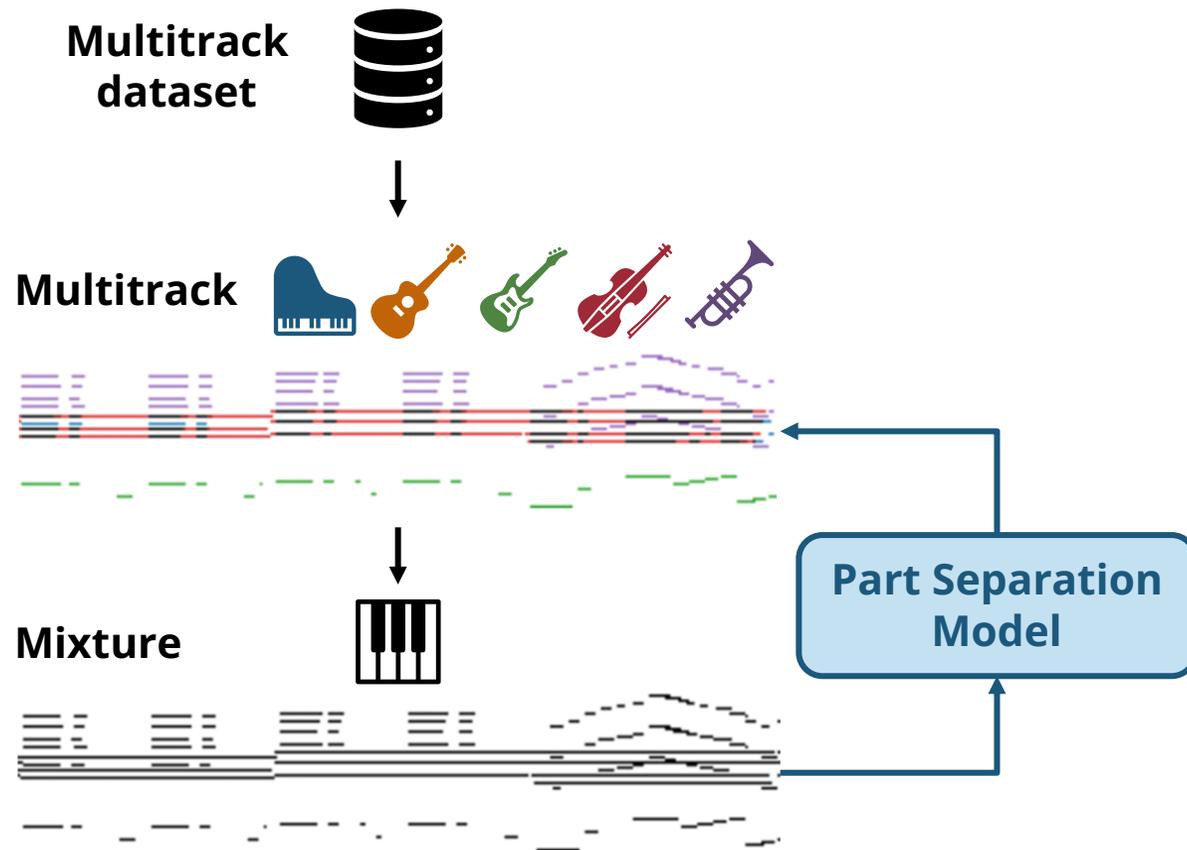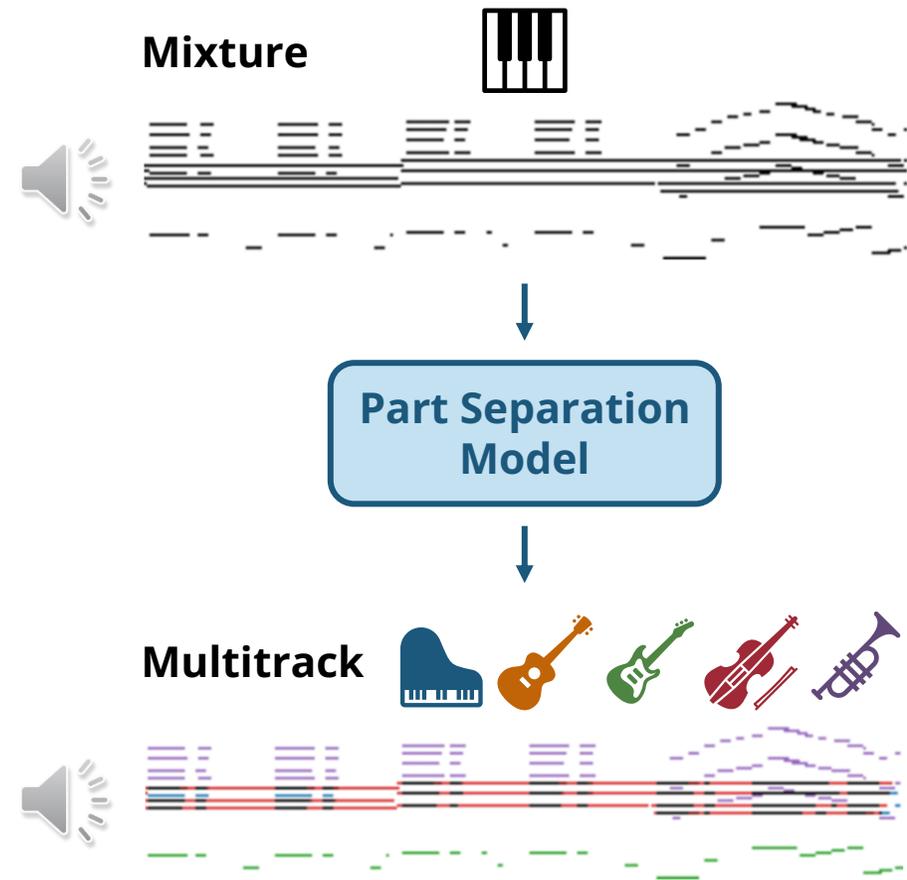
Online vs. Offline Model

