Large Models – COMP4423 Computer Vision

Xiaoyong Wei (魏驍勇)

x1wei@polyu.edu.hk

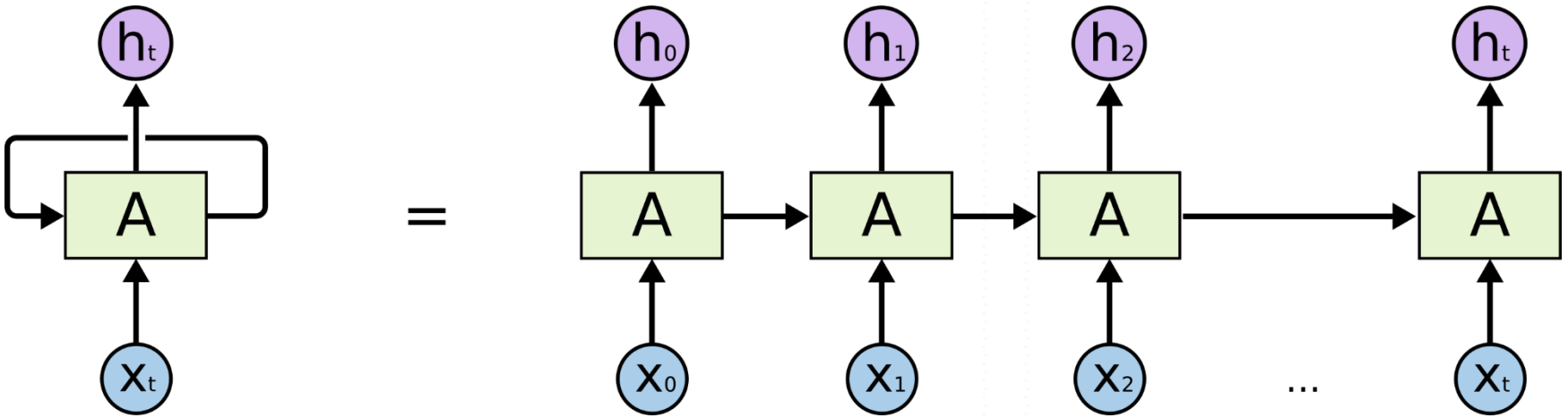
Department of Computing
電子計算學系

The Hong Kong Polytechnic University
香港理工大學

Opening Minds • Shaping the Future
啟迪思維 • 成就未來

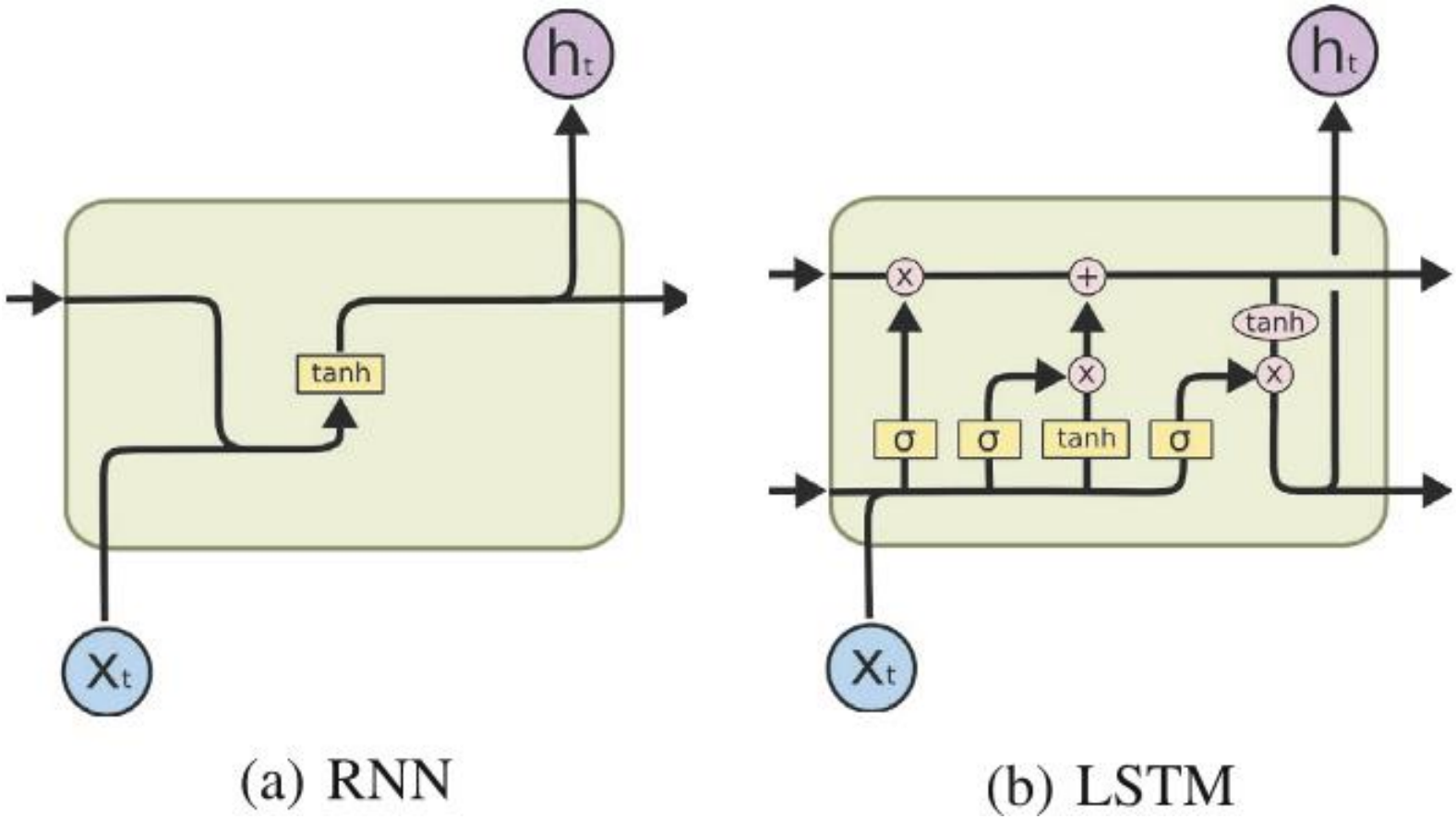
RNN and Image Captioning

Recurrent Neural Networks



Source: Ming Li's Lecture notes

RNN vs. LSTM

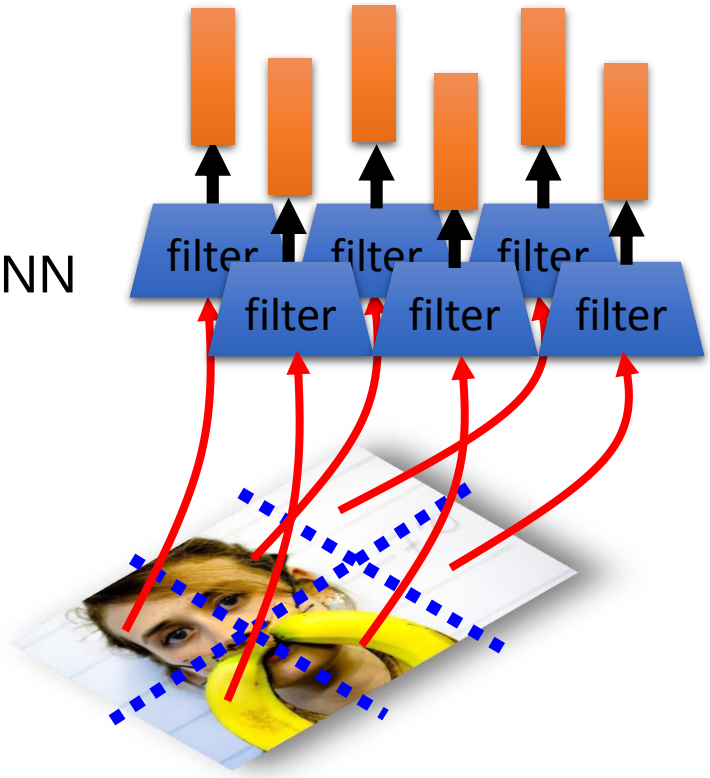


Source: Ming Li's Lecture notes

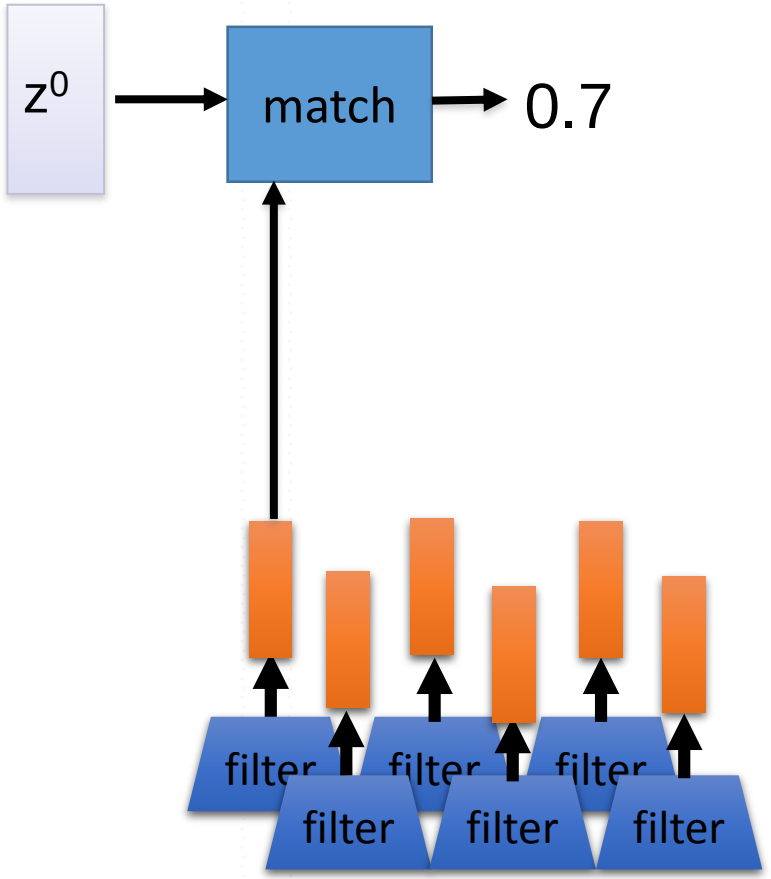
