

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**

- Variational autoencoder (VAE)
- Generative adversarial network (GAN)
- 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



A transformer is a deep



A transformer is a deep learning



A transformer is a deep learning model



A transformer is a deep learning model introduced



A transformer is a deep learning model introduced in



Deep Autoregressive Models

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

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

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

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

Next word

The whole sentence

$$X = (x_0, x_1, \dots, x_N)$$

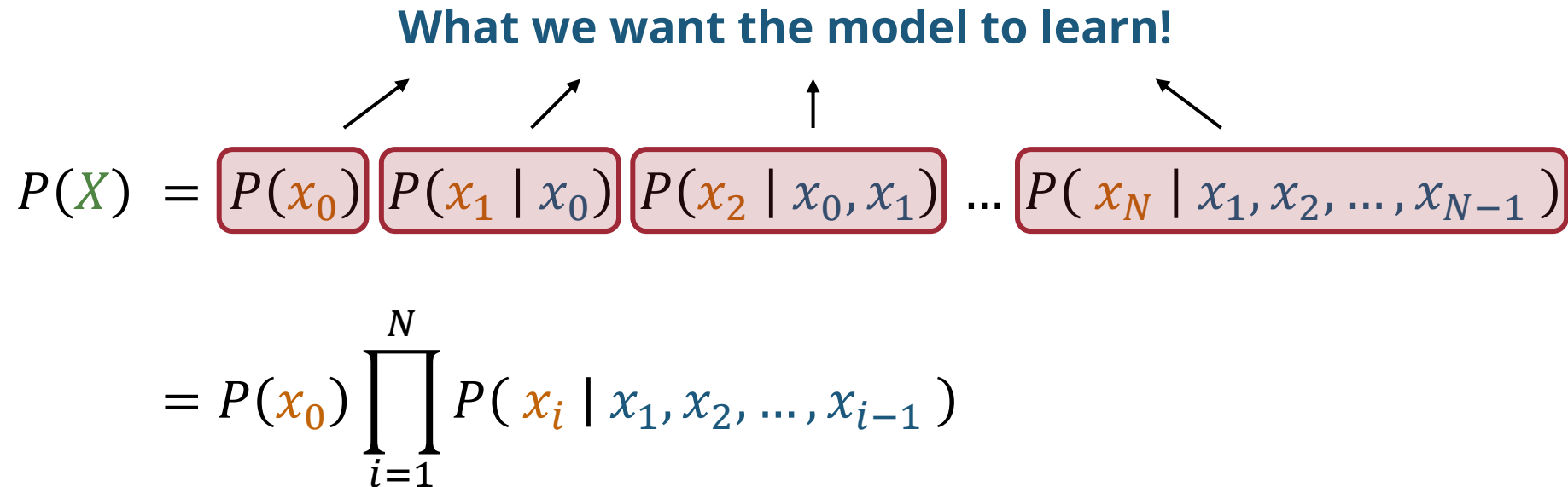
$$P(X) = P(x_0) P(x_1 | x_0) P(x_2 | x_0, x_1) \dots P(x_N | x_1, x_2, \dots, x_{N-1})$$

1st word 2nd word given 1st word 3rd word given 1st & 2nd words Last word given all previous words

Deep Autoregressive Models

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

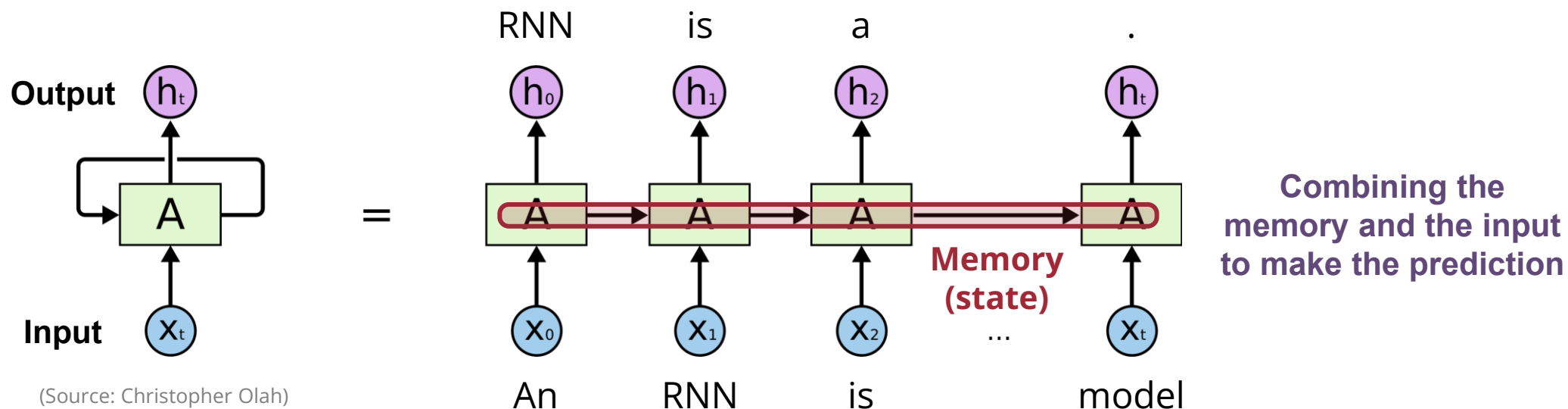
What we want the model to learn!

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


Recurrent Neural Networks (RNNs)

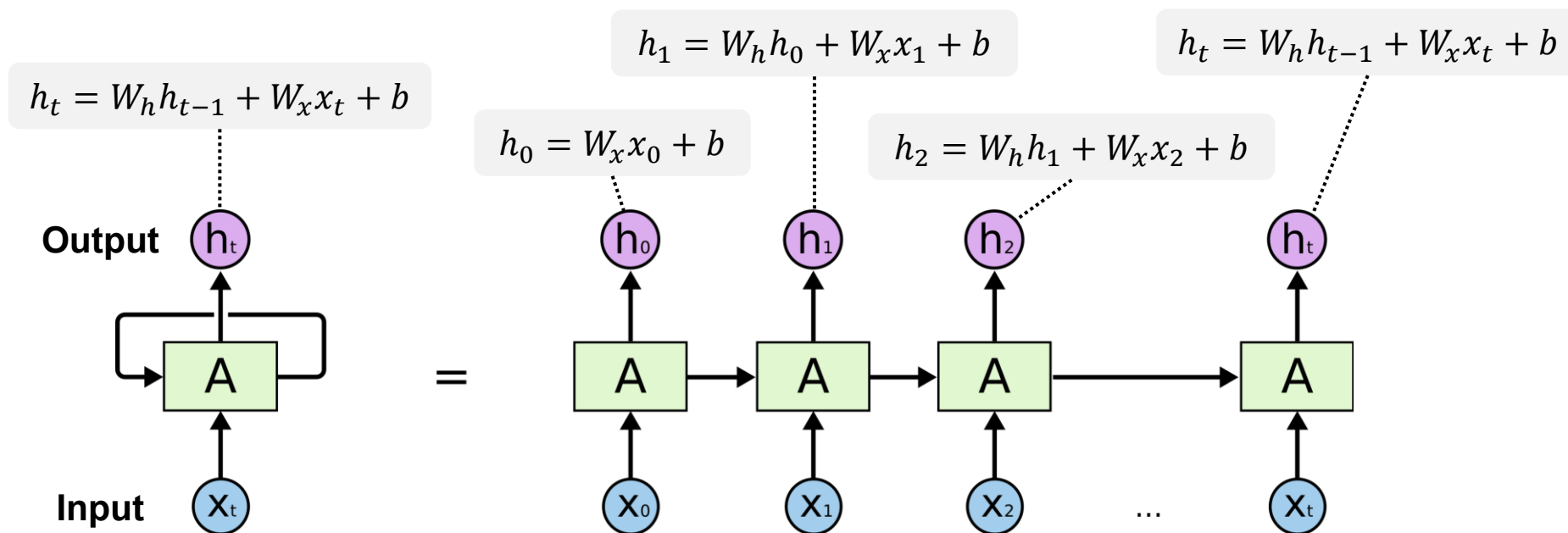
What is an RNN (Recurrent Neural Network)?

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



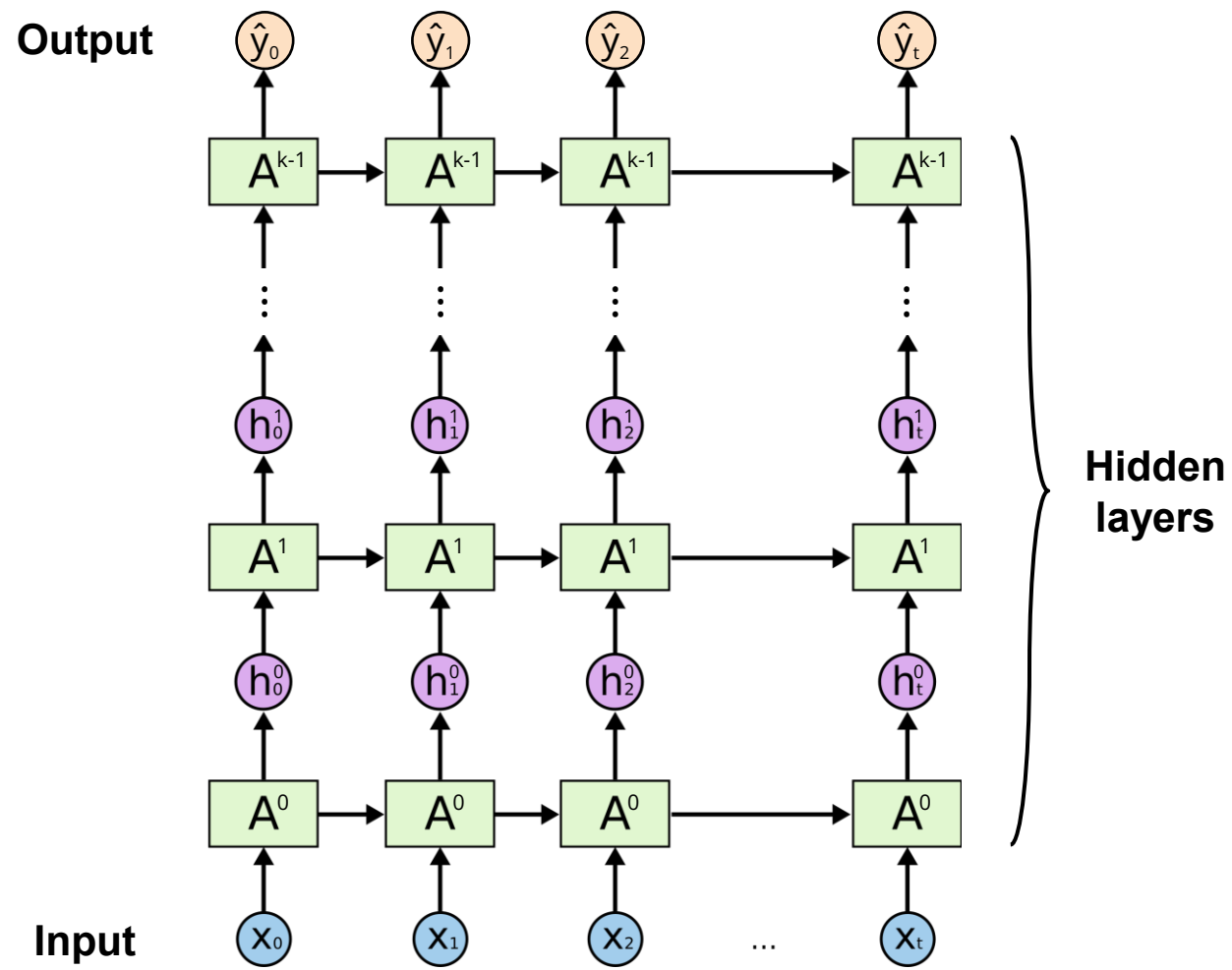
Vanilla RNNs

- The simplest form of RNNs



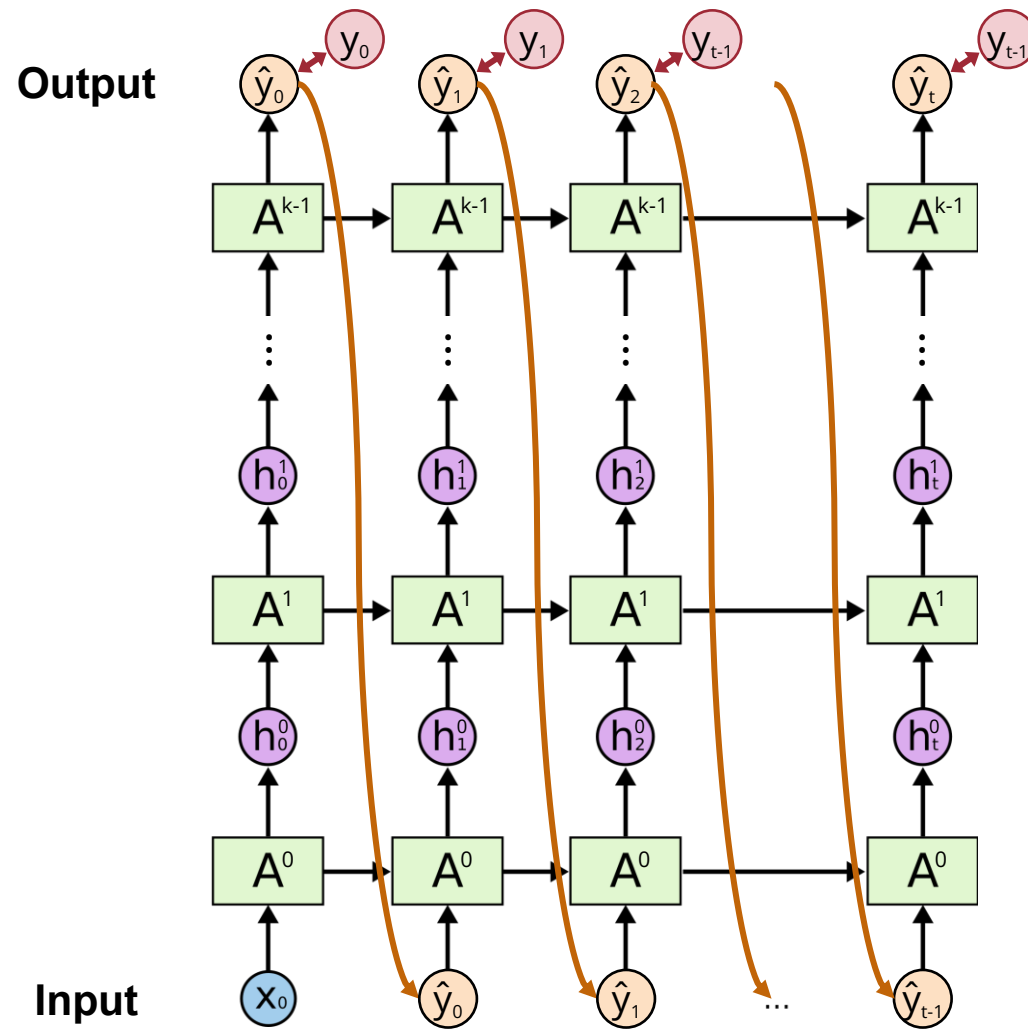
(Source: Christopher Olah)

