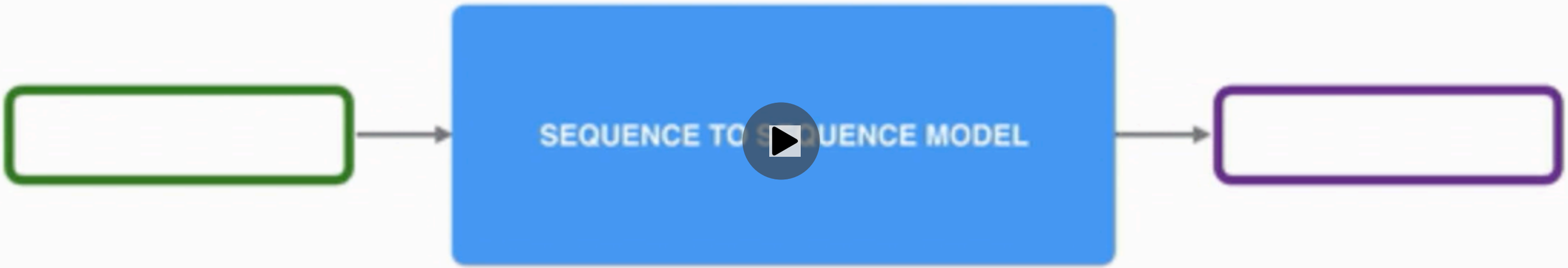
How Transformers Work

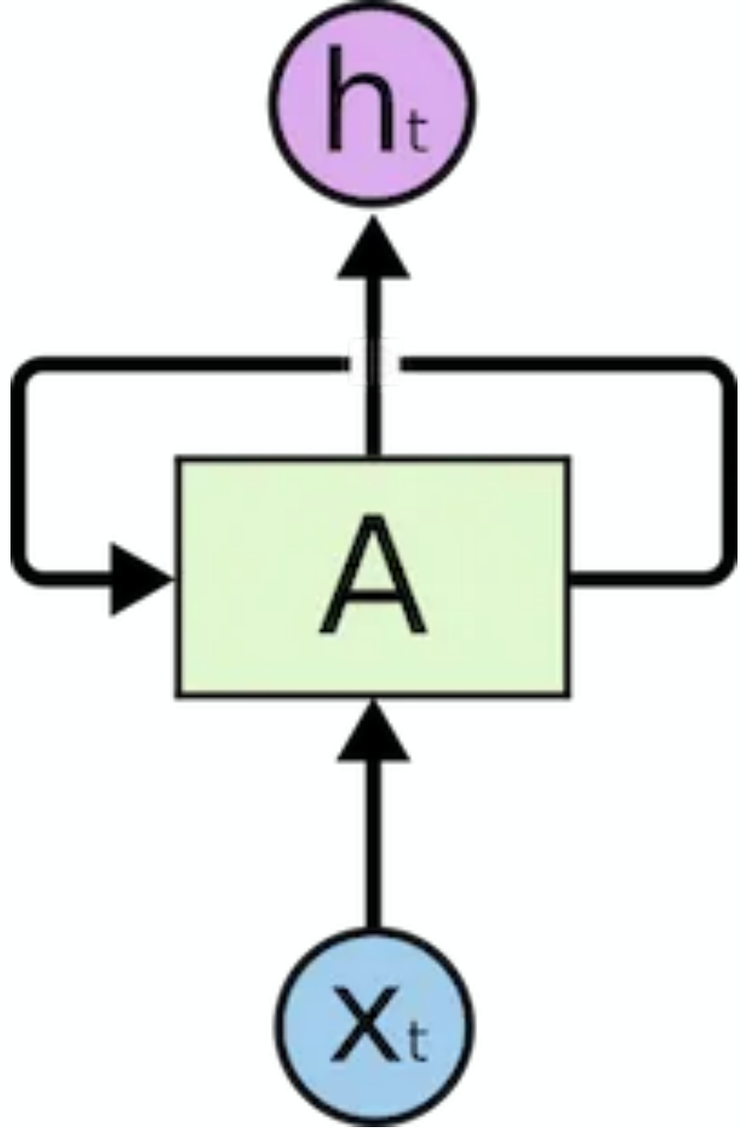
The Neural Network used by OpenAI and DeepMind