Image Captioning

A vector for each region

CNN

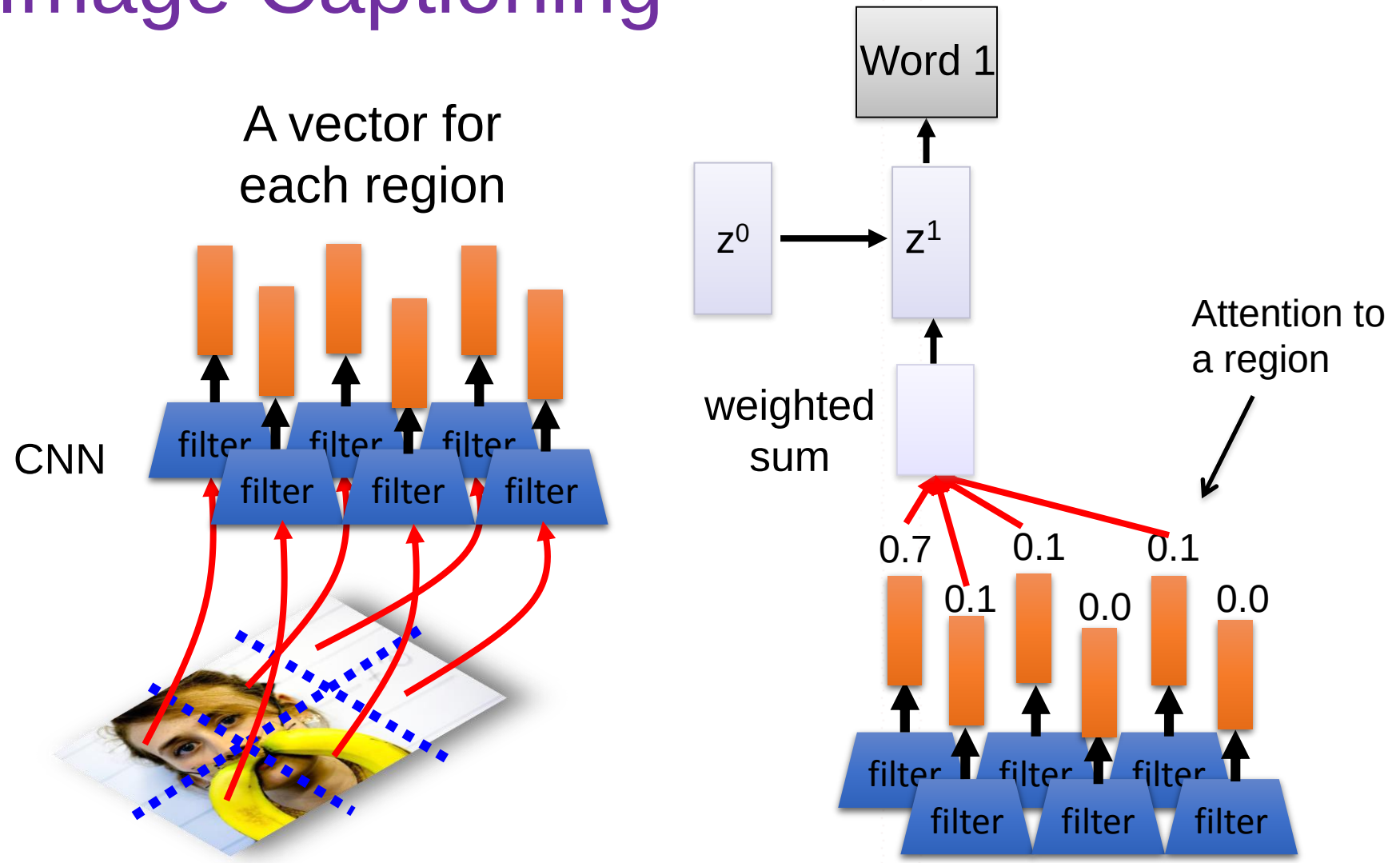


z^0 is initial parameter, it is also learned



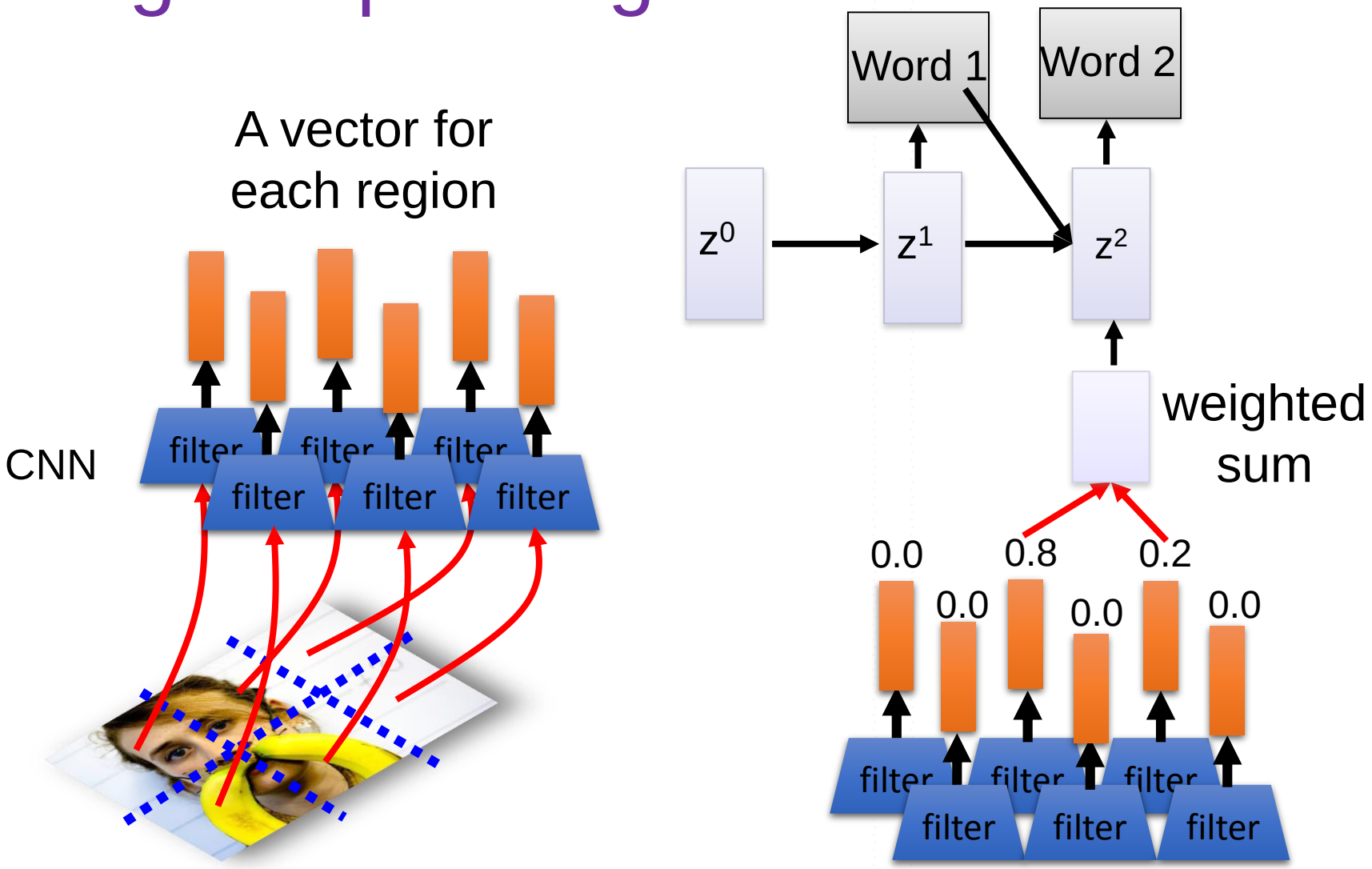
Source: Ming Li's Lecture notes

Image Captioning



Source: Ming Li's Lecture notes

Image Captioning



Source: Ming Li's Lecture notes

Image Captioning

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio, “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”, ICML, 2015

Image Captioning