Vanilla RNNs



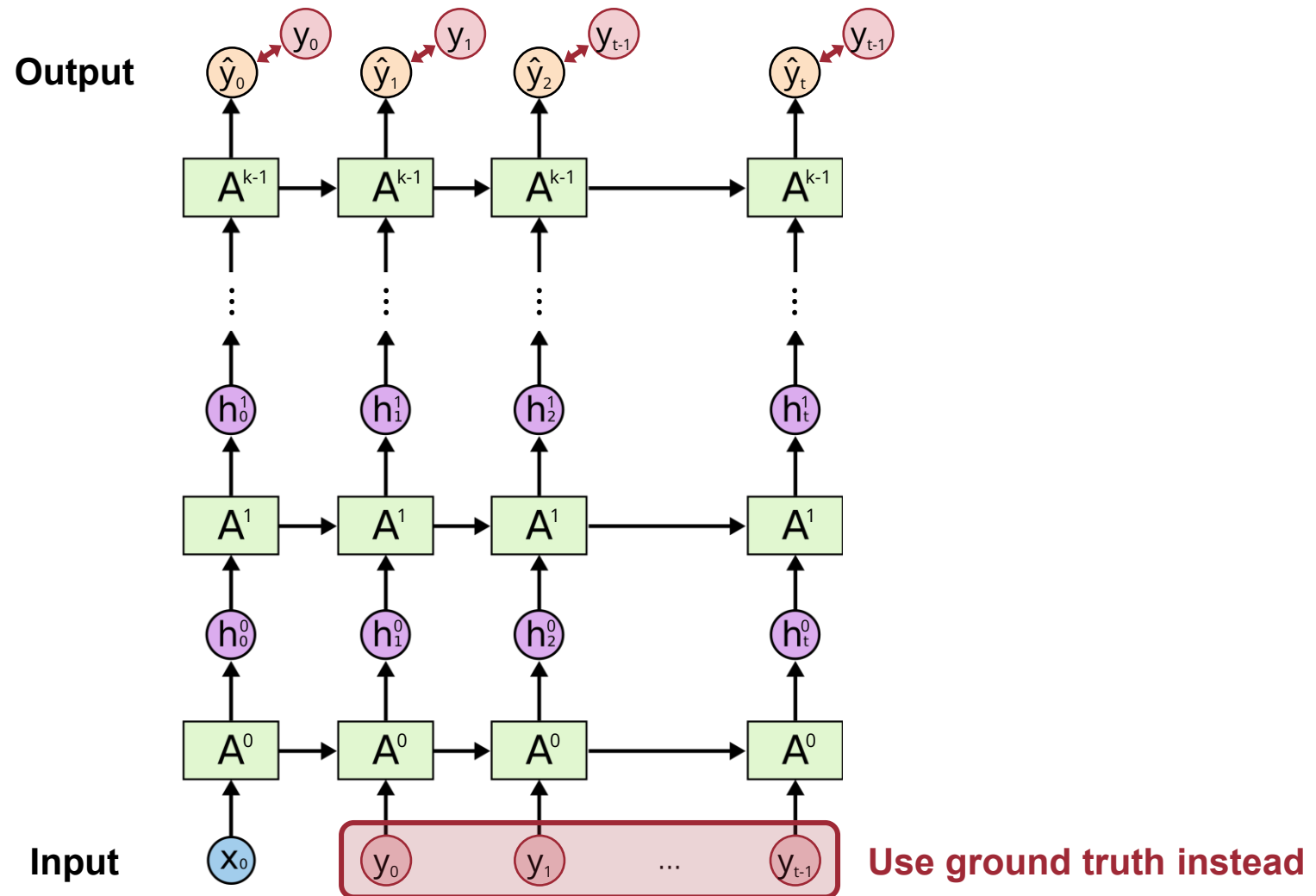
(Source: Christopher Olah)

Vanilla RNNs: Training



(Source: Christopher Olah)

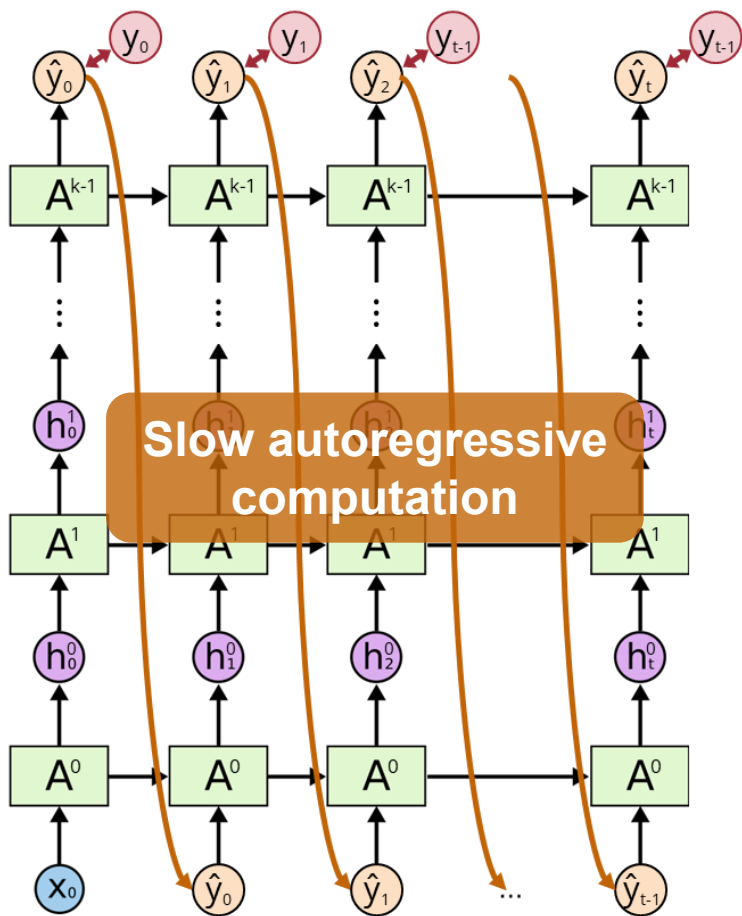
Vanilla RNNs: Training with Teacher Forcing



(Source: Christopher Olah)

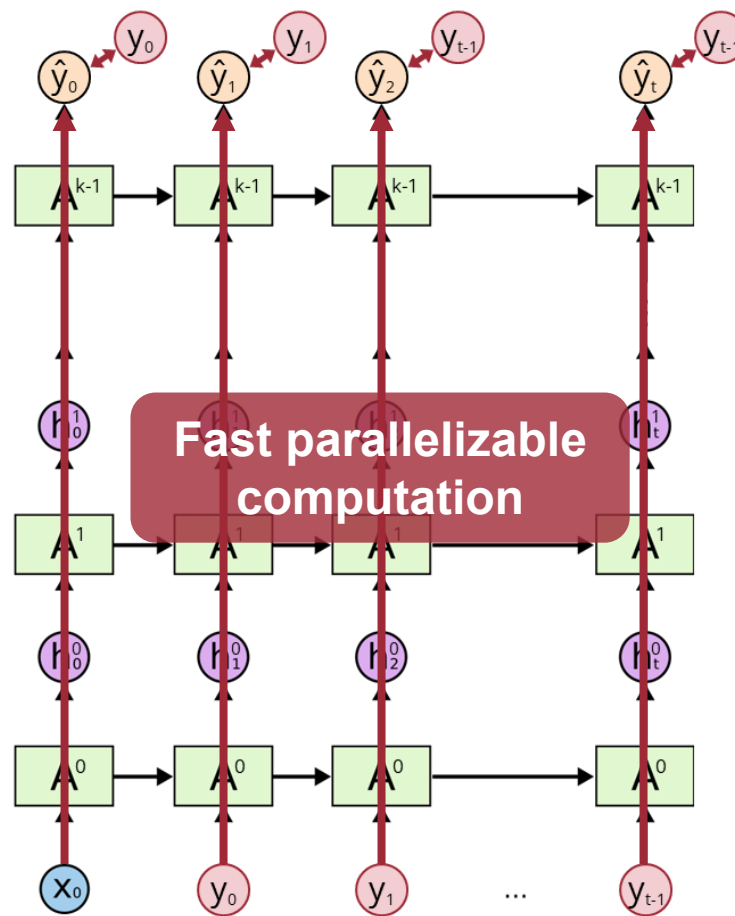
Vanilla RNNs: Training **with** vs. **without** Teacher Forcing

Without Teacher Forcing



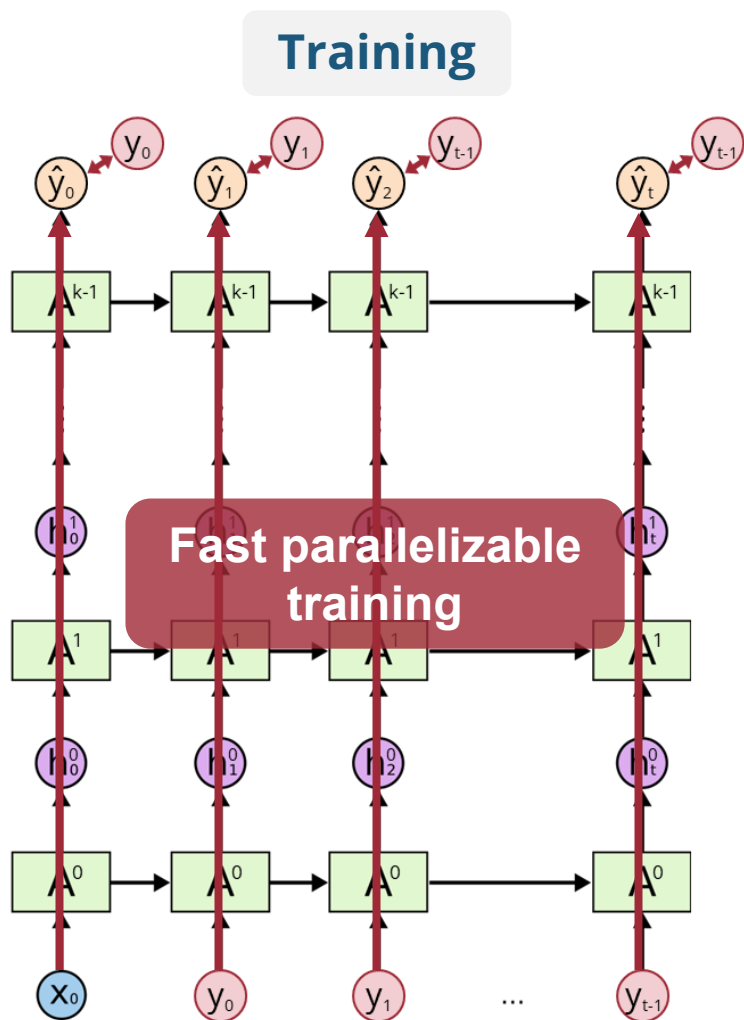
(Source: Christopher Olah)

With Teacher Forcing

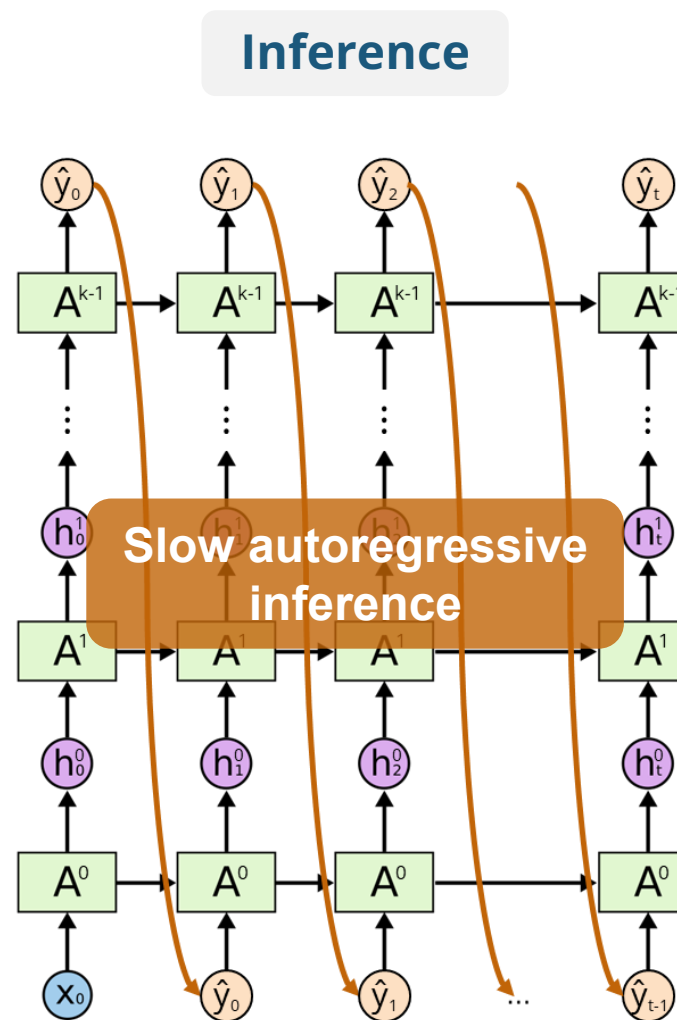


(Source: Christopher Olah)

Vanilla RNNs: Training vs. Inference



(Source: Christopher Olah)