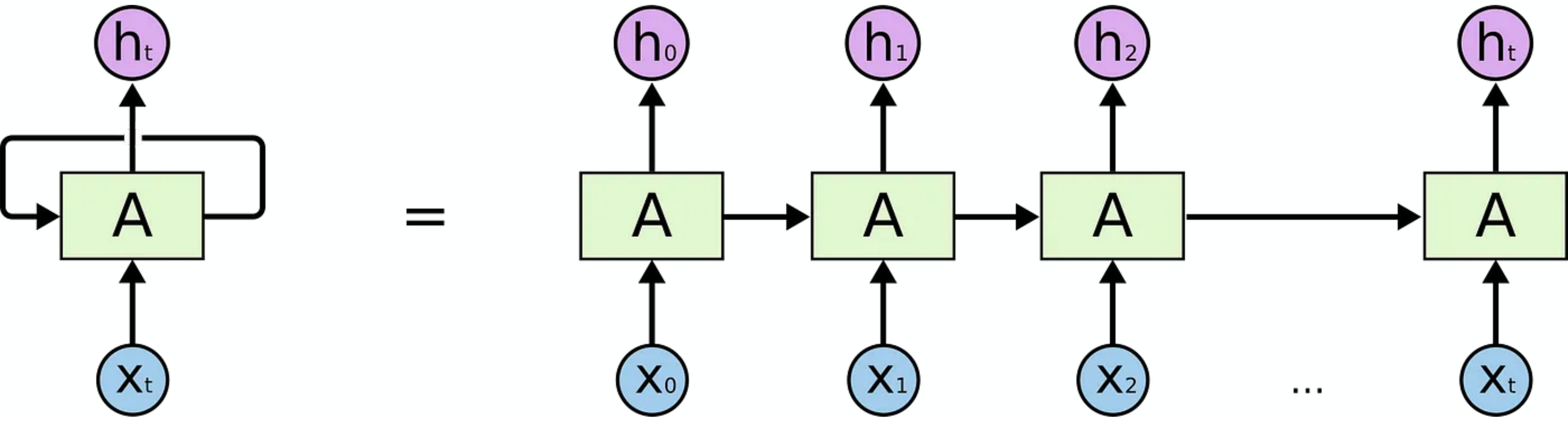
This image represents sequence transduction. The input is represented in green, the model is represented in blue, and the output is represented in purple.

The Transformers are a Japanese [[hardcore punk]] band. The band was formed in 1968, during the height of Japanese music history