Intelligent Musical Instruments

Bass   Guitar   Piano   Strings
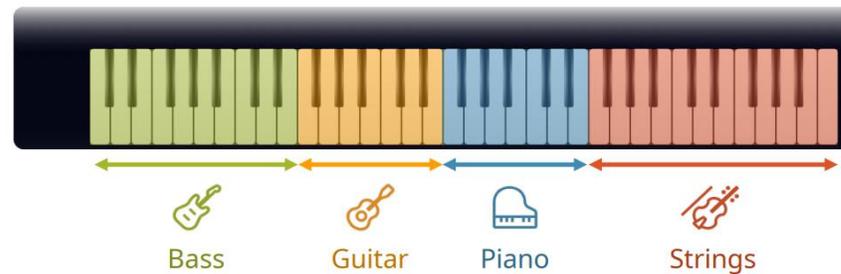
Can only look at the **past**

Assistive Composing Tools

Can look at both the **future** and the **past**
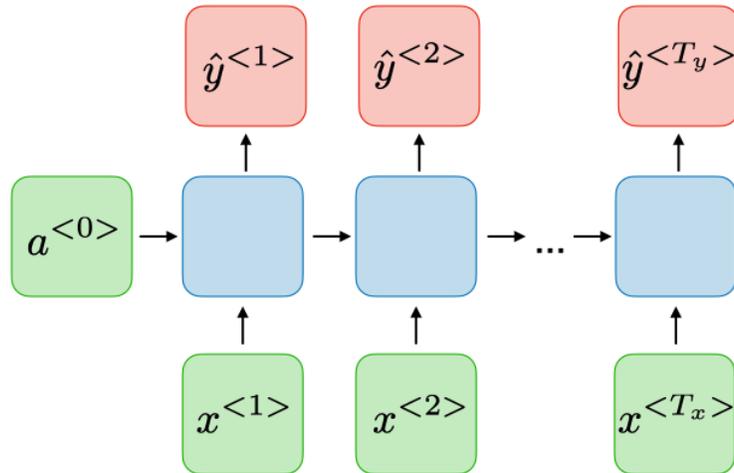
Hao-Wen Dong, Chris Donahue, Taylor Berg-Kirkpatrick, and Julian McAuley, "Towards Automatic Instrumentation by Learning to Separate Parts in Symbolic Multitrack Music," *ISMIR*, 2021.
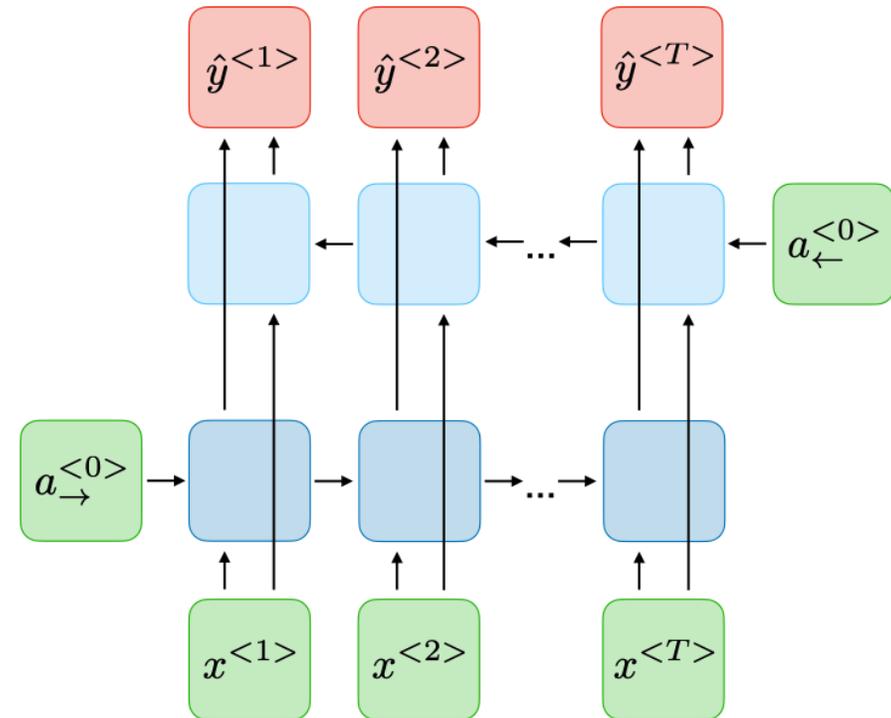
LSTMs vs. BiLSTMs

LSTMs

BiLSTMs

Access to only past information

Access to past & future information

(Source: Amidi & Amidi, 2019)

Arranger (Dong et al., 2021)