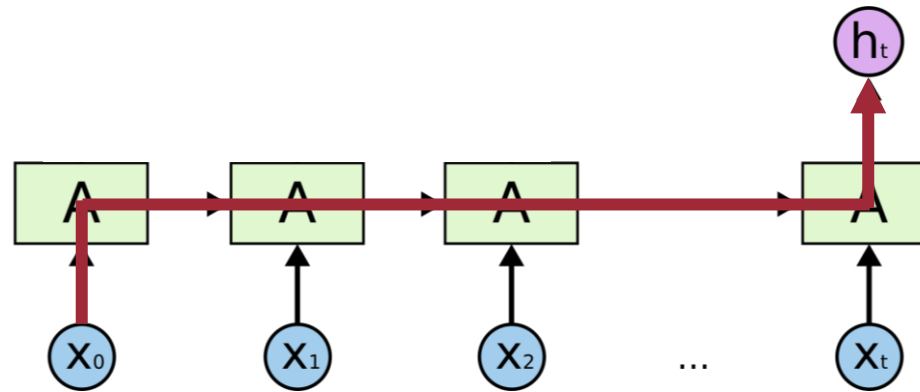


(Source: Christopher Olah)

Long Short-Term Memory (LSTMs)

Backpropagation Through Time

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



(Source: Christopher Olah)

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

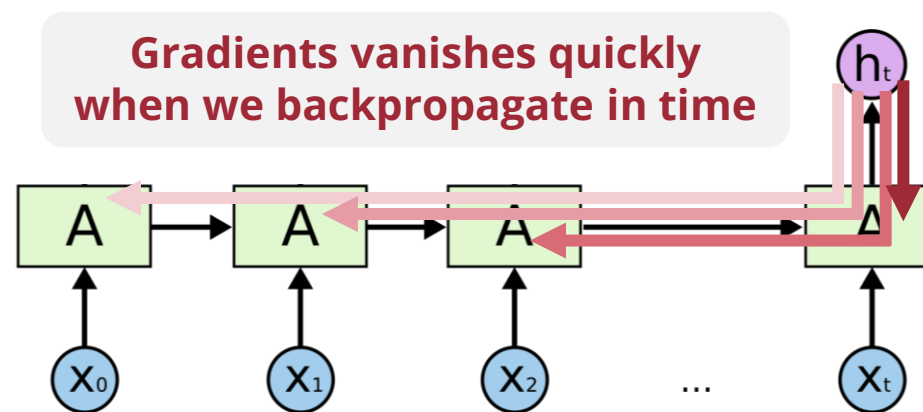
$$h_t = W_h (W_h h_{t-2} + W_x x_{t-1} + b) + W_x x_t + b$$

⋮

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

Vanishing Gradients

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



(Source: Christopher Olah)

All the layers share the same weight matrix

Can still train the model without deeper gradients

Why bother?

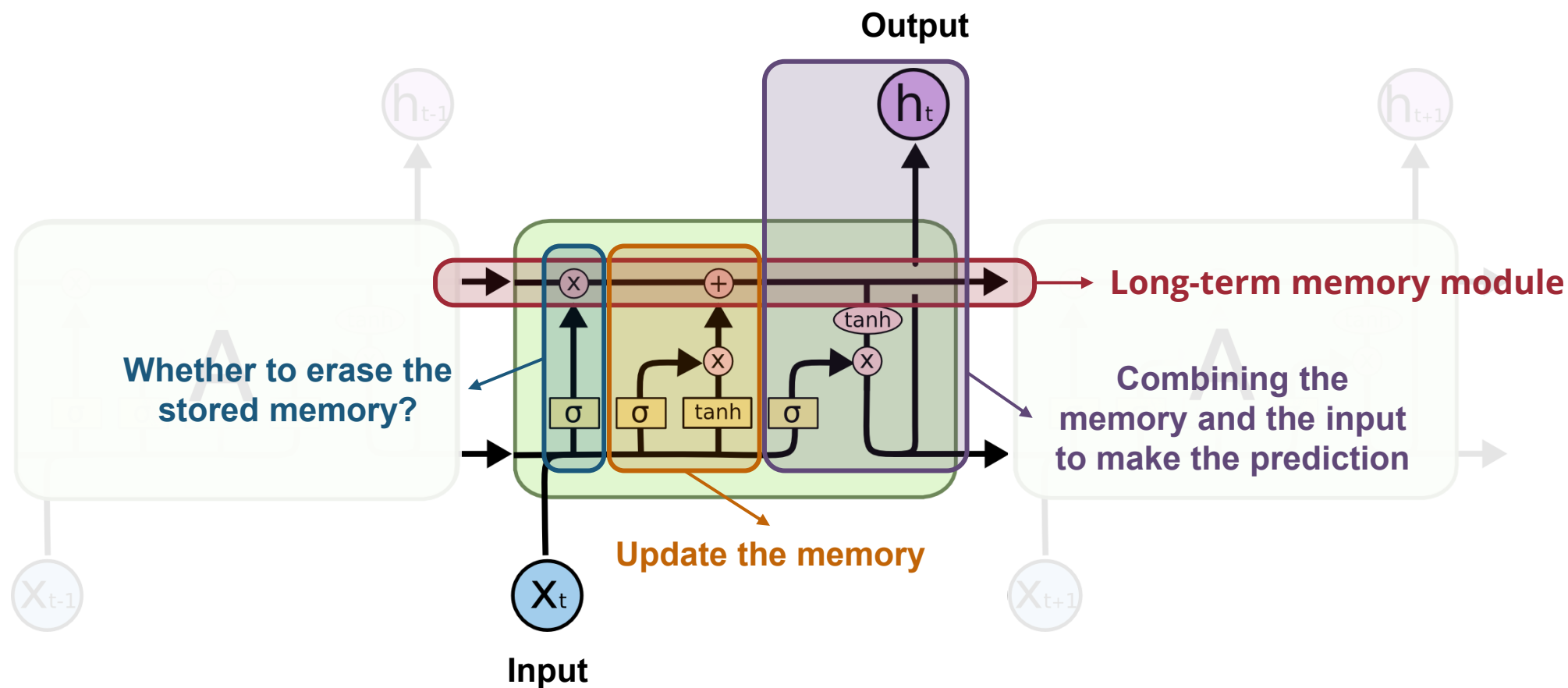
$$h_t = W_h h_{t-1} + W_x x_t + b$$

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

⋮

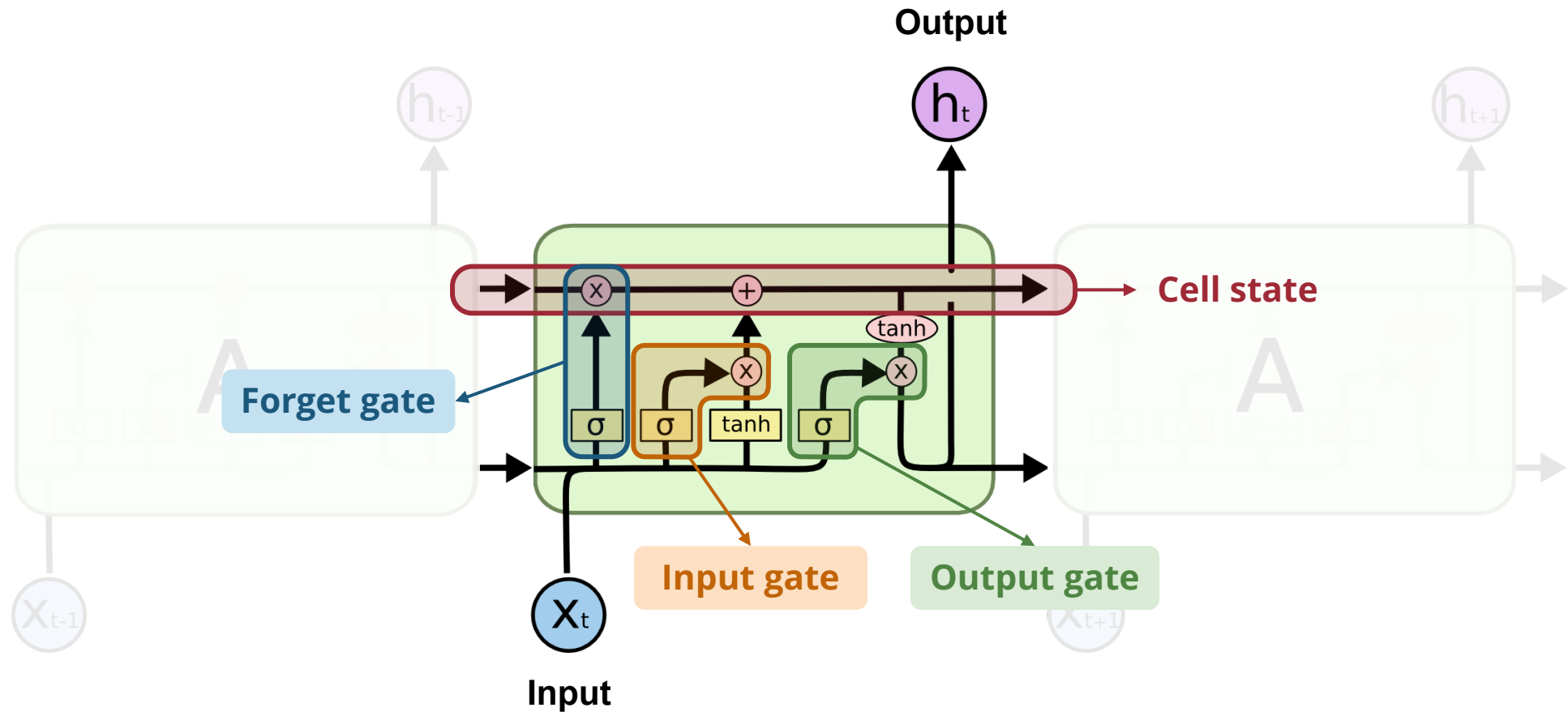
$$h_t = W_h (W_x x_{t-1} + W_h (\dots W_h h_0 + W_x x_1 + b \dots) + b) + W_x x_t + b$$

Demystifying LSTMs (Hochreiter & Schmidhuber, 1997)



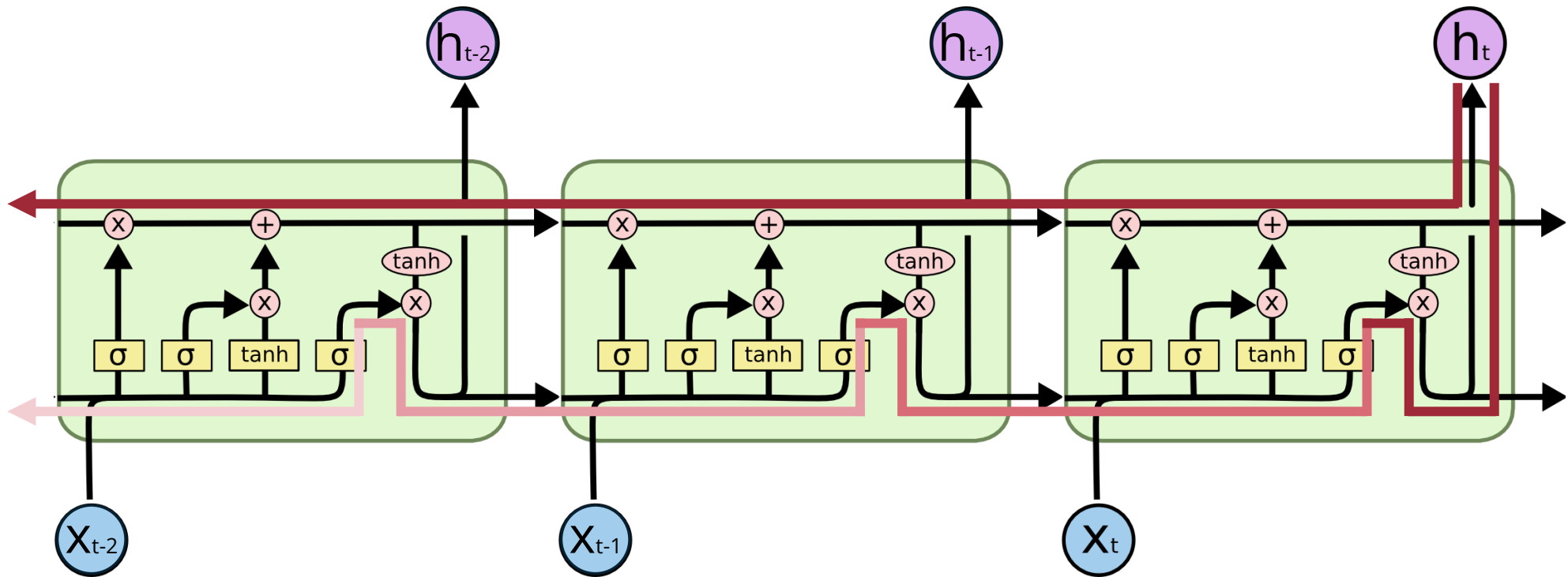
(Source: Christopher Olah)

Demystifying LSTMs (Hochreiter & Schmidhuber, 1997)



(Source: Christopher Olah)

How can LSTMs Help Alleviate Vanishing Gradients?