A large white bird standing in a forest.

A woman holding a clock in her hand.

A man wearing a hat and a hat on a skateboard.

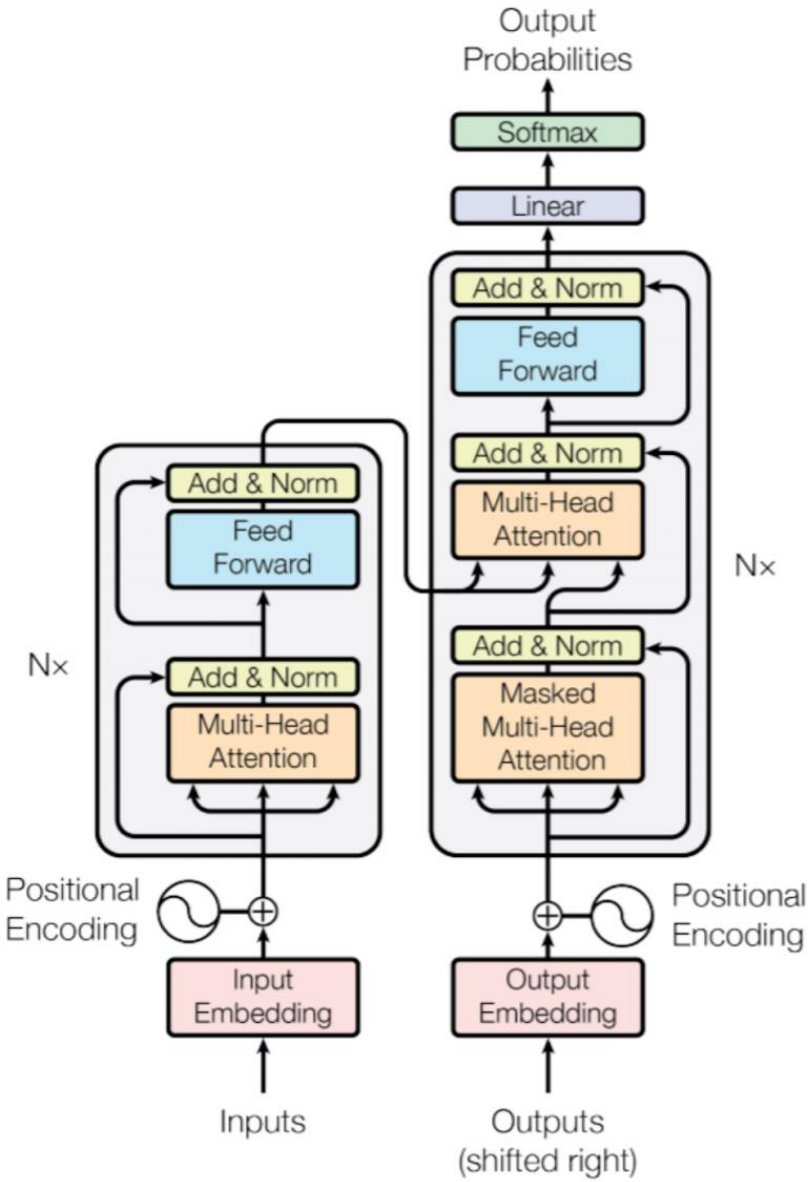
A person is standing on a beach with a surfboard.

A woman is sitting at a table with a large pizza.

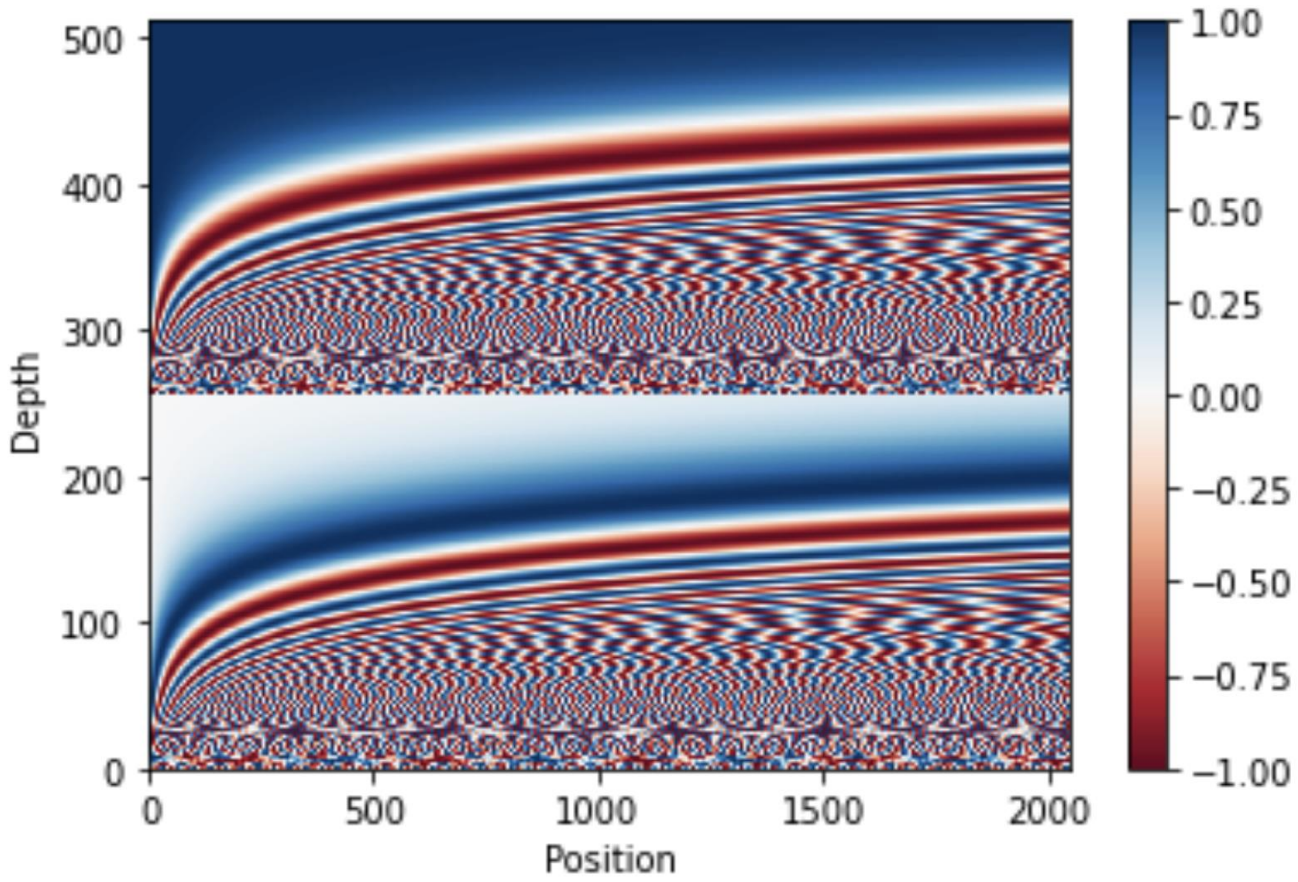
A man is talking on his cell phone while another man watches.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio, “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”, ICML, 2015