piano, guitar, bass, strings, brass

Original

LSTM
(w/o entry hints)
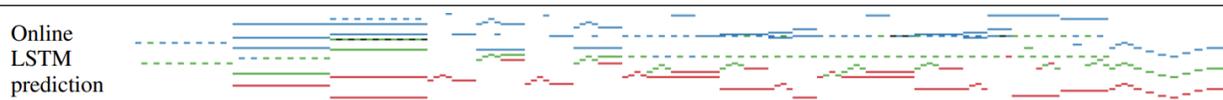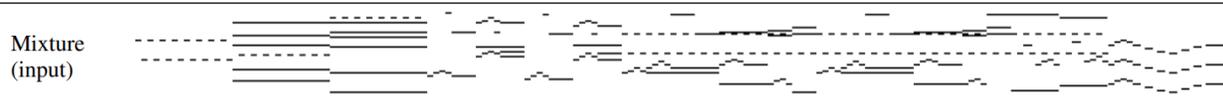
BiLSTM
(w/ entry hints)

(Source: Dong et al., 2021)

Hao-Wen Dong, Chris Donahue, Taylor Berg-Kirkpatrick, and Julian McAuley, "Towards Automatic Instrumentation by Learning to Separate Parts in Symbolic Multitrack Music," *ISMIR*, 2021.

62

Arranger (Dong et al., 2021)

Bach Chorales

Musical score

Ground truth
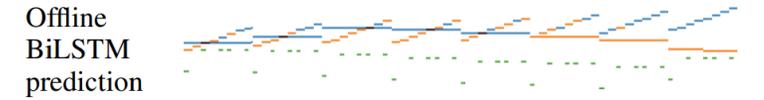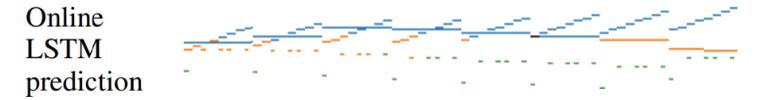
Online LSTM prediction

Offline BiLSTM prediction

(Audio available. [1] Colors: soprano, alto, tenor, bass.)