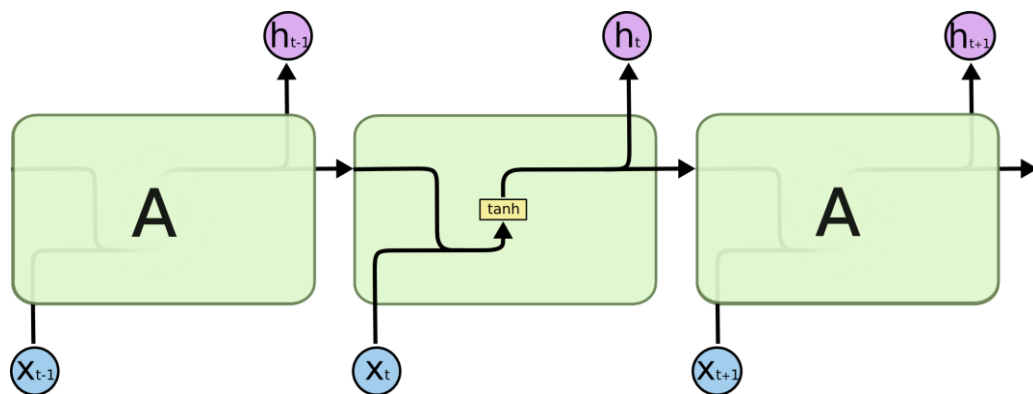
(Source: Christopher Olah)

LSTMs does not completely solve vanishing gradients

Vanilla RNNs vs. LSTMs

Vanilla RNN

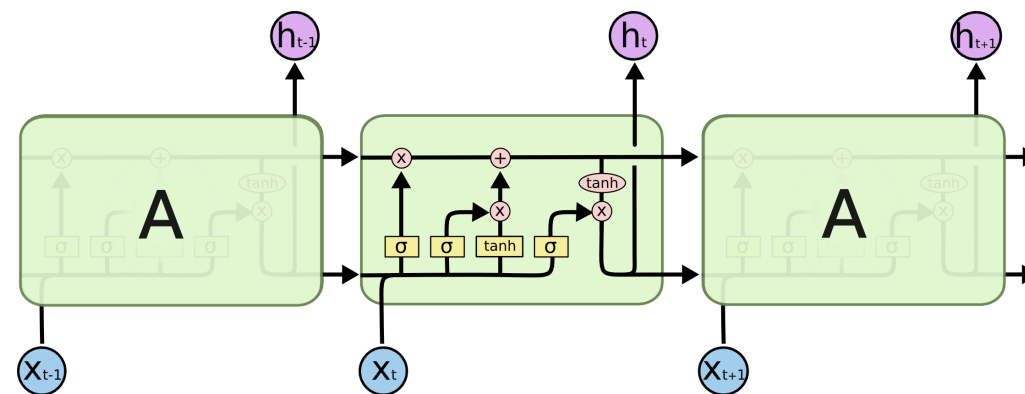
- Simplest form of RNNs
- Limited long-term memory
- Harder to train (due to gradient vanishing)



(Source: Christopher Olah)

LSTM

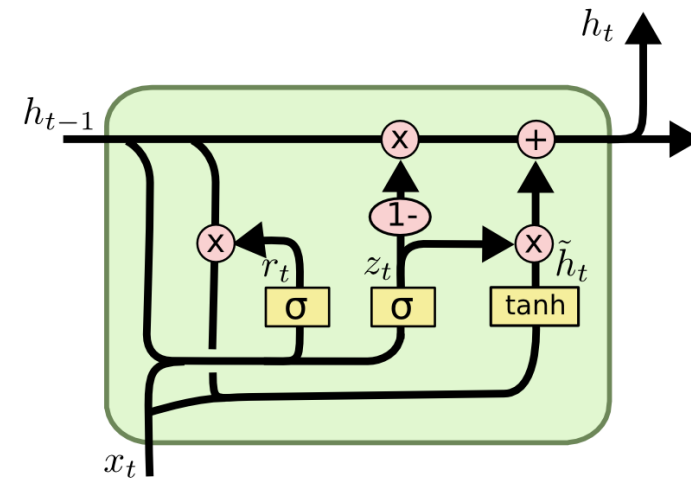
- Improved memory module
- Better long-term memory
- Easier to train



(Source: Christopher Olah)

Gated Recurrent Units (GRUs)

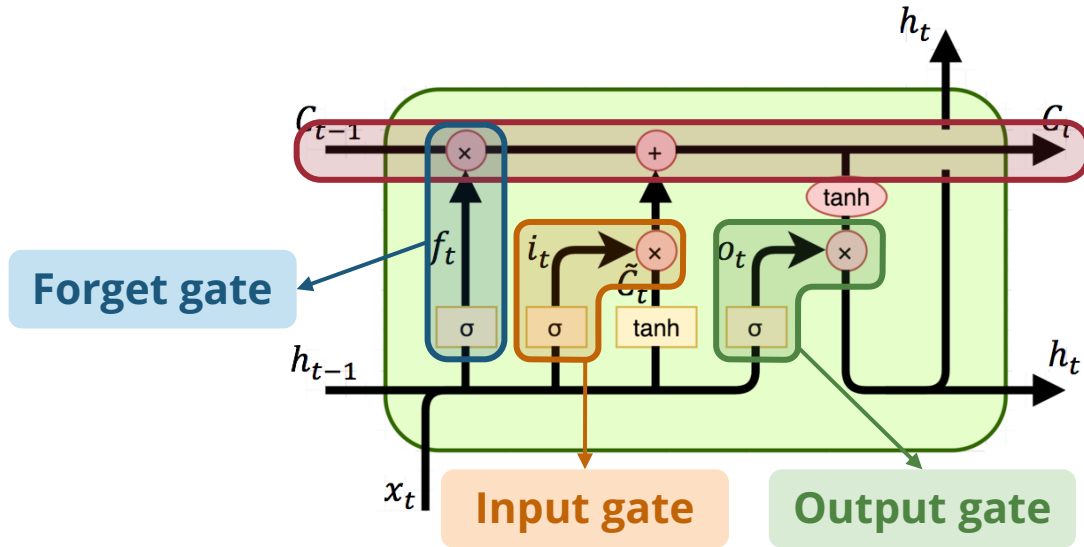
- A **simplified** version of LSTM
- An LSTM consists of
 - **Forget** gate
 - **Input** gate
 - **Output** gate
- An GRU consists of
 - **Reset** gate
 - **Update** gate



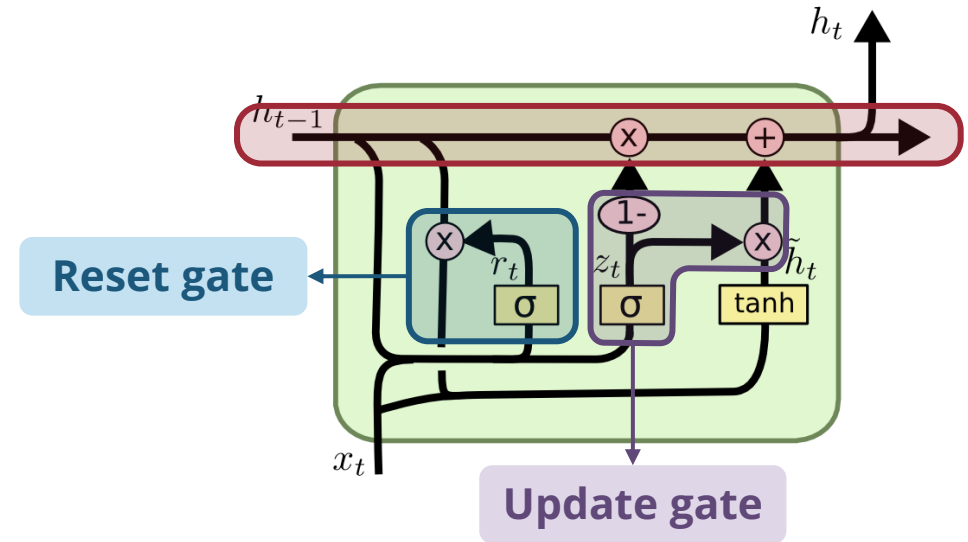
(Source: dprogrammer.org)

LSTMs vs. GRUs

LSTM



GRU



(Source: dprogrammer.org)

Generating Music like Languages

Large Language Models (LLMs)

- The models behind ChatGPT!

SA

You

What's so cool about **AI for music**? Give me a brief answer



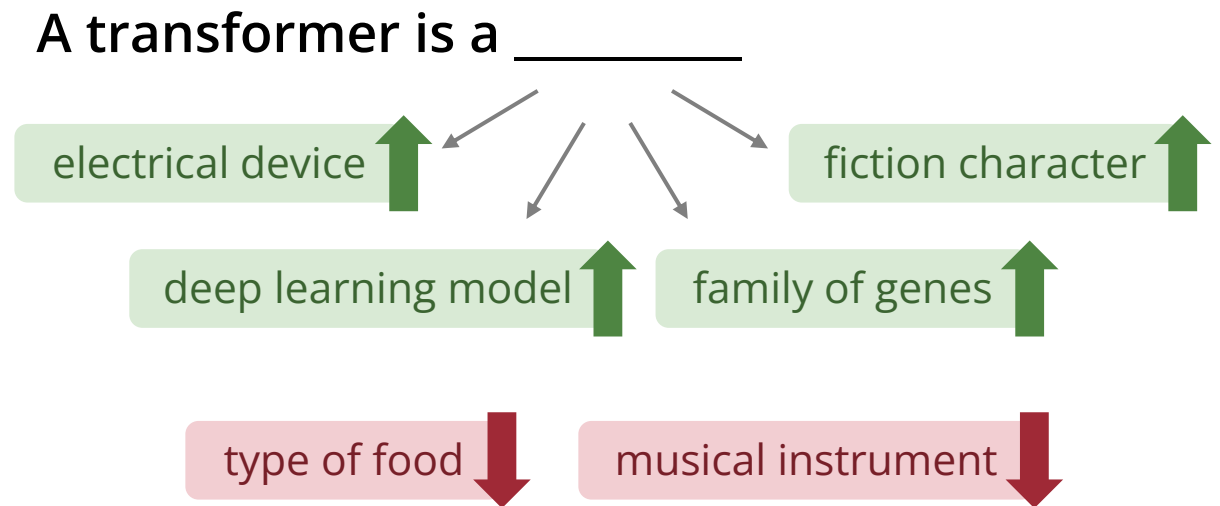
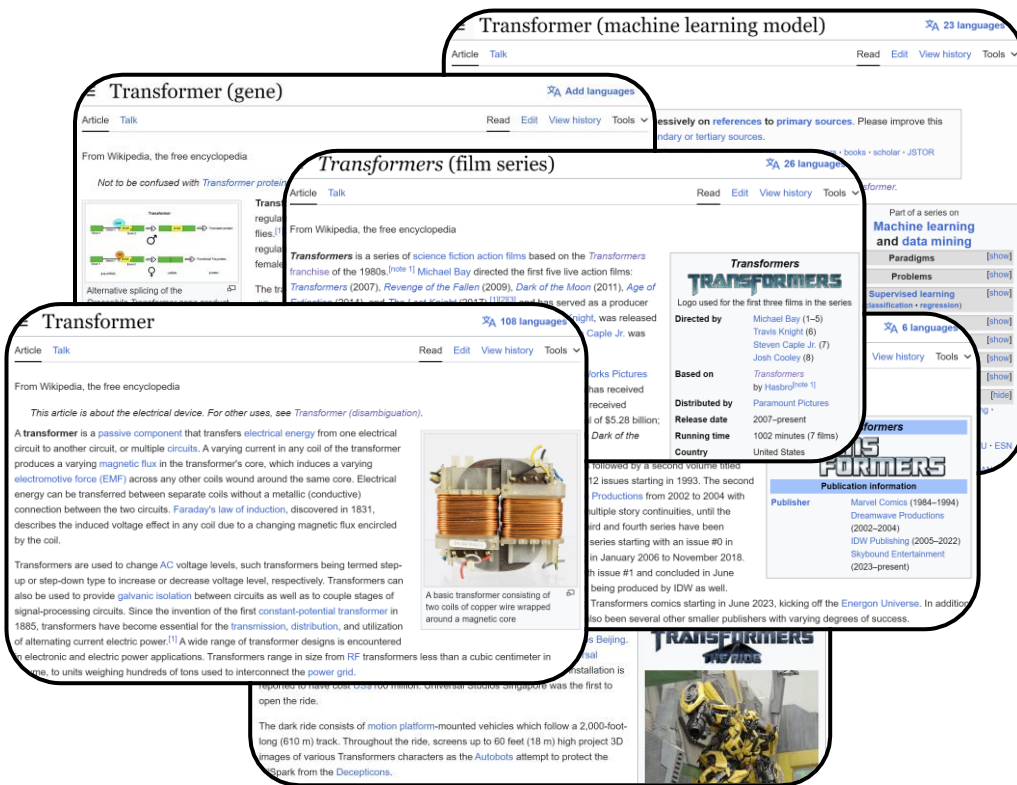
ChatGPT

Word-by-word generation

AI in music is cool because it can compose original pieces, provide personalized recommendations, automate music production tasks, enhance creativity for artists, enable interactive performances, analyze music trends, and even create virtual artists or bands, expanding the possibilities in music creation and enjoyment.

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

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

Next word Previous words

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

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

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

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

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

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

Music Language Models (Mathematically)

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

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

Next note Previous notes

$$\begin{array}{l} P(G \mid C C G G A A) \uparrow \\ P(A \mid C C G G A A) \uparrow \\ P(C \mid C C G G A A) \uparrow \\ P(F \mid C C G G A A) \uparrow \\ P(Ab \mid C C G G A A) \downarrow \\ P(A\# \mid C C G G A A) \downarrow \end{array}$$

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

An Example of ABC Notation

Ah! vous dirai-je, maman
(Twinkle, twinkle, little star)

anon. (France)

The image shows a musical score for the song 'Ah! vous dirai-je, maman' (Twinkle, twinkle, little star). The score is written in treble clef with a common time signature (C). The tempo is marked as ♩ = 120. The score consists of three staves. The first staff has an orange box around the first four notes (C4, D4, E4, F4) and a green box around the fifth note (G4). A red circle highlights the eighth note (A4) in the first staff. Lines connect these boxes and the red circle to the corresponding ABC notation in the right-hand box: the orange box connects to 'CCGG', the green box to 'AAG2', and the red circle to 'AAG2'.