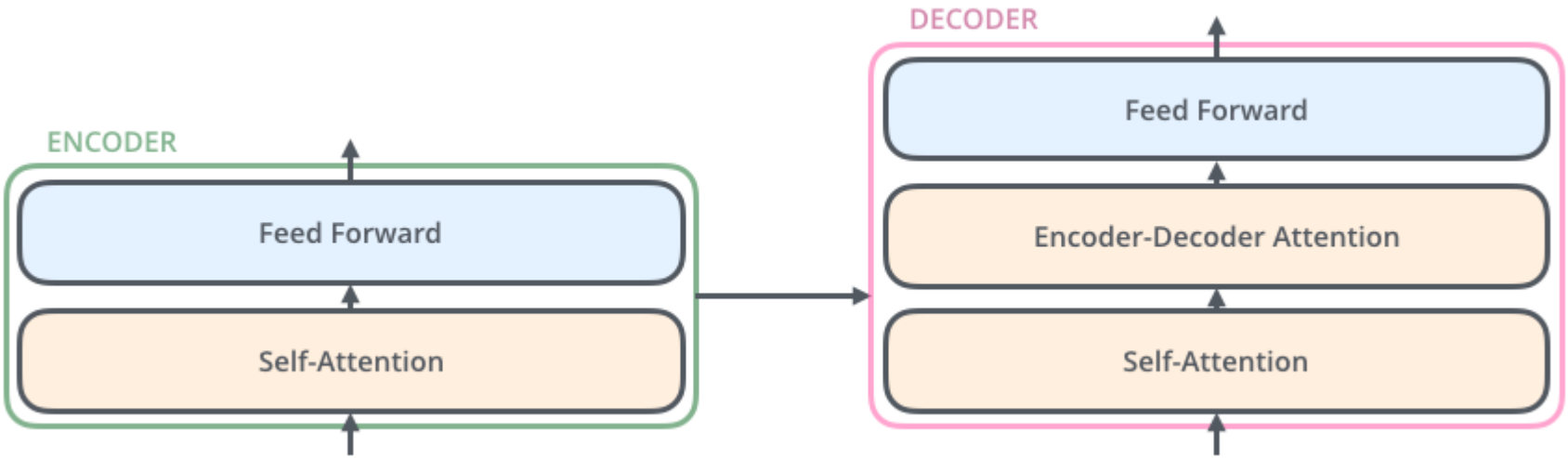
Transformers


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

Positional Encoding

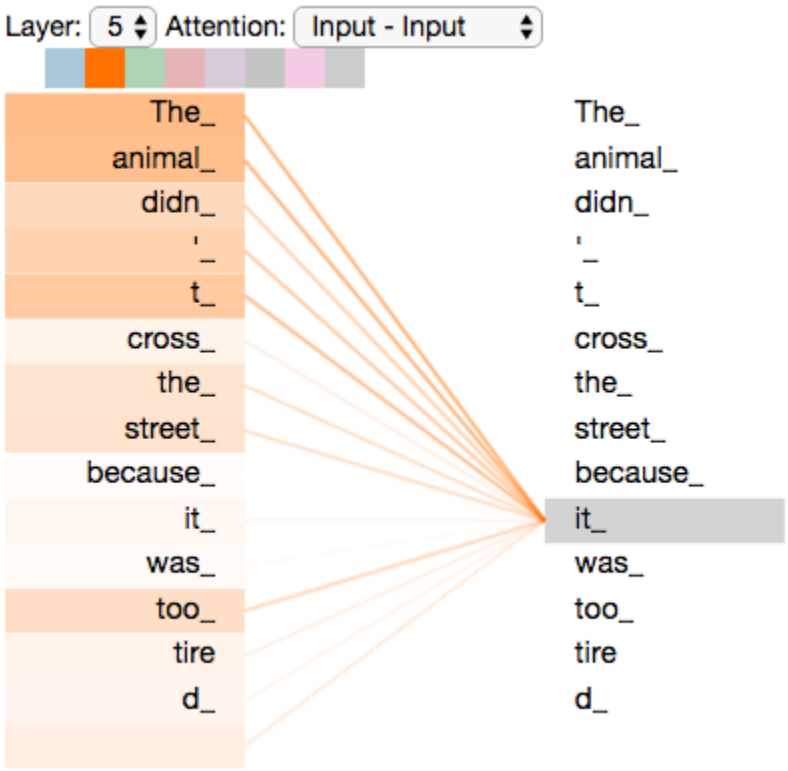


Simplified Version

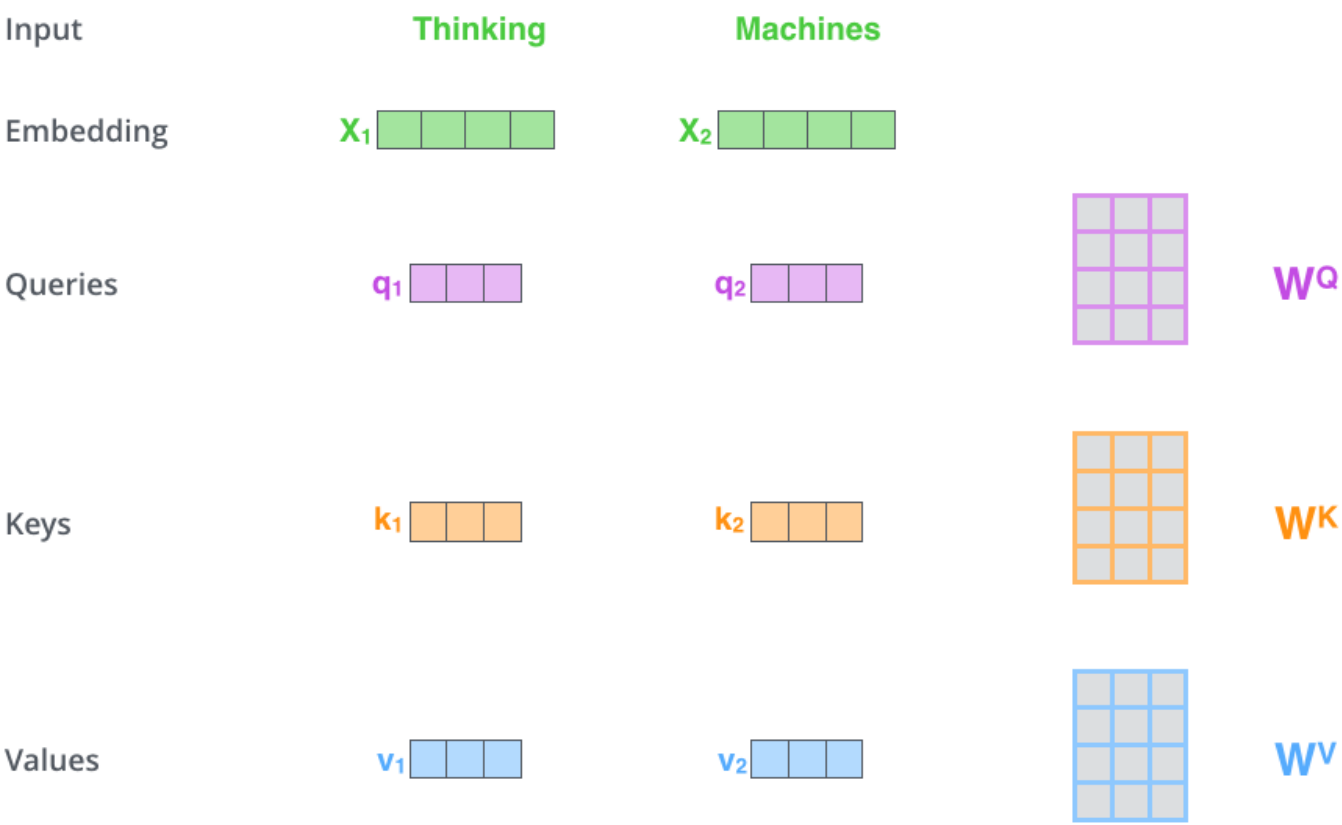


http://jalammar.github.io/illustrated-transformer/

Self Attention

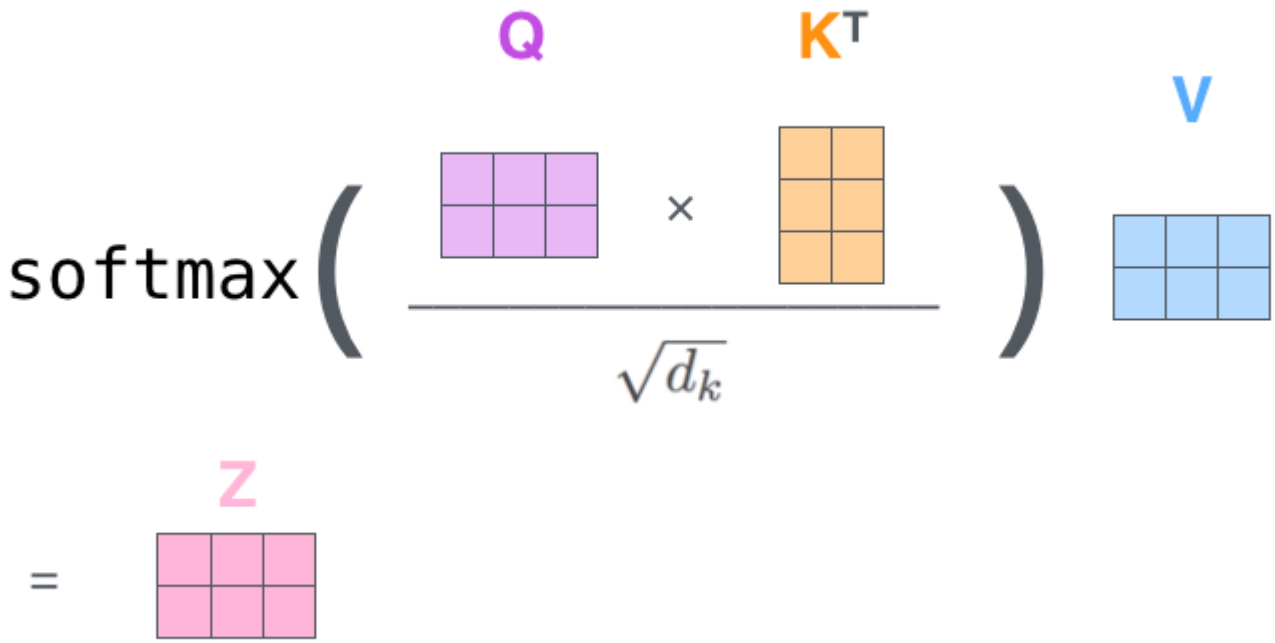


Self Attention



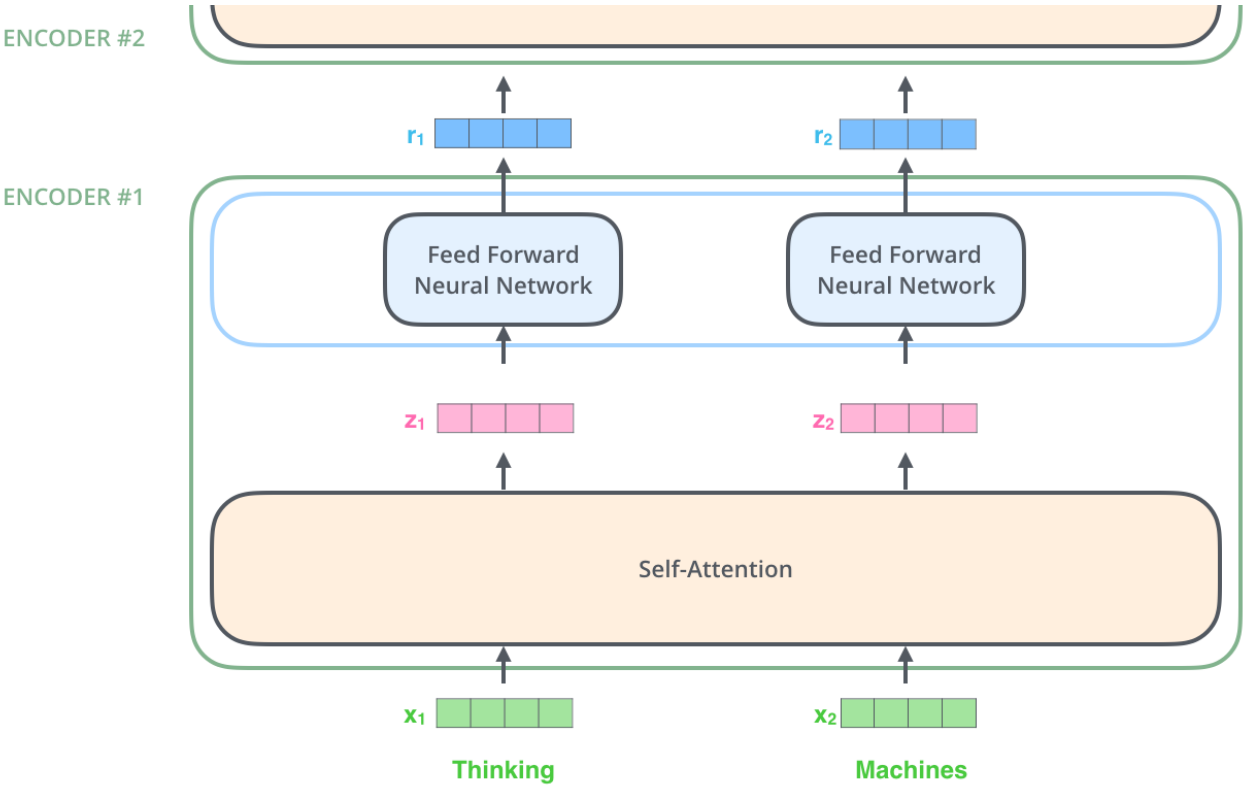
http://jalammar.github.io/illustrated-transformer/