Recurrent Neural Networks



The input is represented as x_t



An unrolled recurrent neural network

chine Translation

EQUENCE MODEL

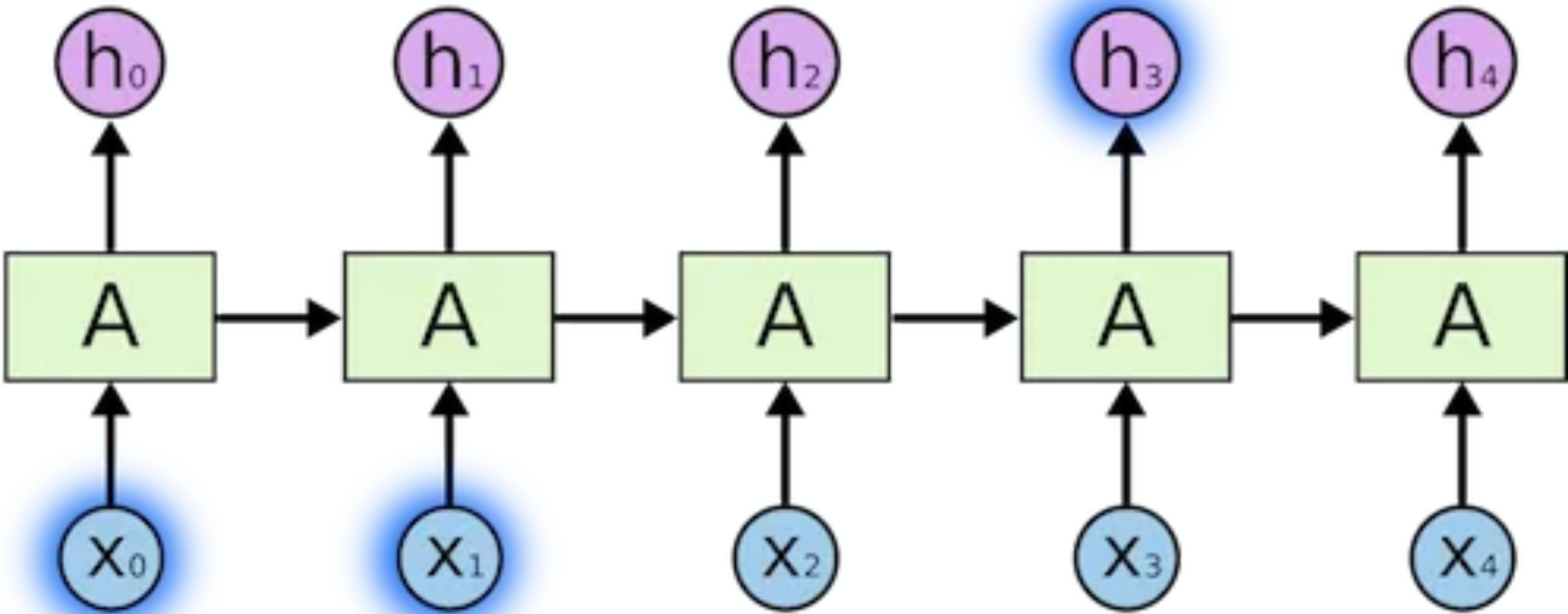
Encoding Stage

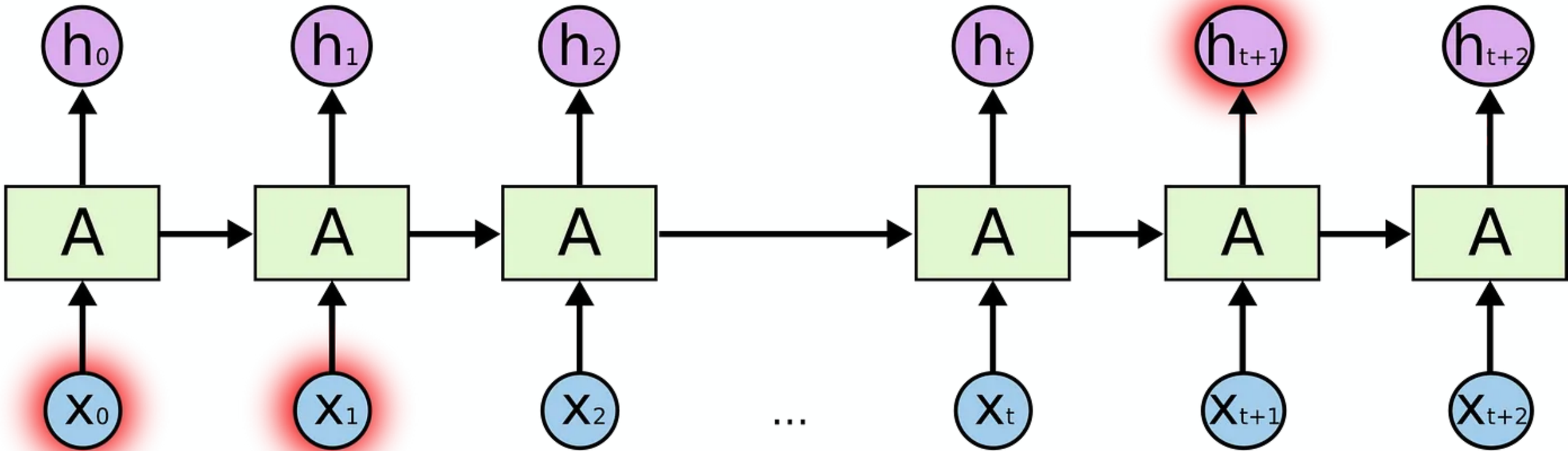
Decoding Sta

suis