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

*folk***RNN**
generate a folk tune with a recurrent neural network

PRESS TO GENERATE TUNE

Compose

MODEL
thesession.org (w/ :| :)

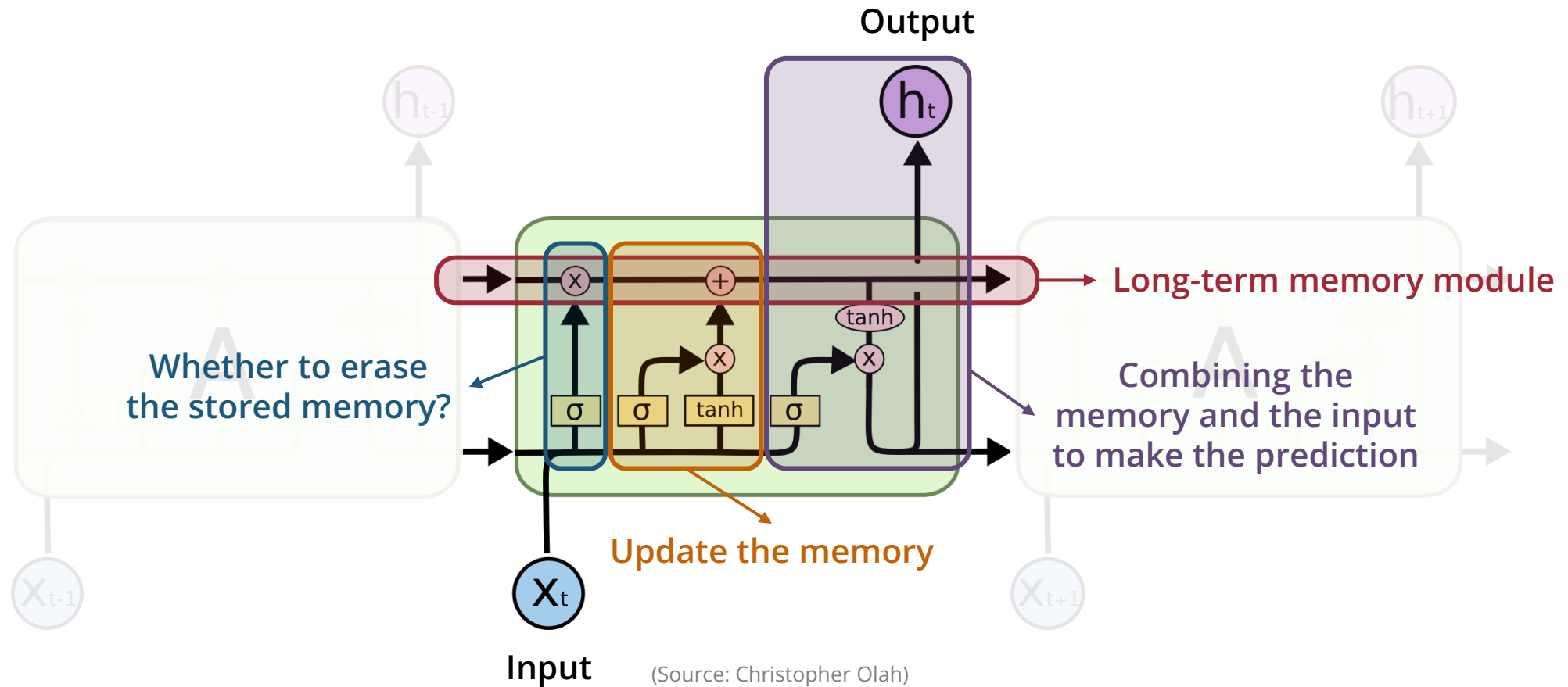
TEMPERATURE SEED
1 62063

METER MODE
4/4 C Major

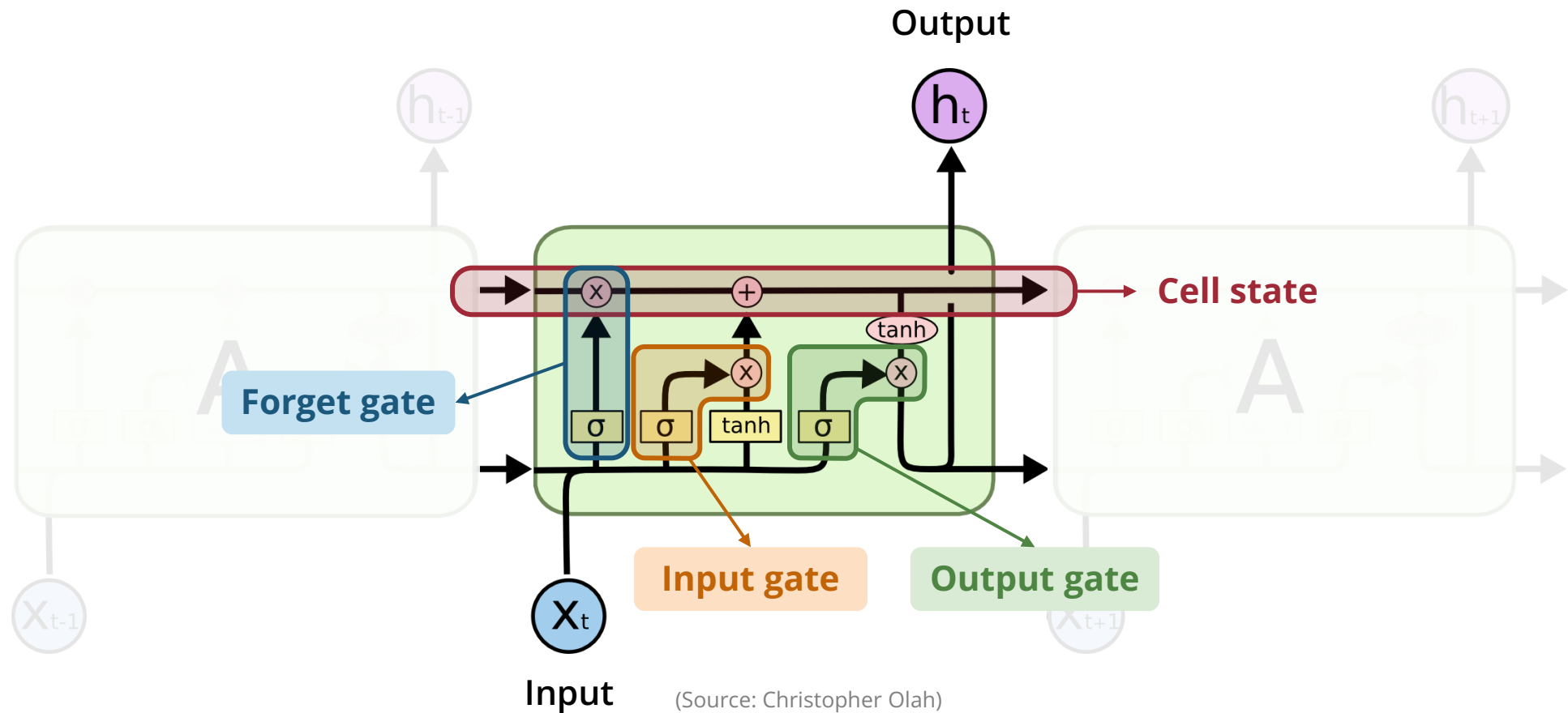
INITIAL ABC
Enter start of tune in ABC notation

folkrrnn.org

Demystifying LSTMs (Hochreiter & Schmidhuber, 1997)



Demystifying LSTMs (Hochreiter & Schmidhuber, 1997)



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**

*folk***RNN**
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

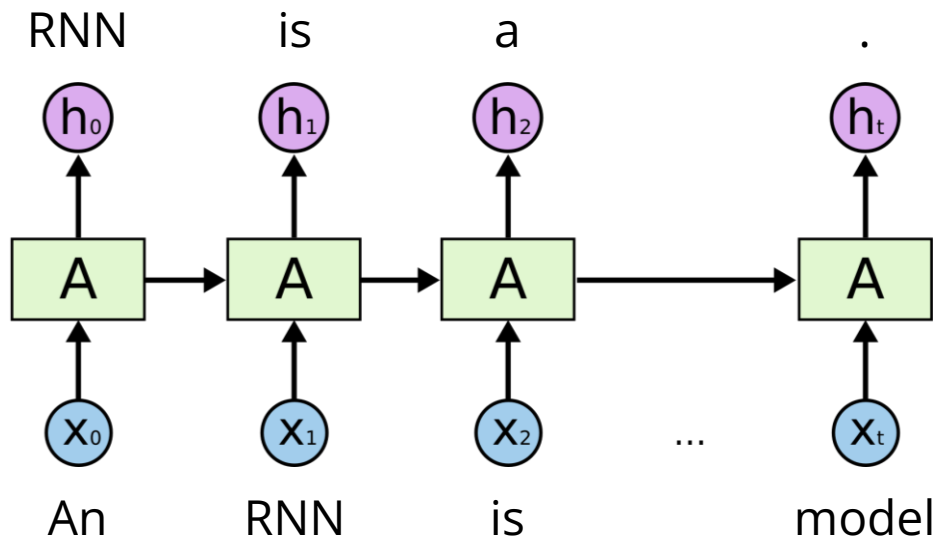
INITIAL ABC
Enter start of tune in ABC notation

folkrrnn.org

Word-level vs Character-level RNNs

Word-level RNNs

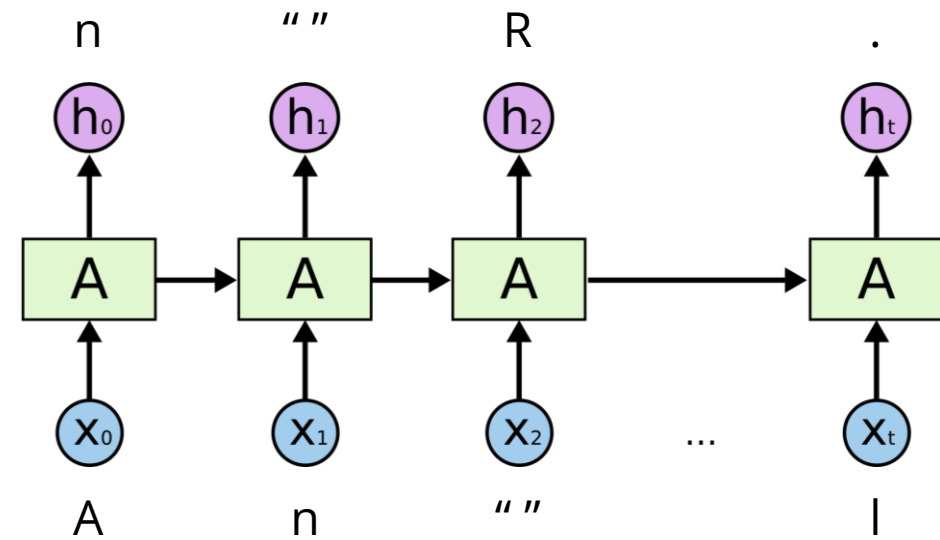
- Predicting word by word
- Most common



(Source: Christopher Olah)

Character-level RNNs

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



(Source: Christopher Olah)

| Limitations of ABC Notations

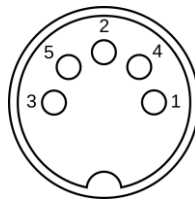
- Limited expressiveness
- Monophonic tunes only

MIDI-like Representation

MIDI (Musical Instrument Digital Interface)



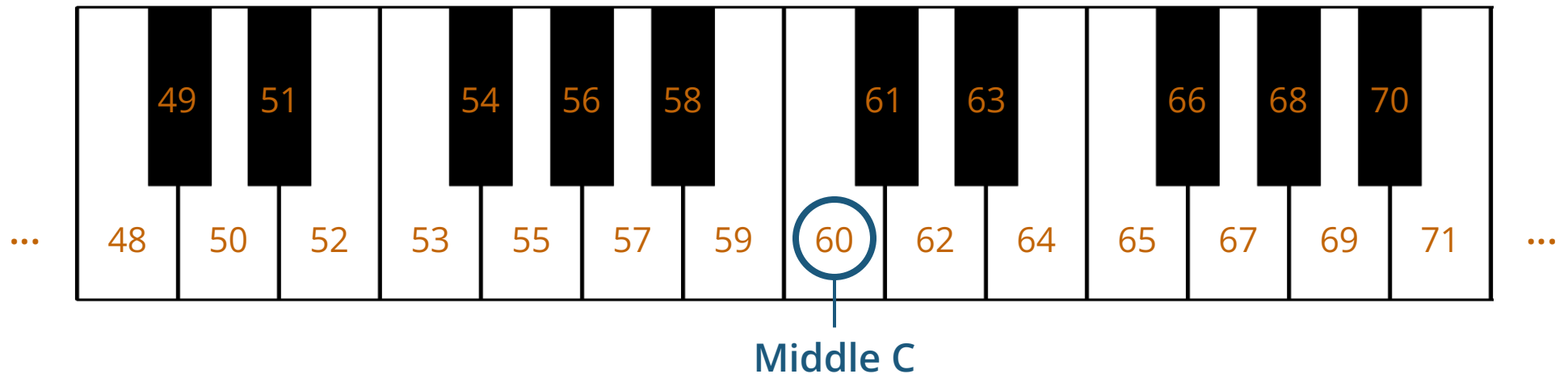
- A communication **protocol** between devices
- MIDI Messages
 - Note on
 - Note off
 - Delta time
 - Program change
 - Control change
 - Pitch bend change



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

Sunshine on the Meadow



The image shows two staves of musical notation in 4/4 time. The first staff has a treble clef and a 4/4 time signature. The first two notes are circled: the first is in a blue circle with an orange 'X' over it, and the second is in a green circle. A red arrow points from the first note to the second, indicating a time shift. The second staff continues the melody.

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

Clair de Lune
from “Suite Bergamasque” L. 75
3rd Movement
Claude Debussy
(1862–1918)

Andante très expressif

Piano

pp *con sordina*

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.

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

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

Next word Previous words

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

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

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

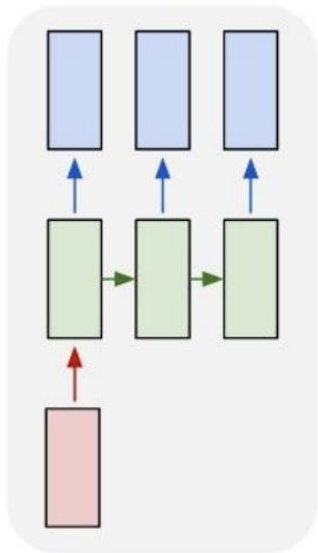
$P(\text{model} \mid \text{A transformer is a})$ ↑