Self Attention



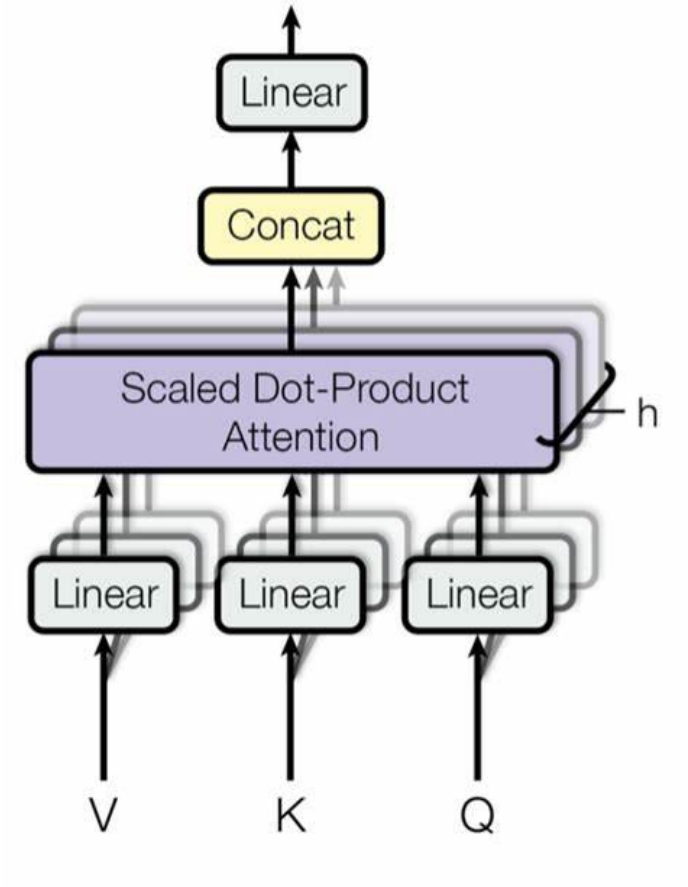
http://jalammar.github.io/illustrated-transformer/