String Quartets

Musical score

Mixture (input)

Ground truth

Online LSTM prediction

Offline BiLSTM prediction

(Audio available. [1] Colors: first violin, second violin, viola, cello.)

Game Music

Ground truth

Online LSTM prediction

Offline BiLSTM prediction
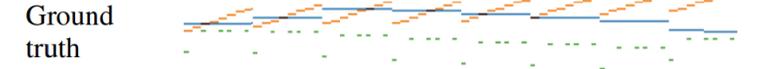
(Audio available. [1] Colors: pulse wave I, pulse wave II, triangle wave.)

Pop Music

Ground truth

Online LSTM prediction

Offline BiLSTM prediction
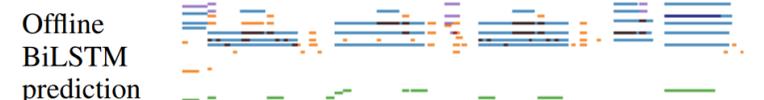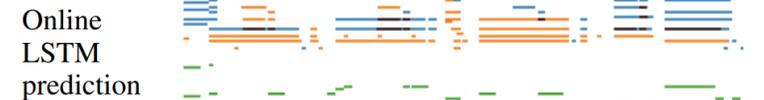
(Audio available. [1] Colors: piano, guitar, bass, strings, brass.)

(Source: Dong et al., 2021)

63

Recap

Deep Autoregressive Models

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

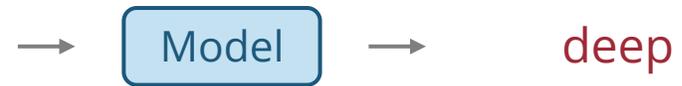|		
A transformer is a	→ Model →	deep
A transformer is a deep	→ Model →	learning
A transformer is a deep learning	→ Model →	model
A transformer is a deep learning model	→ Model →	introduced
A transformer is a deep learning model introduced	→ Model →	in
A transformer is a deep learning model introduced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

Deep Autoregressive Models

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

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

Next word   Previous words

The whole sentence
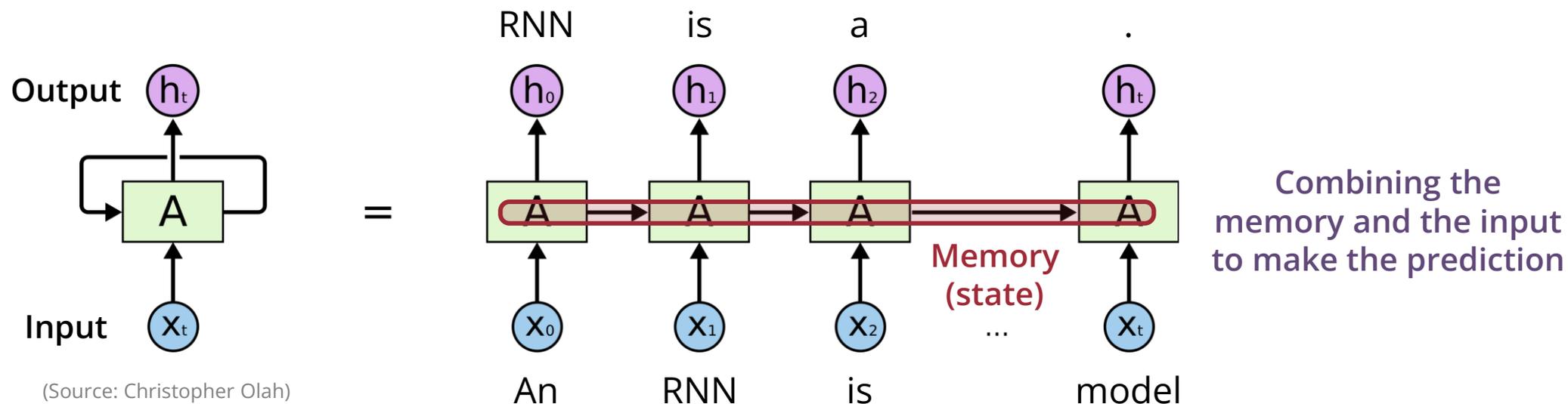$X = (x_0, x_1, \ldots, x_N)$