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

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

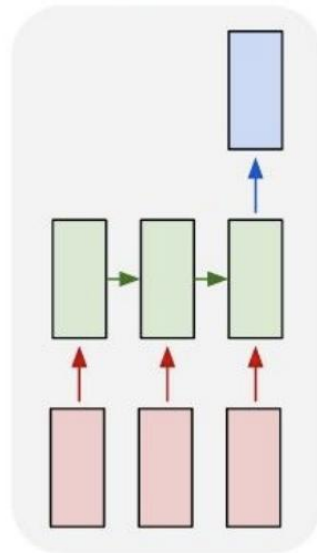
| Types of RNNs

one to many



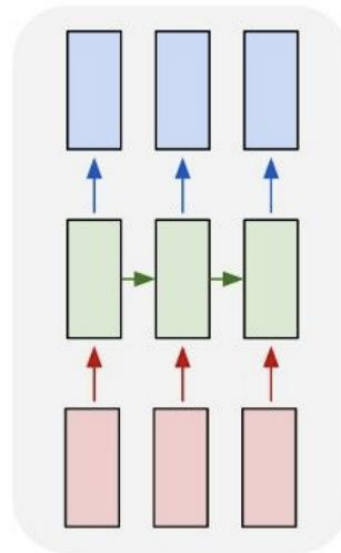
Text generation
Music generation

many to one



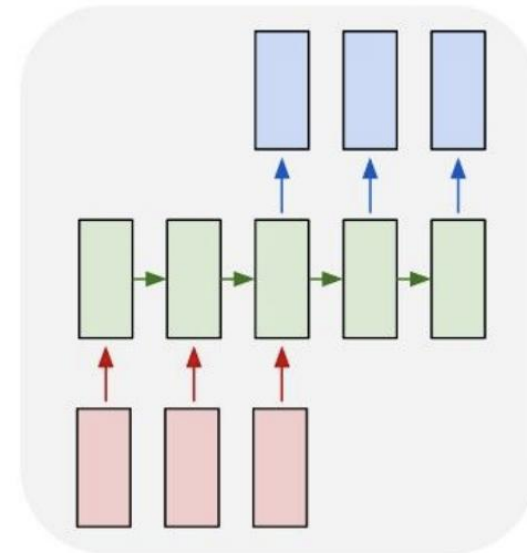
Sentiment classification
Genre classification

many to many



Name entity recognition
Performance rendering

many to many

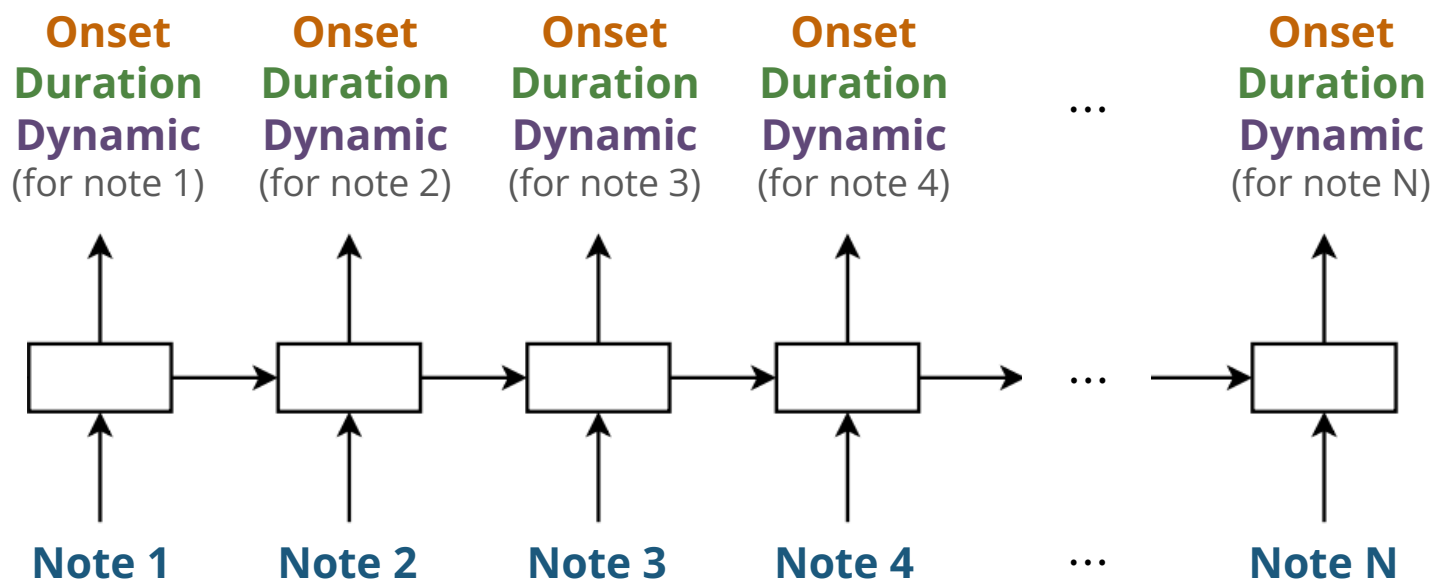


Machine translation
Music accompaniment
Style Transfer

(Source: CS231n)

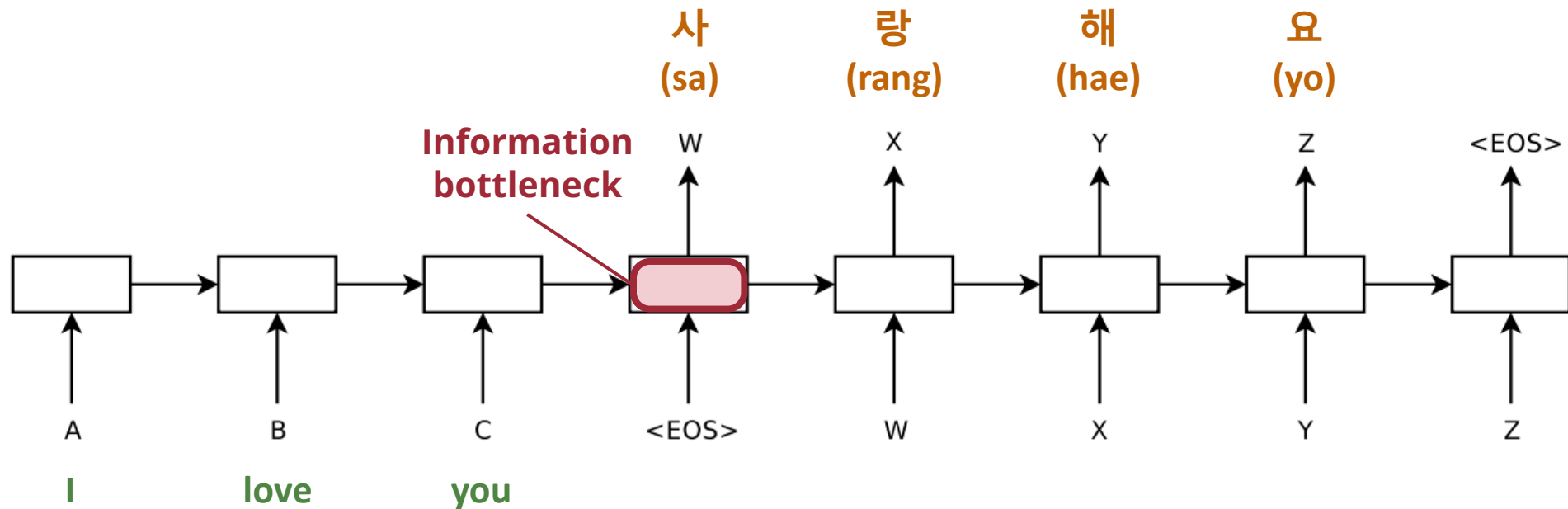
Many-to-Many RNNs for Performance Rendering

- Inputs and outputs are **aligned sequences**



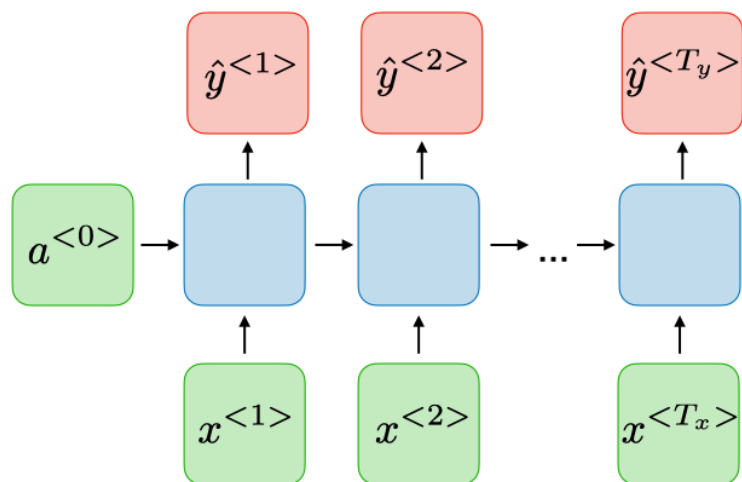
Sequence-to-Sequence Model (Seq2seq)

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



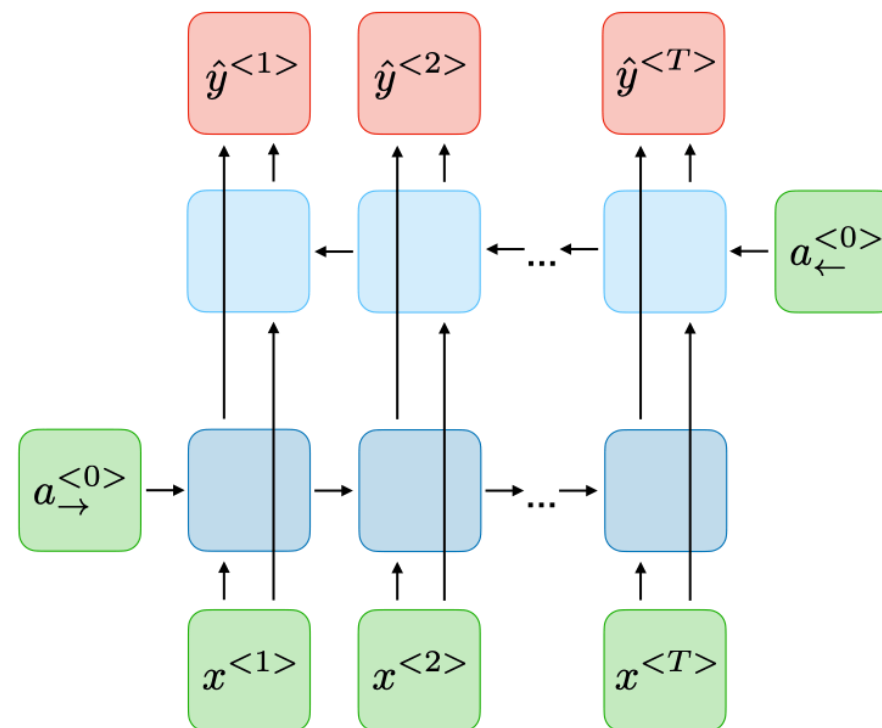
Unidirectional vs Bidirectional RNNs

Unidirectional RNNs



Access to **only past** information

Bidirectional RNNs

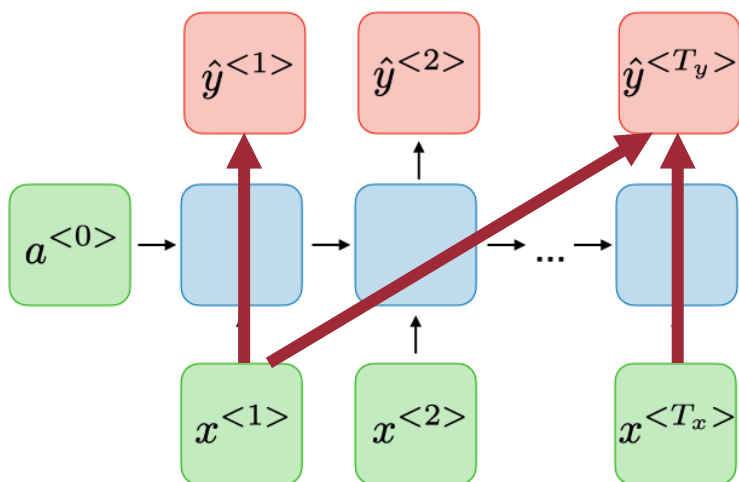


Access to **past & future** information

(Source: Amidi & Amidi, 2019)

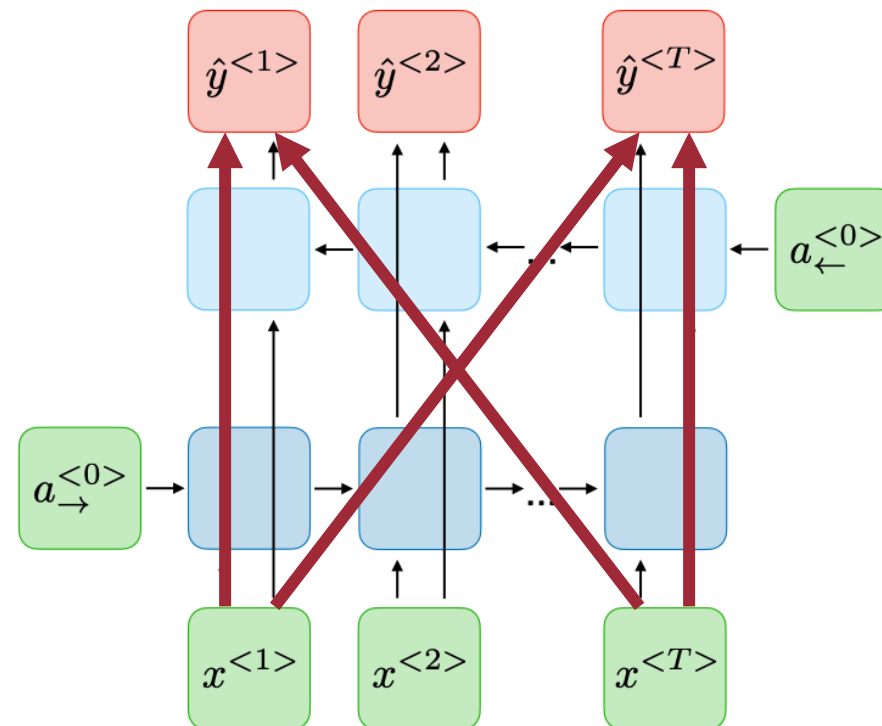
Unidirectional vs Bidirectional RNNs

Unidirectional RNNs



Access to **only past** information

Bidirectional RNNs



Access to **past & future** information

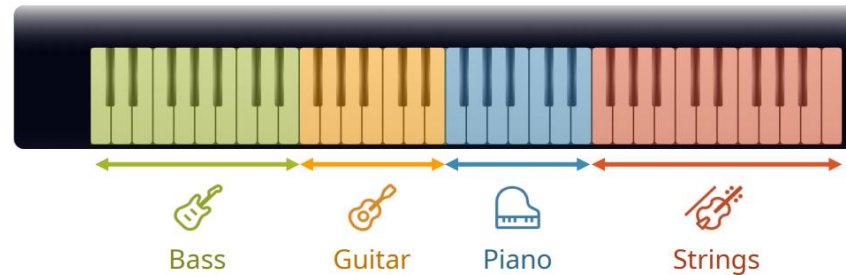
(Source: Amidi & Amidi, 2019)