Encoder

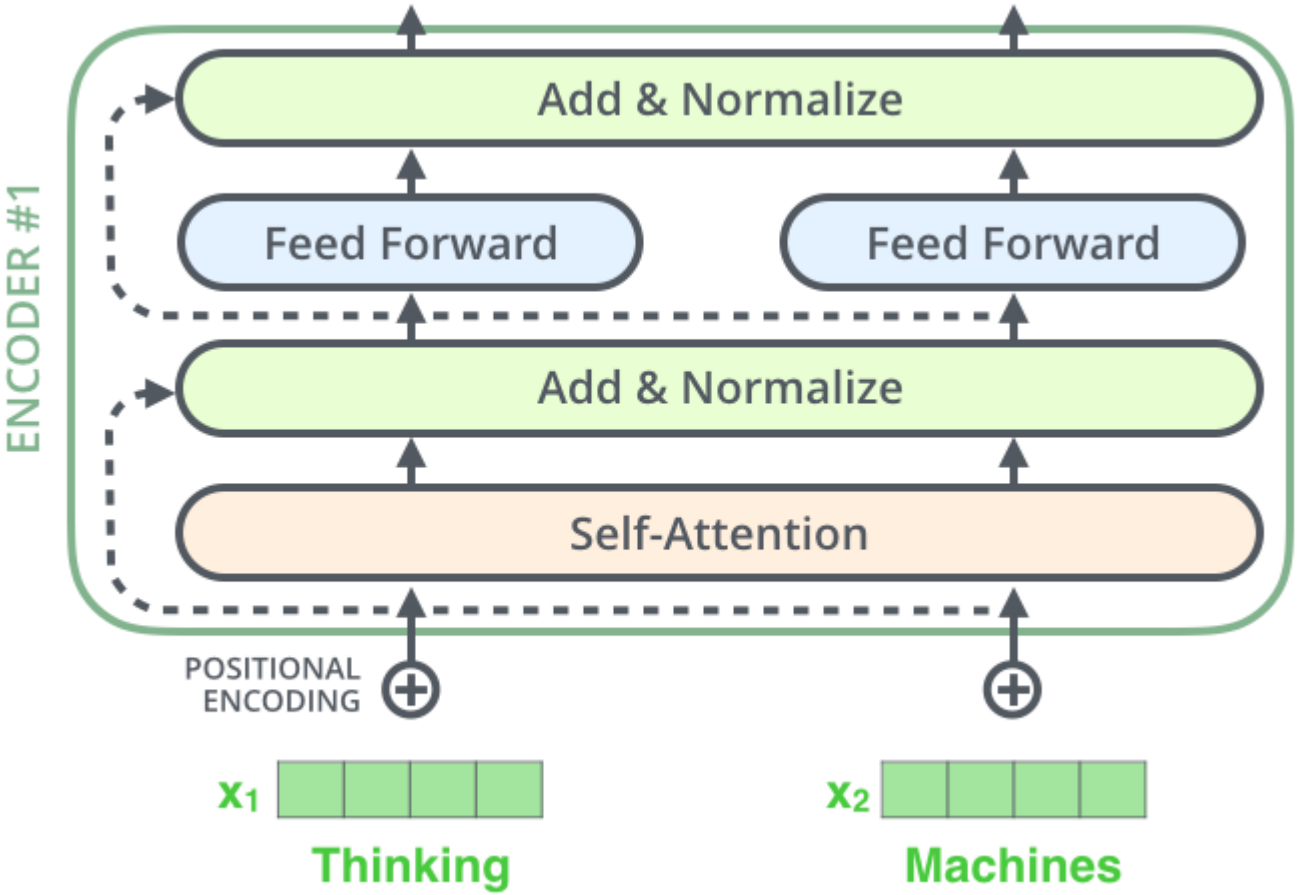


http://jalammar.github.io/illustrated-transformer/

Multi-Head Attention

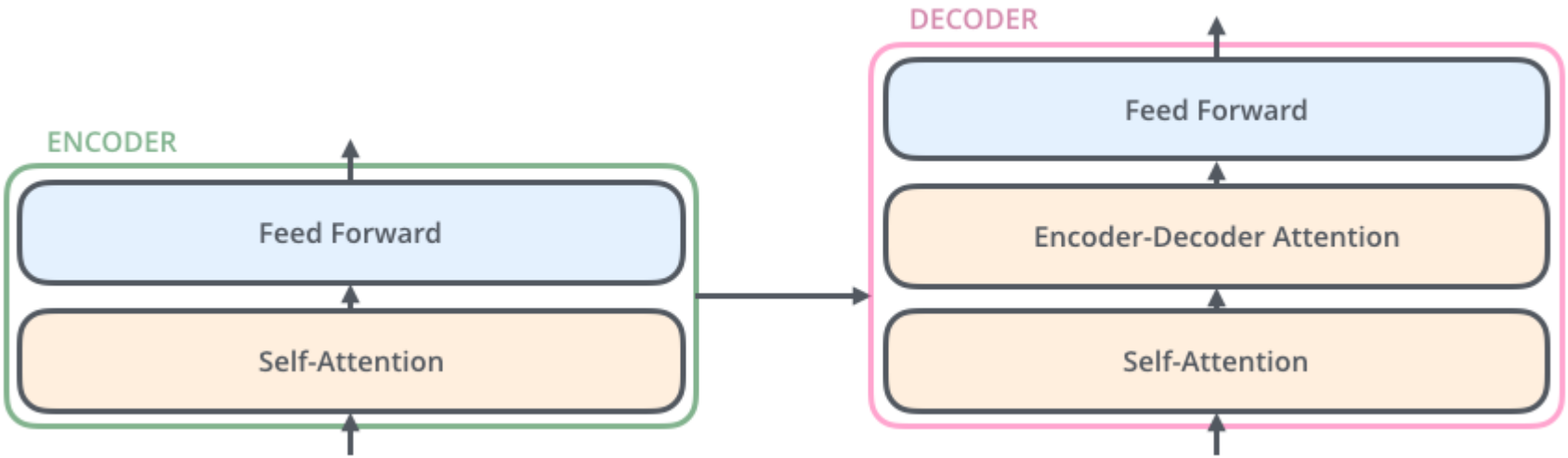


Residuals