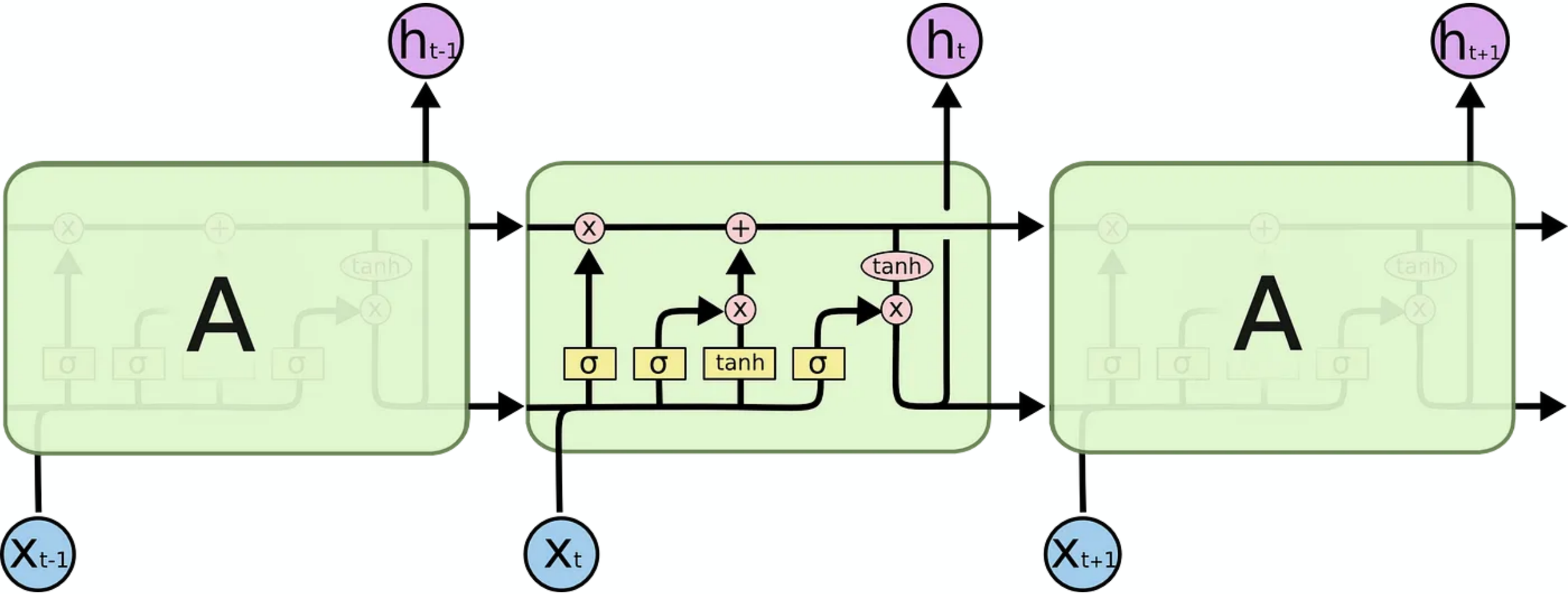
étudiant

The Problem of Long-Term Dependencies


h0
h1
h2
h3
h4
A
A
A
A
A
X0
X1
X2
X3
X4


h0
h1
h2
ht
ht+1
ht+2
A
A
A
A
A
A
x0
x1
x2
...
xt
xt+1
xt+2

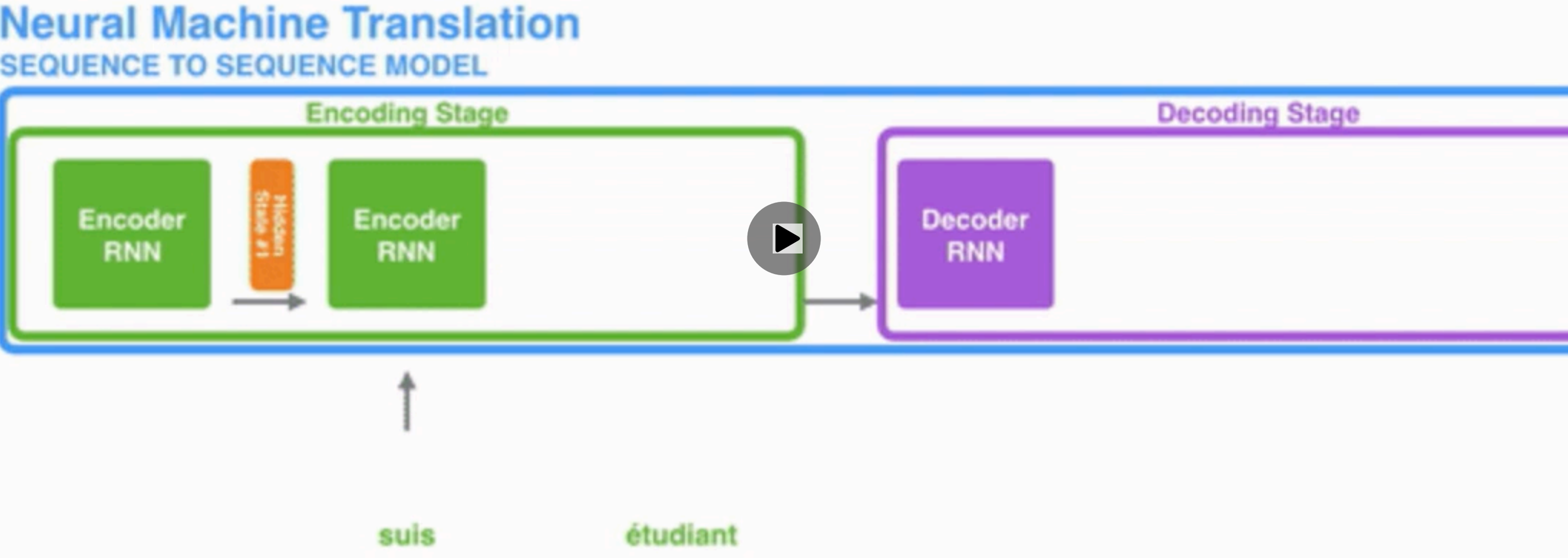
Long-Short Term Memory (LSTM)


h_{t-1}
h_t
h_{t+1}
A
A
A
tanh
tanh
σ
σ
tanh
σ
X_{t-1}
X_t
X_{t+1}

The Problem with LSTMs

- Sequential computation inhibits parallelization
- No explicit modeling of long and short range dependencies
- “Distance” between positions is linear