Automatic Instrumentation

Automatic Instrumentation

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

Intelligent Musical Instruments



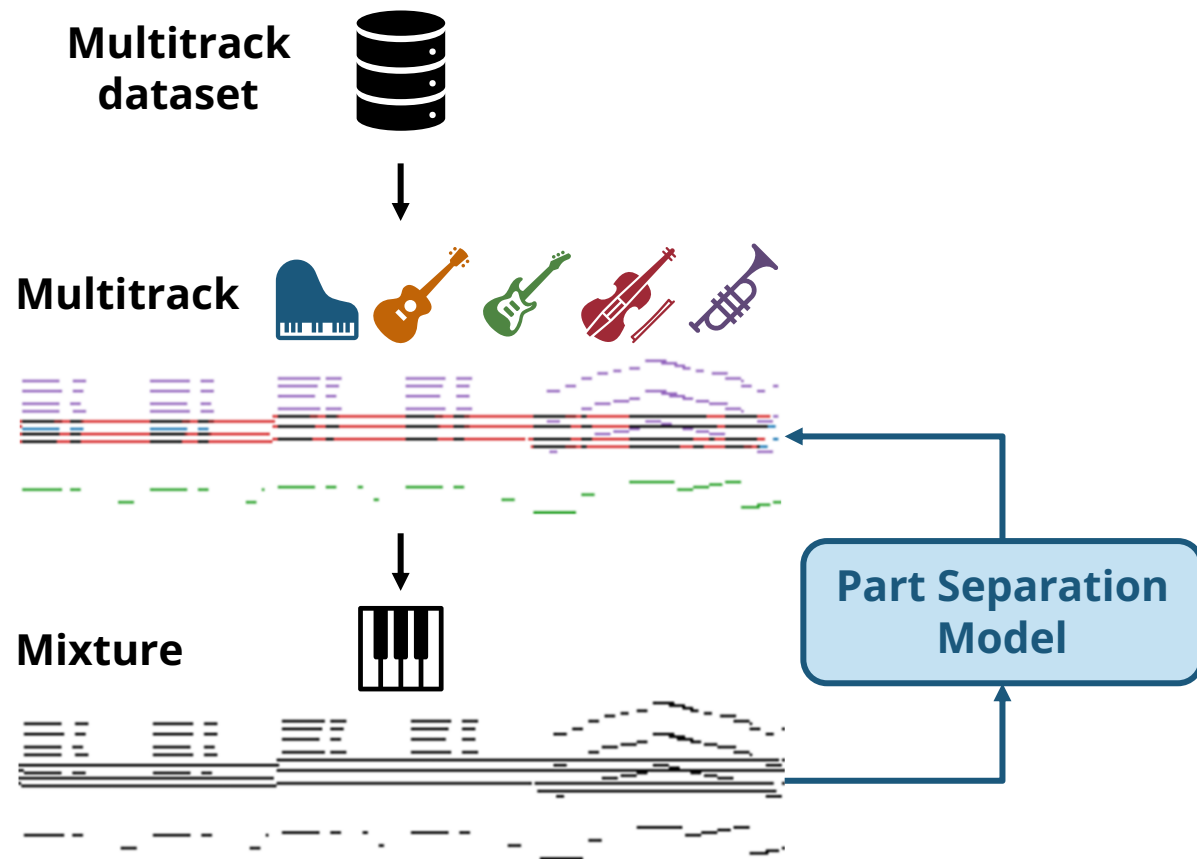
Assistive Composing Tools



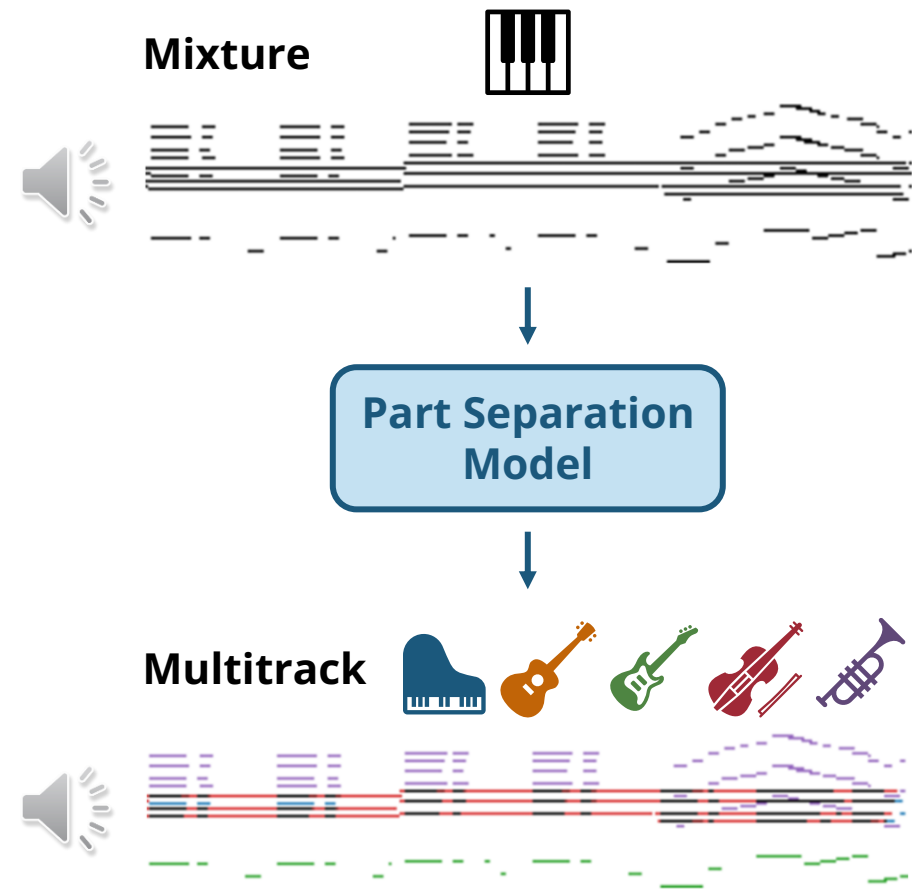
How can we acquire paired data?

Arranger (Dong et al., 2021)

Training

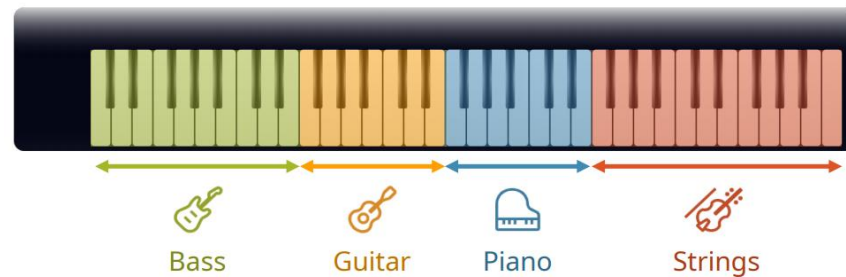


Inference



Online vs. Offline Model

Intelligent Musical Instruments



Can only look at the **past**

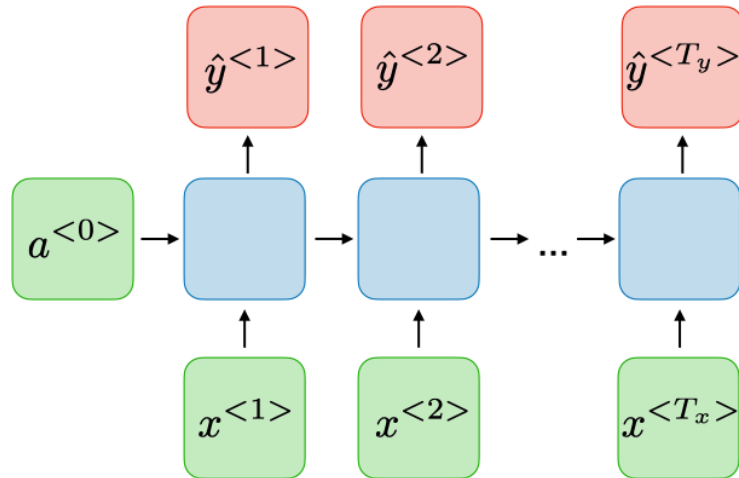
Assistive Composing Tools



Can look at both the **future** and the **past**

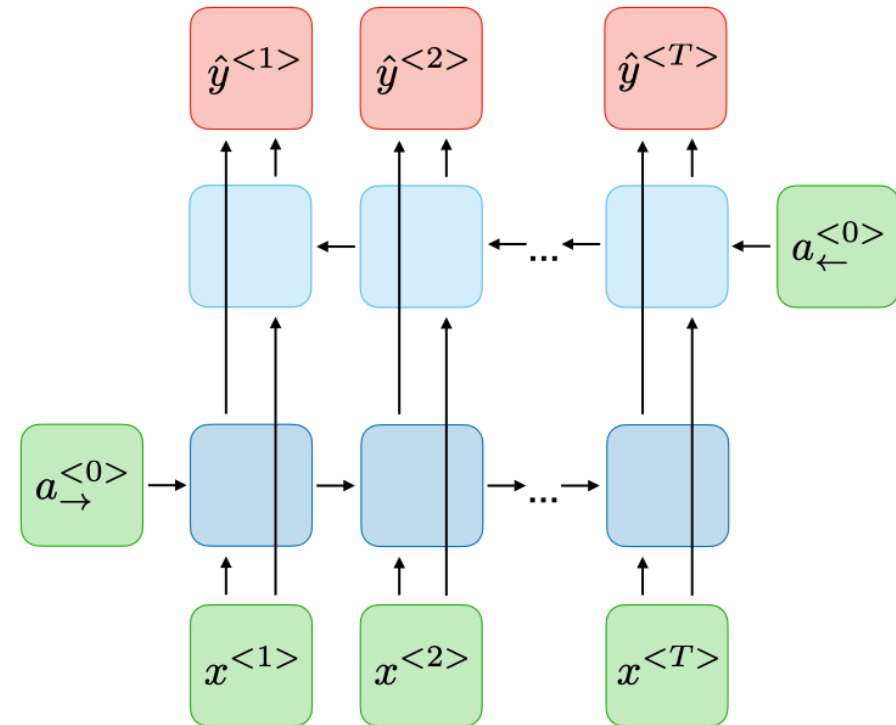
LSTMs vs. BiLSTMs

LSTMs



Access to **only past** information

BiLSTMs



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)

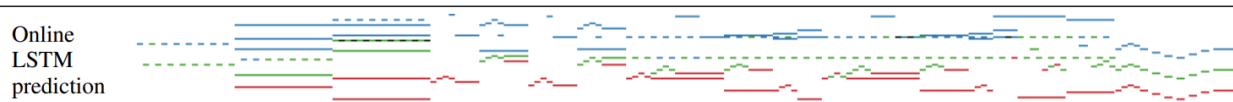
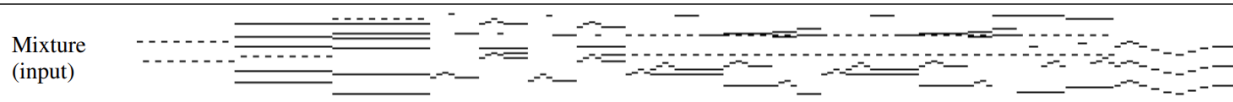
Arranger (Dong et al., 2021)

Bach Chorales



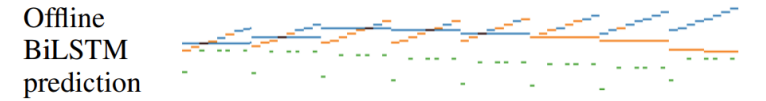
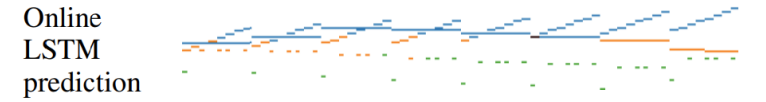
(Audio available.¹ Colors: soprano, alto, tenor, bass.)

String Quartets



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

Game Music



(Audio available.¹ Colors: pulse wave I, pulse wave II, triangle wave.)

Pop Music



(Audio available.¹ Colors: piano, guitar, bass, strings, brass.)

(Source: Dong et al., 2021)