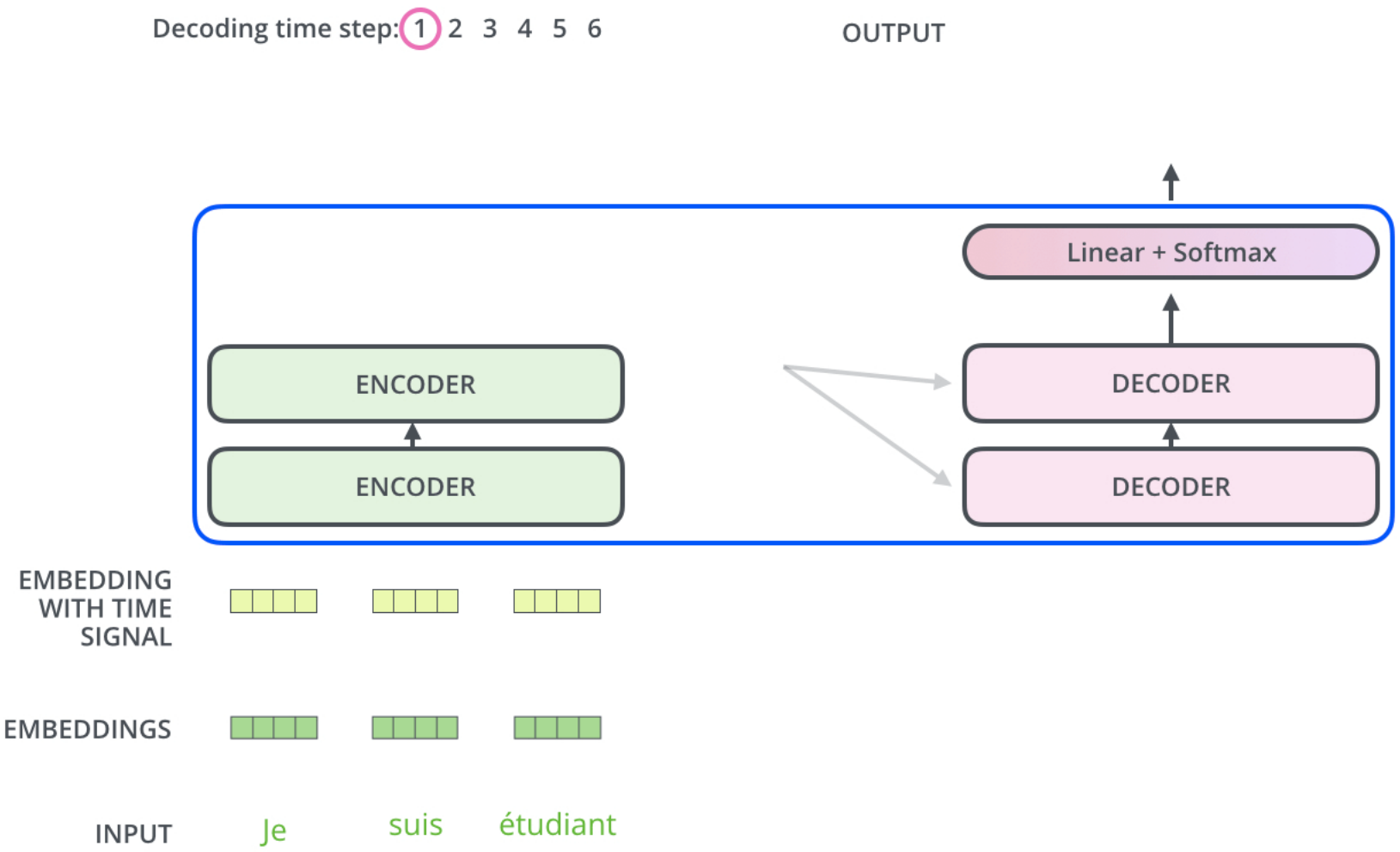
http://jalammar.github.io/illustrated-transformer/

Simplified Version



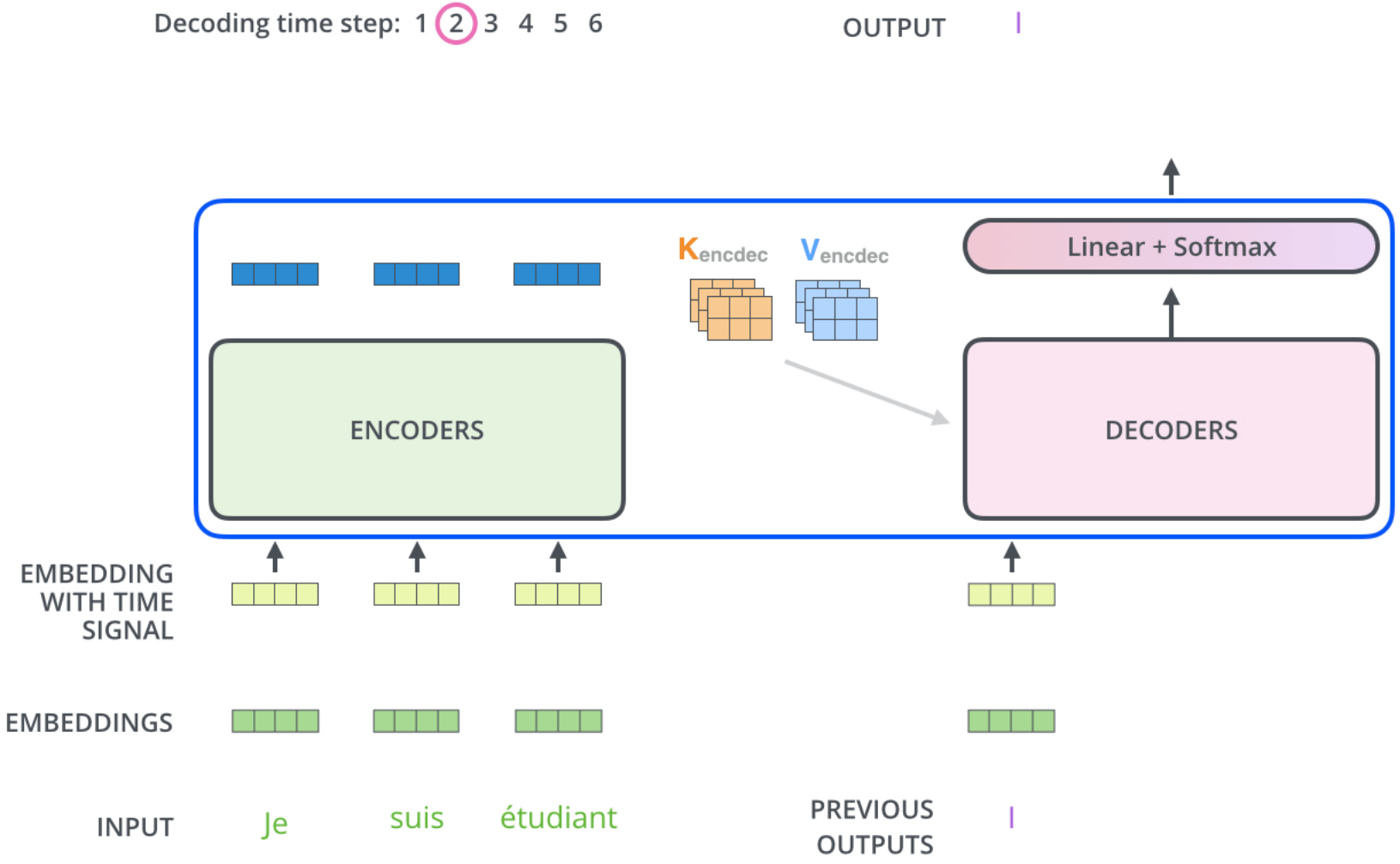
http://jalammar.github.io/illustrated-transformer/