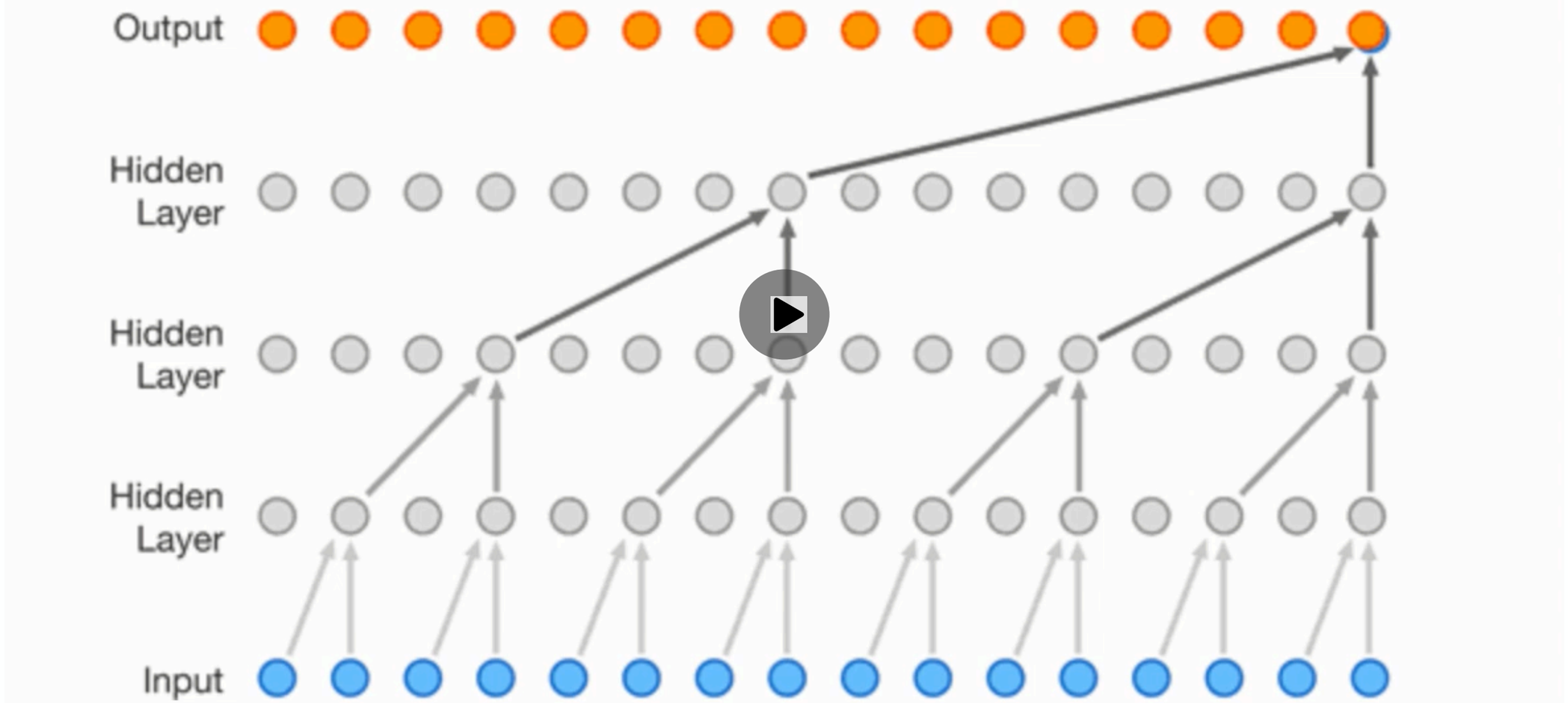
Attention



The green step is called the **encoding stage** and the purple step is the **decoding stage.**

Convolutional Neural Networks

- Trivial to parallelize (per layer)
- Exploits local dependencies
- Distance between positions is logarithmic



Wavenet, model is a Convolutional Neural Network (CNN).