Recap

Deep Autoregressive Models

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

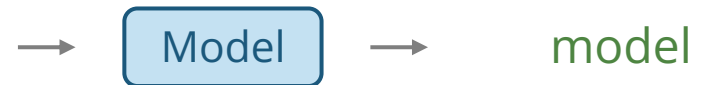
A transformer is a



A transformer is a deep



A transformer is a deep learning



A transformer is a deep learning model



A transformer is a deep learning model introduced



A transformer is a deep learning model introduced in



Deep Autoregressive Models

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

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

Next word

The whole sentence

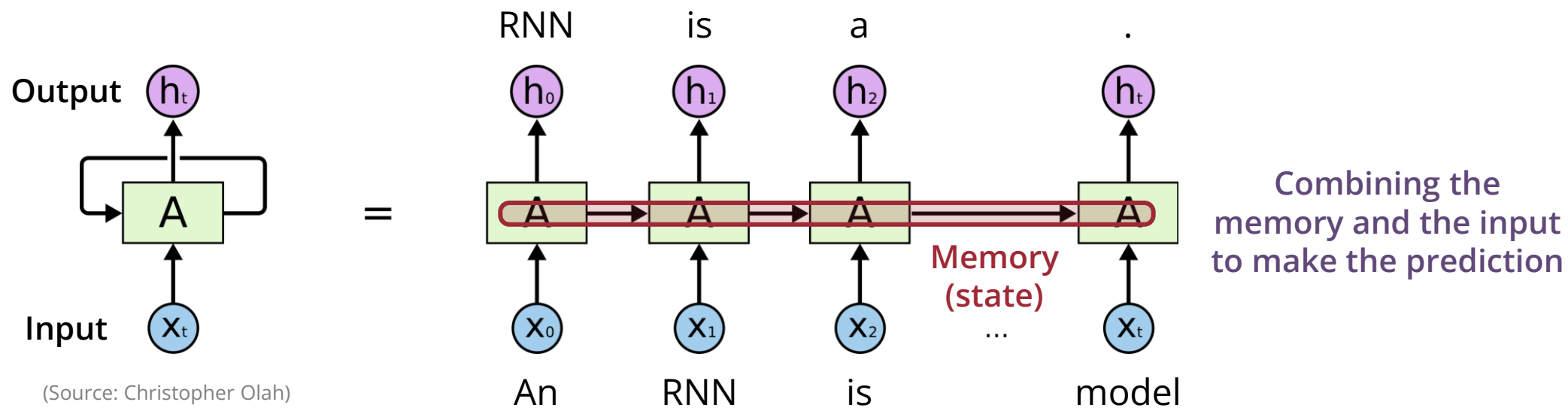
$$X = (x_0, x_1, \dots, x_N)$$

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

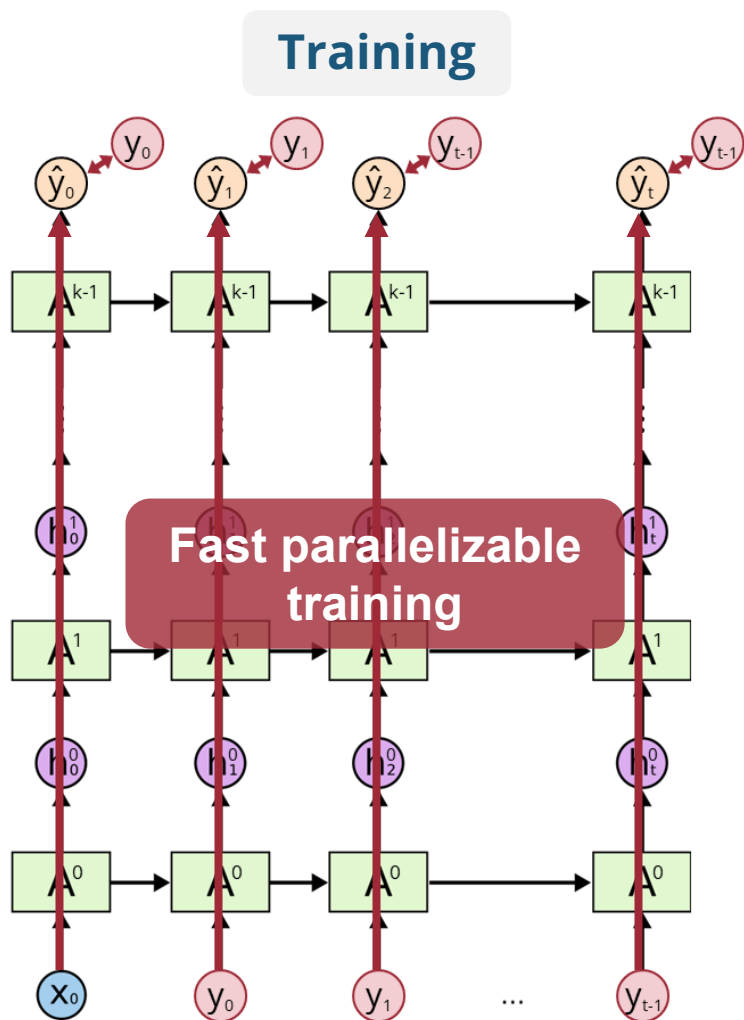
1st word 2nd word given 1st word 3rd word given 1st & 2nd words Last word given all previous words

What is an RNN (Recurrent Neural Network)?

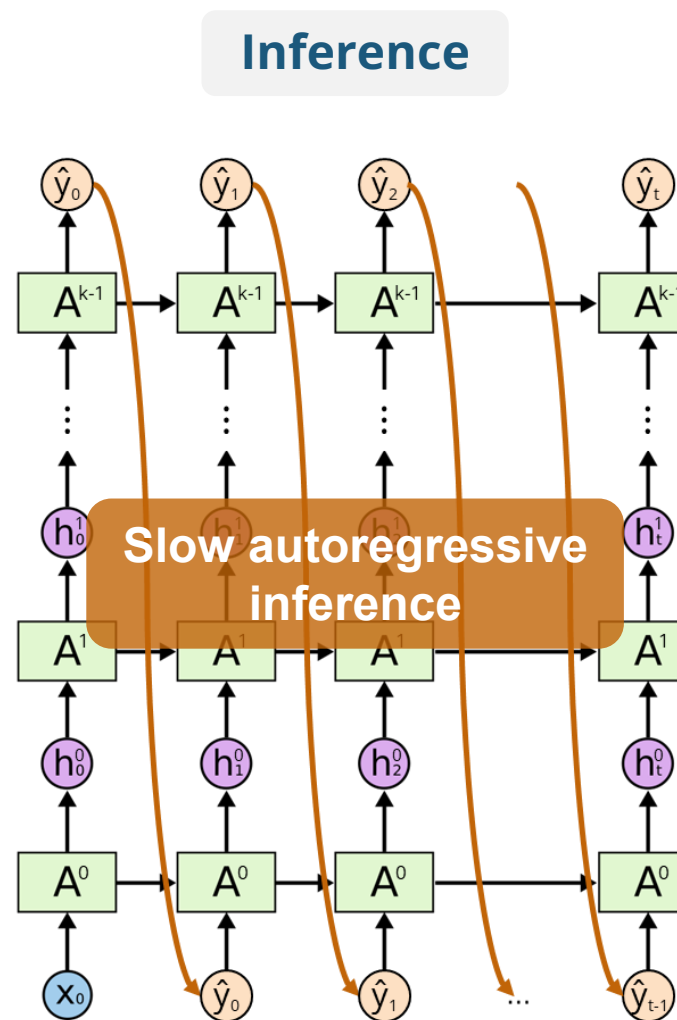
- A type of neural networks that have **loops**
- Widely used for **modeling sequences** (e.g., in natural language processing)



Vanilla RNNs: Training vs. Inference

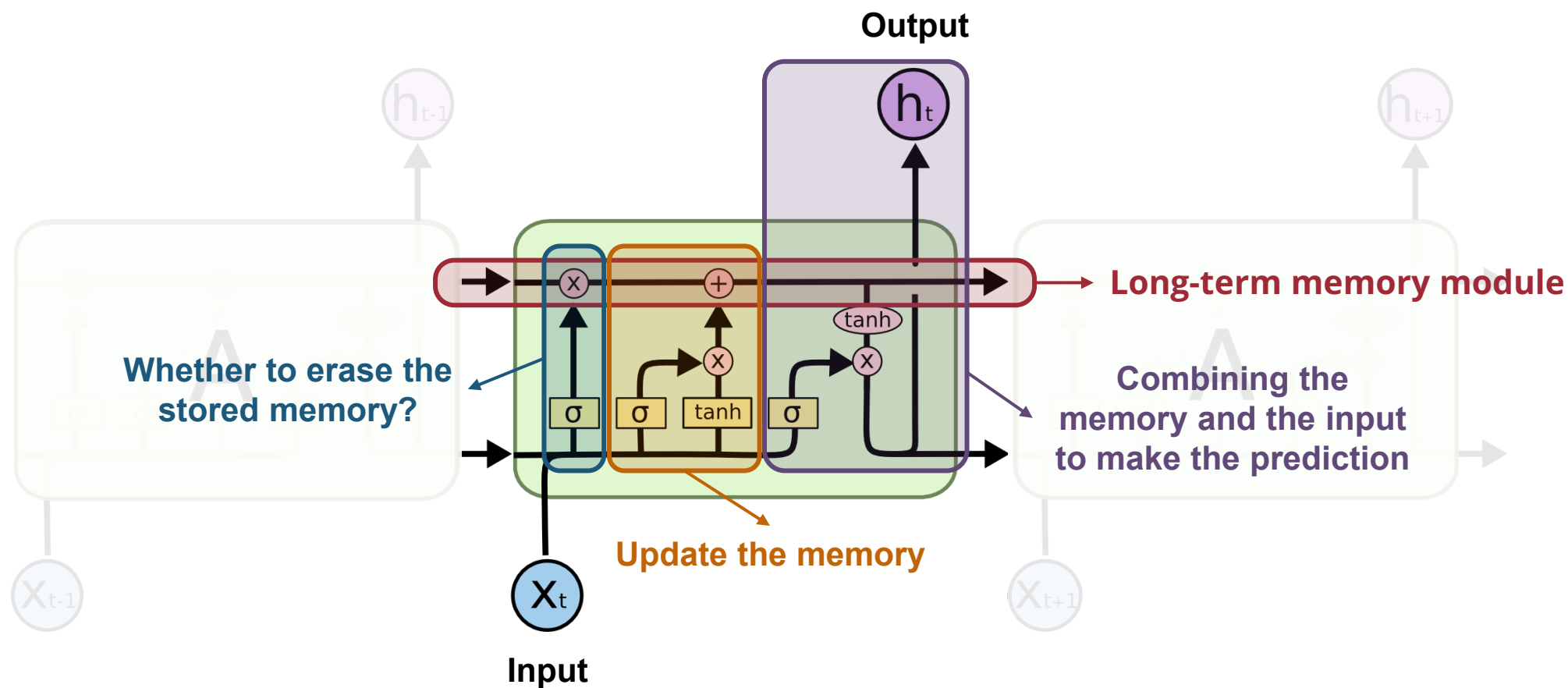


(Source: Christopher Olah)



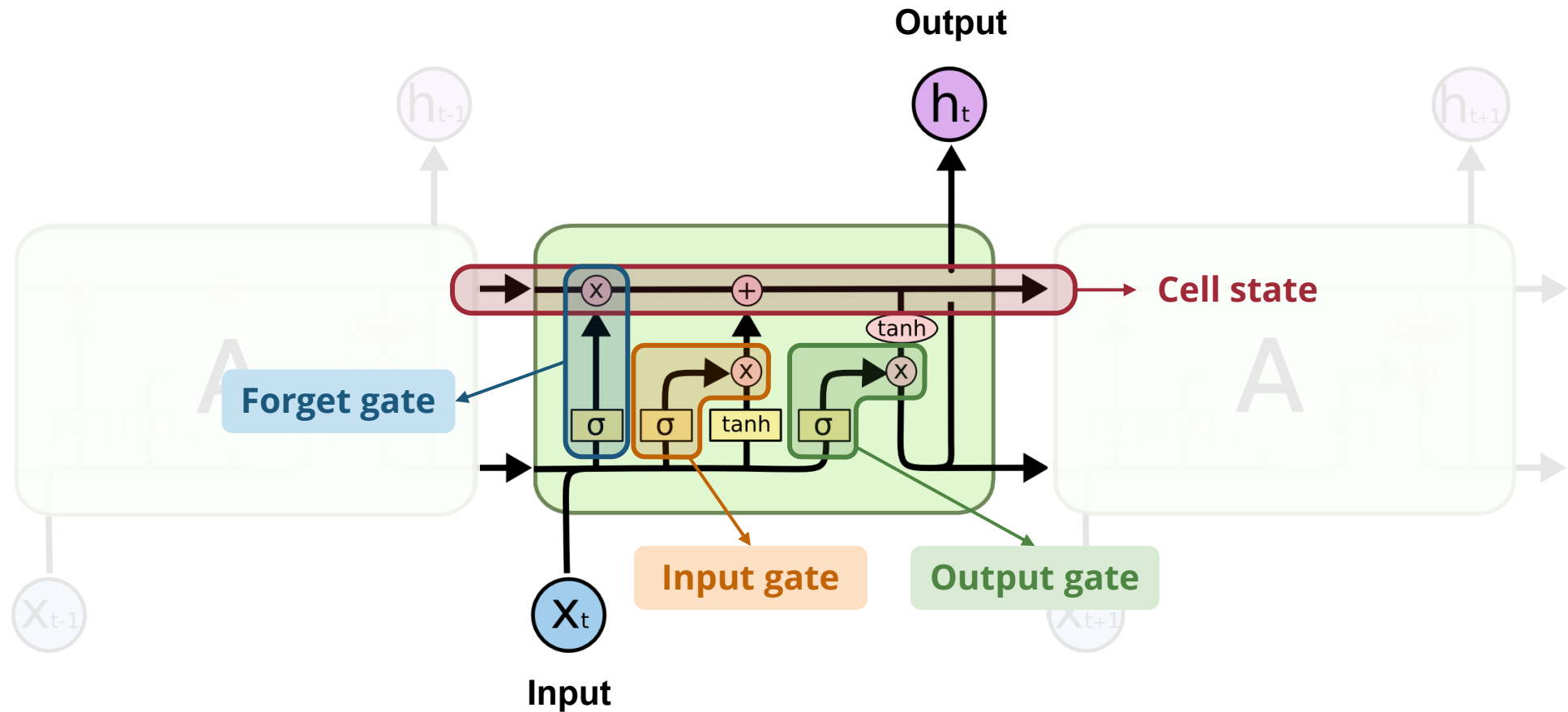
(Source: Christopher Olah)

Demystifying LSTMs (Hochreiter & Schmidhuber, 1997)



(Source: Christopher Olah)

Demystifying LSTMs (Hochreiter & Schmidhuber, 1997)



(Source: Christopher Olah)

Music Language Models (Mathematically)

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

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

Next note Previous notes

$$\begin{array}{l} P(G \mid C C G G A A) \uparrow \\ P(A \mid C C G G A A) \uparrow \\ P(C \mid C C G G A A) \uparrow \\ P(F \mid C C G G A A) \uparrow \\ P(Ab \mid C C G G A A) \downarrow \\ P(A\# \mid C C G G A A) \downarrow \end{array}$$

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

*folk***RNN**
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

folkrrnn.org

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

Clair de Lune
from “Suite Bergamasque” L. 75
3rd Movement
Claude Debussy
(1862–1918)

Andante très expressif

Piano

pp *con sordina*

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)



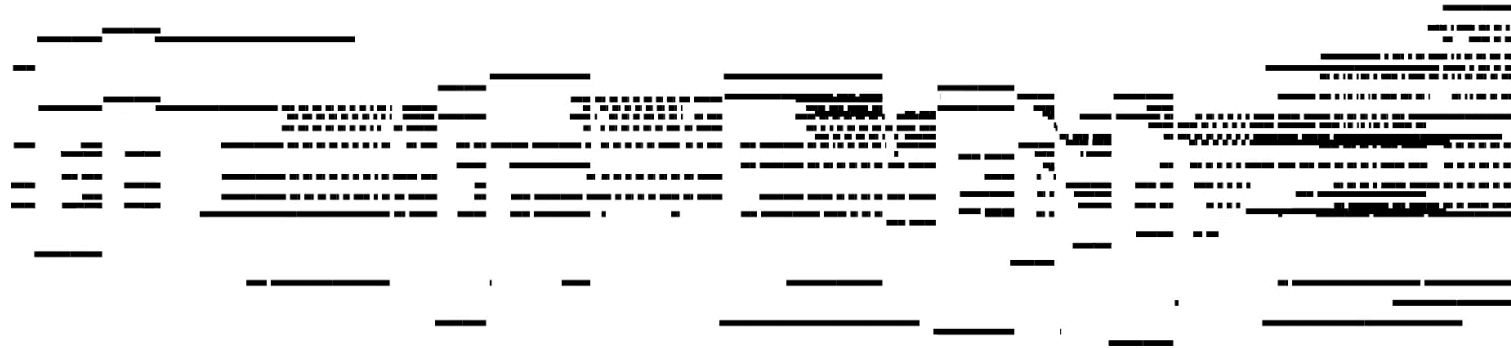
BiLSTM
(w/ entry hints)



(Source: Dong et al., 2021)

Next Lecture

Transformers



(Source: Huang et al., 2018)