2nd word given 1st word   Last word given all previous words

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

1st word   3rd word given 1st & 2nd words

66

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)

Output — Input diagram with **Memory (state)** and outputs RNN, is, a, . / inputs An, RNN, is, model

Combining the memory and the input to make the prediction

Vanilla RNNs: Training vs. Inference

(Source: Christopher Olah)

(Source: Christopher Olah)

Demystifying LSTMs (Hochreiter & Schmidhuber, 1997)

Output

h_t

Long-term memory module

Whether to erase the stored memory?

Update the memory

Combining the memory and the input to make the prediction

x_t

Input

(Source: Christopher Olah)

Demystifying LSTMs (Hochreiter & Schmidhuber, 1997)

(Source: Christopher Olah)

70

Music Language Models (Mathematically)

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

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

Next note   Previous notes

$P(\ G \mid C\ C\ G\ G\ A\ A\ )$ ⬆

$P(\ A \mid C\ C\ G\ G\ A\ A\ )$ ⬆

$P(\ C \mid C\ C\ G\ G\ A\ A\ )$ ⬆

$P(\ F \mid C\ C\ G\ G\ A\ A\ )$ ⬆

$P(\ A\flat \mid C\ C\ G\ G\ A\ A\ )$ ⬇

$P(\ A\sharp \mid C\ C\ G\ G\ A\ A\ )$ ⬇

Designing a Machine-readable Music Language

- How can we "represent" music in a way that machines understand?

Folk RNN (Sturm et al., 2015)

- **Data**
  - 23,958 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.