Transformers

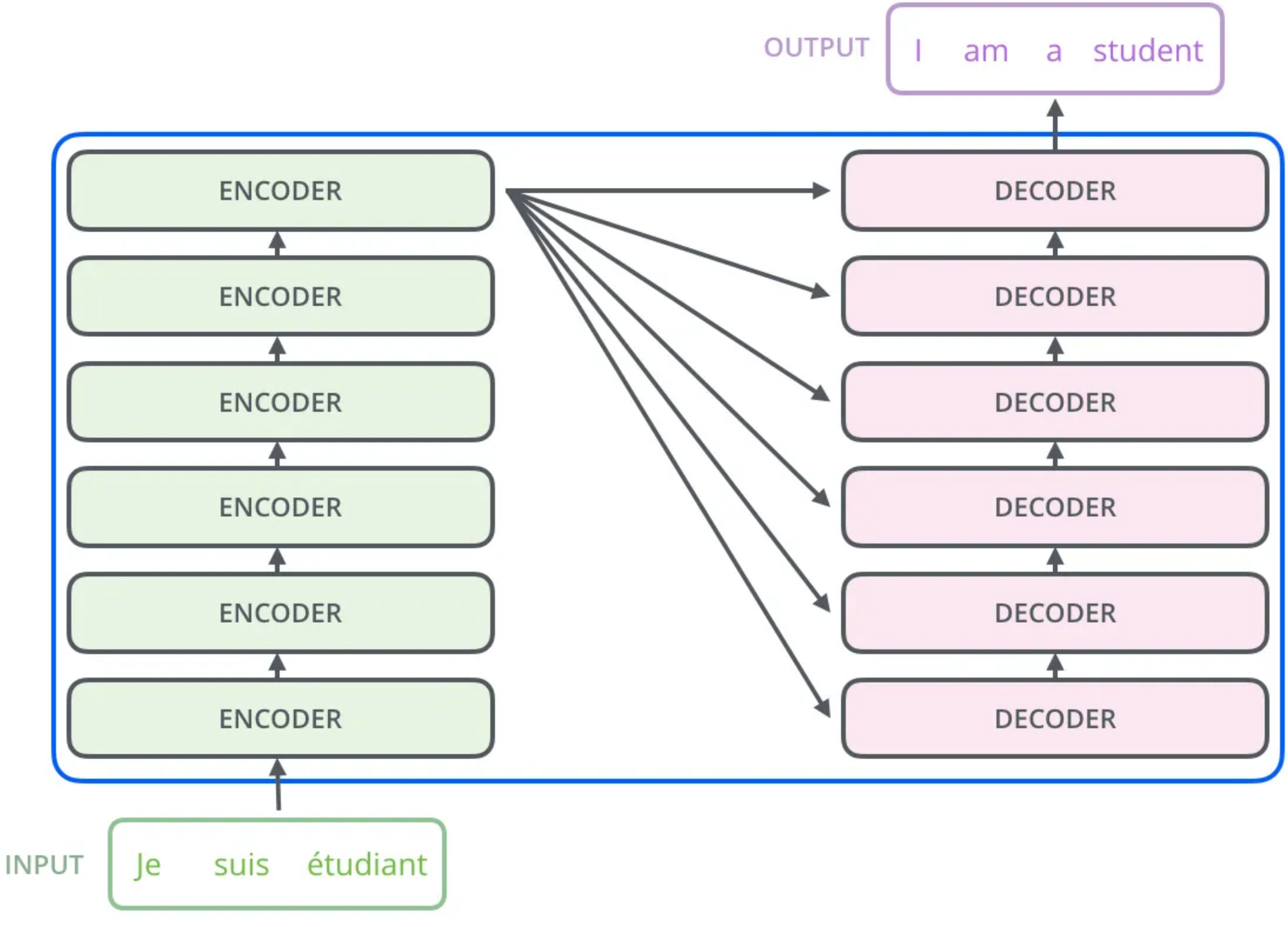
INPUT

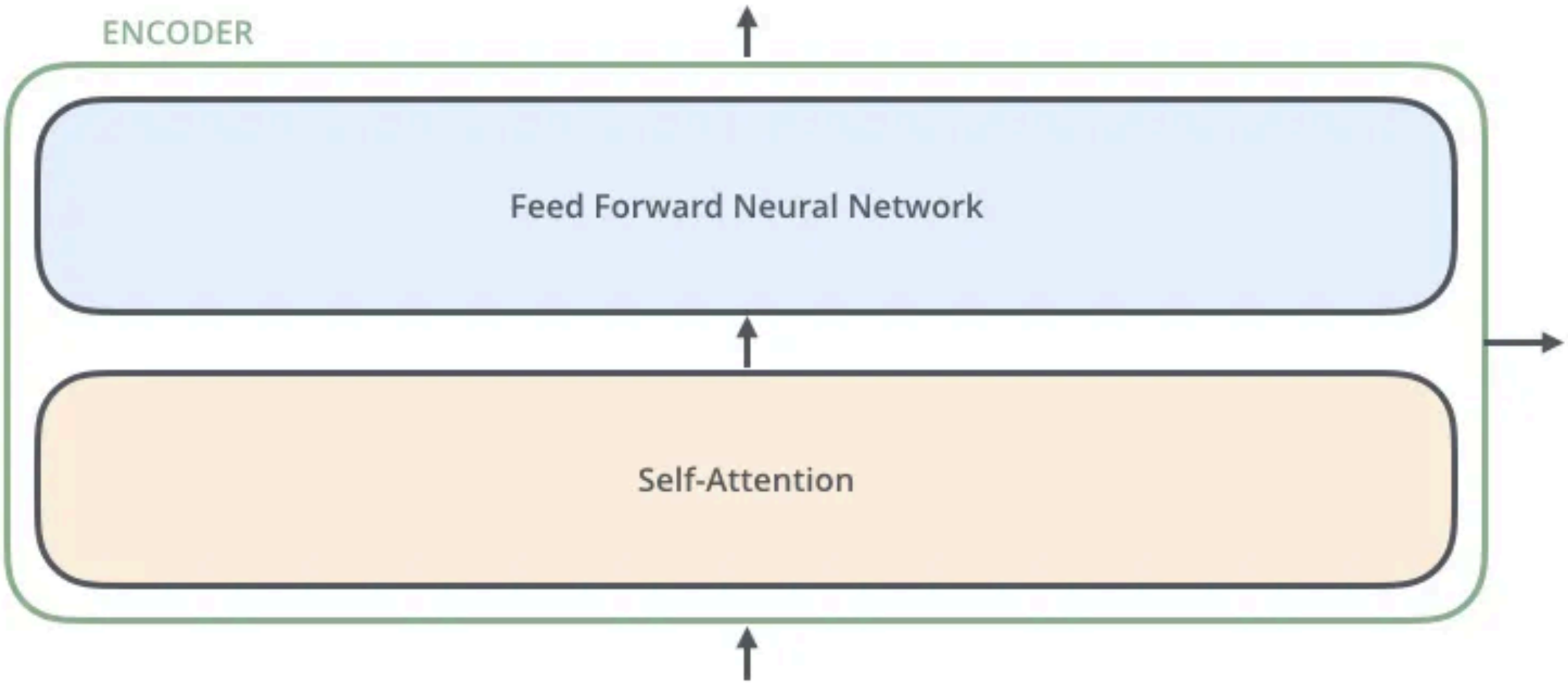
Je suis étudiant

THE TRANSFORMER

OUTPUT

I am a student


OUTPUT
I am a student
ENCODER
ENCODER
ENCODER
ENCODER
ENCODER
ENCODER
DECODER
DECODER
DECODER
DECODER
DECODER
DECODER
INPUT
Je suis étudiant


ENCODER
Feed Forward Neural Network
Self-Attention

DECODER

Feed Forward

Encoder-Decoder Attention

Self-Attention

Positional Encoding

Bibliography

1. The Unreasonable Effectiveness of Recurrent Neural Networks
2. Understanding LSTM Networks
3. Visualizing A Neural Machine Translation Model
4. The Illustrated Transformer
5. The Transformer — Attention is all you need
6. The Annotated Transformer
7. Attention is all you need attentional neural network models
8. Self-Attention For Generative Models
9. OpenAI GPT-2: Understanding Language Generation through Visualization
10. WaveNet: A Generative Model for Raw Audio
11. How Transformers Work