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, ...
```

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) ⟵ Expressive timing
  - 32 Set-Velocity events ⟵ Expressive 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.

A.I. Duet (Mann et al, 2016)
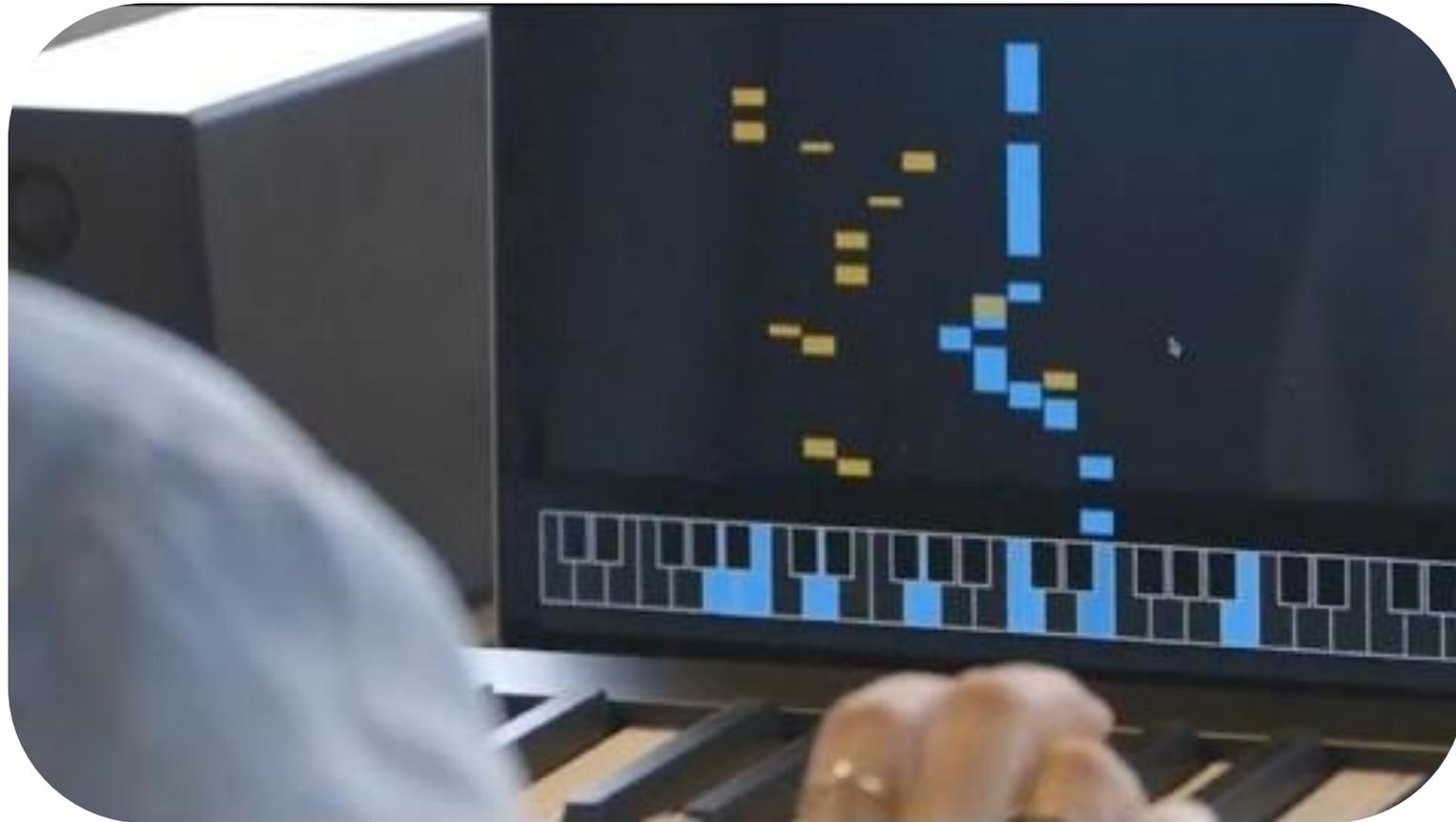
youtu.be/0ZE1bfPtvZo
experiments.withgoogle.com/ai/ai-duet/view

Arranger (Dong et al., 2021)

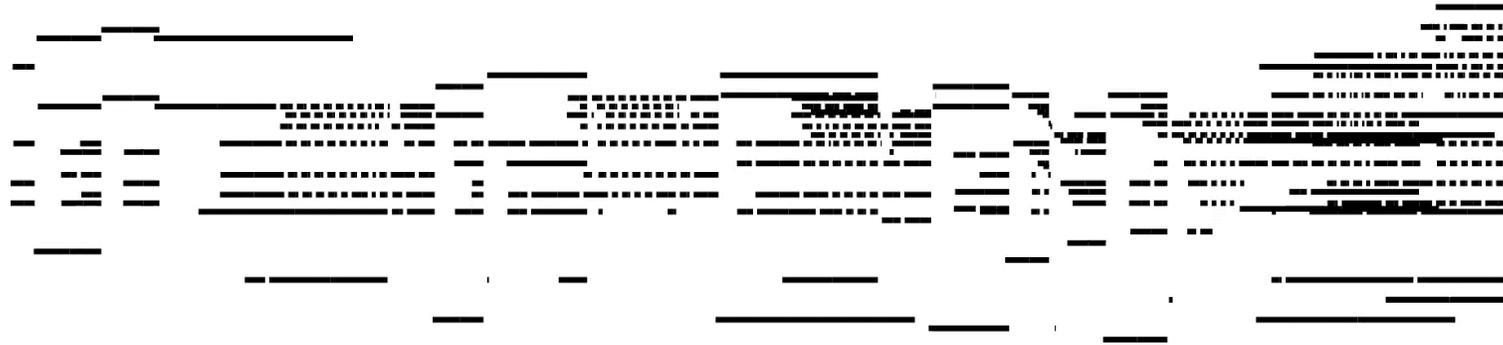piano, guitar, bass, strings, brass

Original

LSTM
(w/o entry hints)

BiLSTM
(w/ entry hints)

(Source: Dong et al., 2021)

Hao-Wen Dong, Chris Donahue, Taylor Berg-Kirkpatrick, and Julian McAuley, "Towards Automatic Instrumentation by Learning to Separate Parts in Symbolic Multitrack Music," *ISMIR*, 2021.

77

Next Lecture

Transformers

(Source: Huang et al., 2018)

UNIVERSITY OF MICHIGAN