Decoding



http://jalammar.github.io/illustrated-transformer/

Decoding



http://jalammar.github.io/illustrated-transformer/

Transformers Models

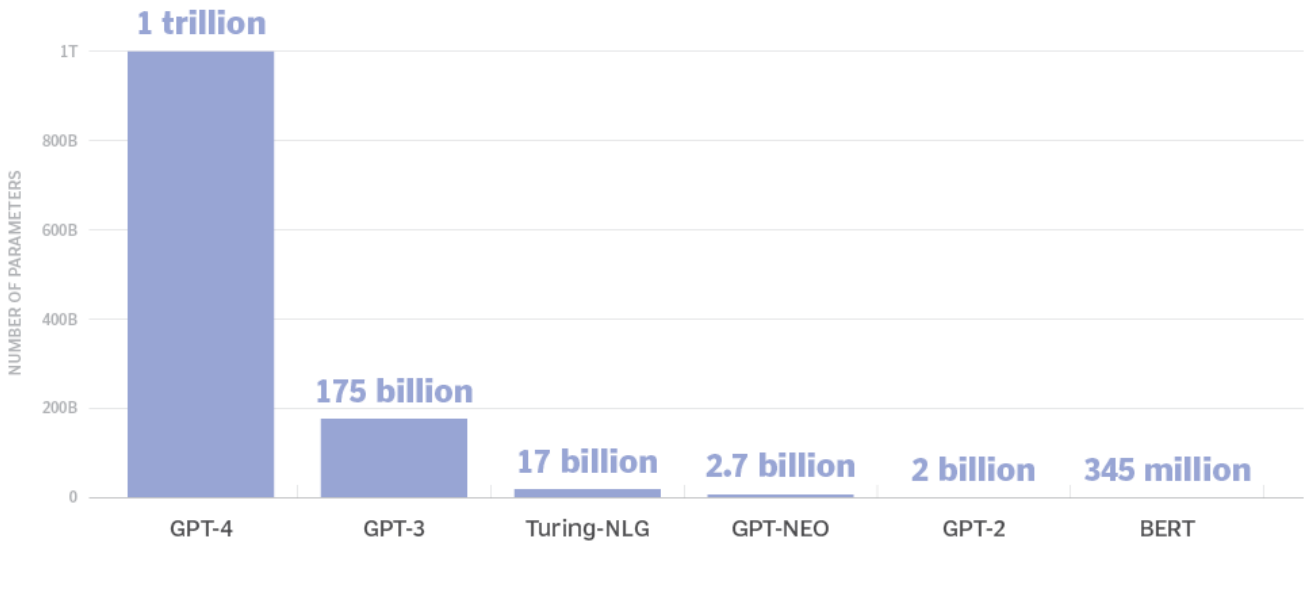
>BERT

>GPT2, GPT3, GPT3.5(ChatGPT), GPT4

>ViT

Large Language Models

They are LARGE



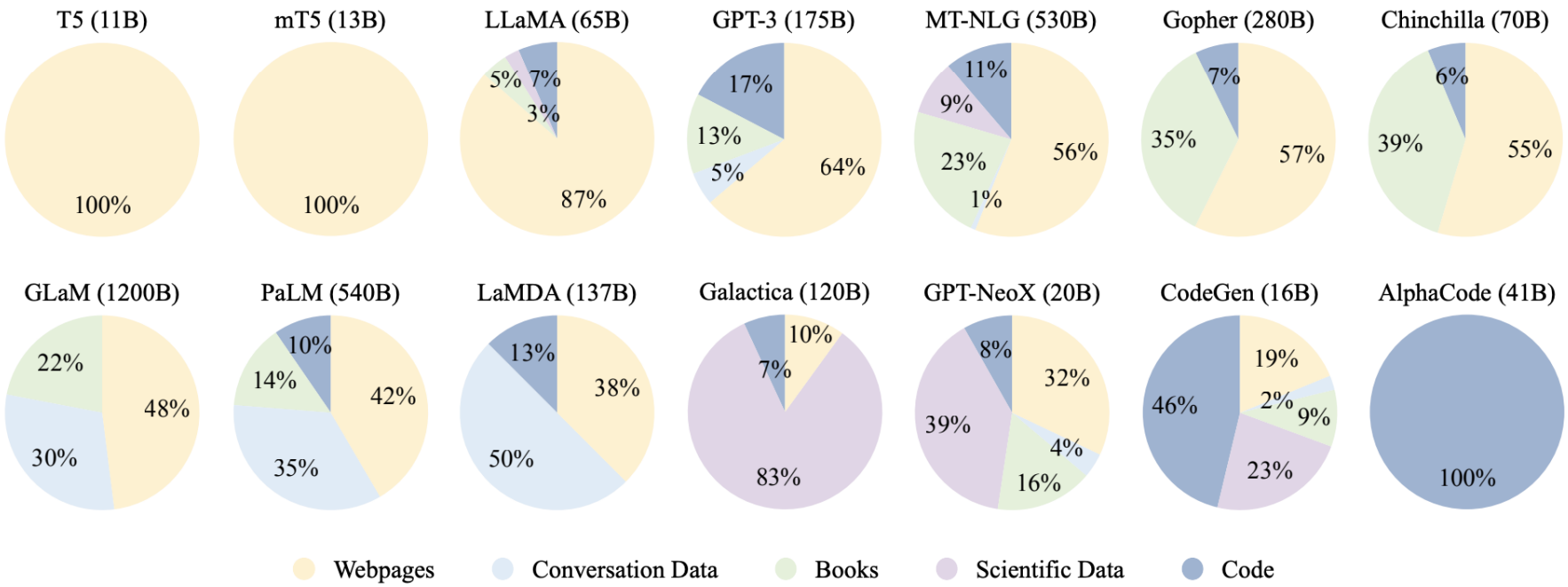
https://www.techtarget.com/whatis/definition/large-language-model-LLM

Why are they so LARGE?

- **Emergent abilities** of LLMs are formally defined as “the abilities that are not present in small models but arise in large models”

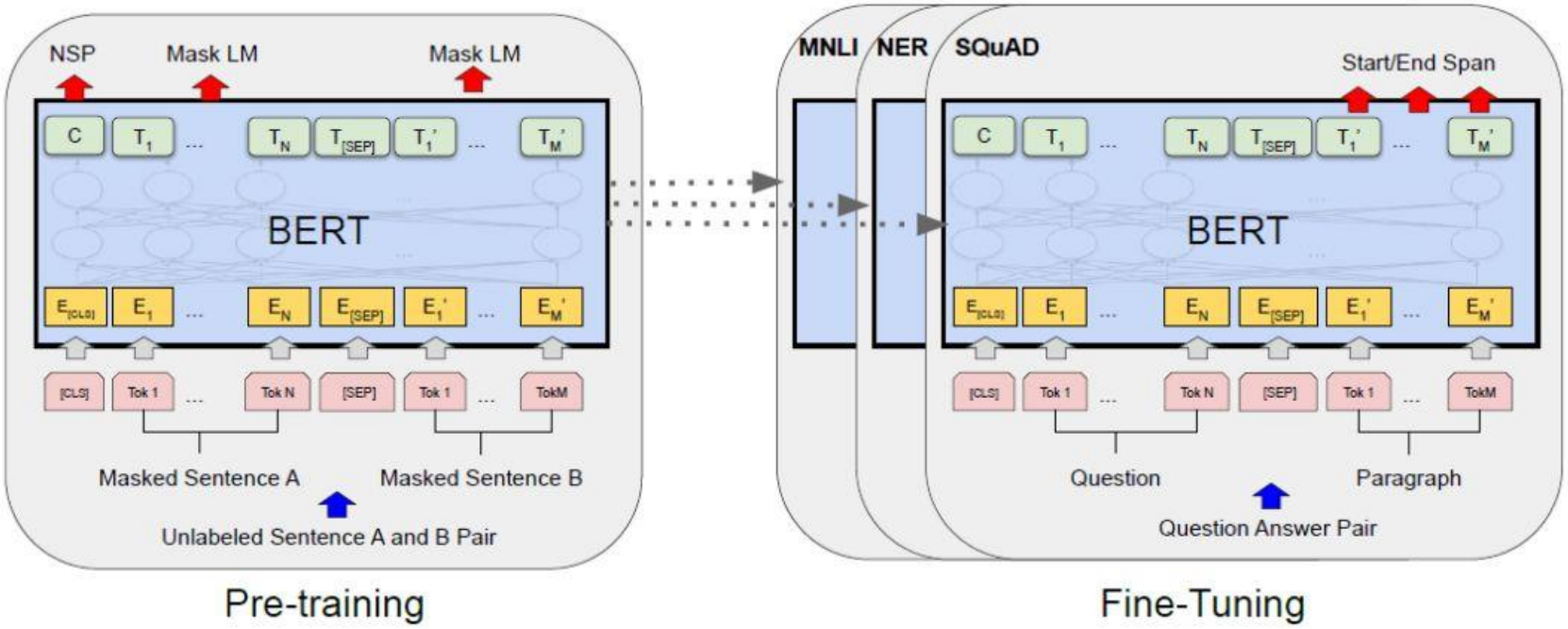
How can they be so LARGE?

> **Large corpora: Books, CommonCrawel, Reddit links, Wikipedia, Code, Conversation, …**

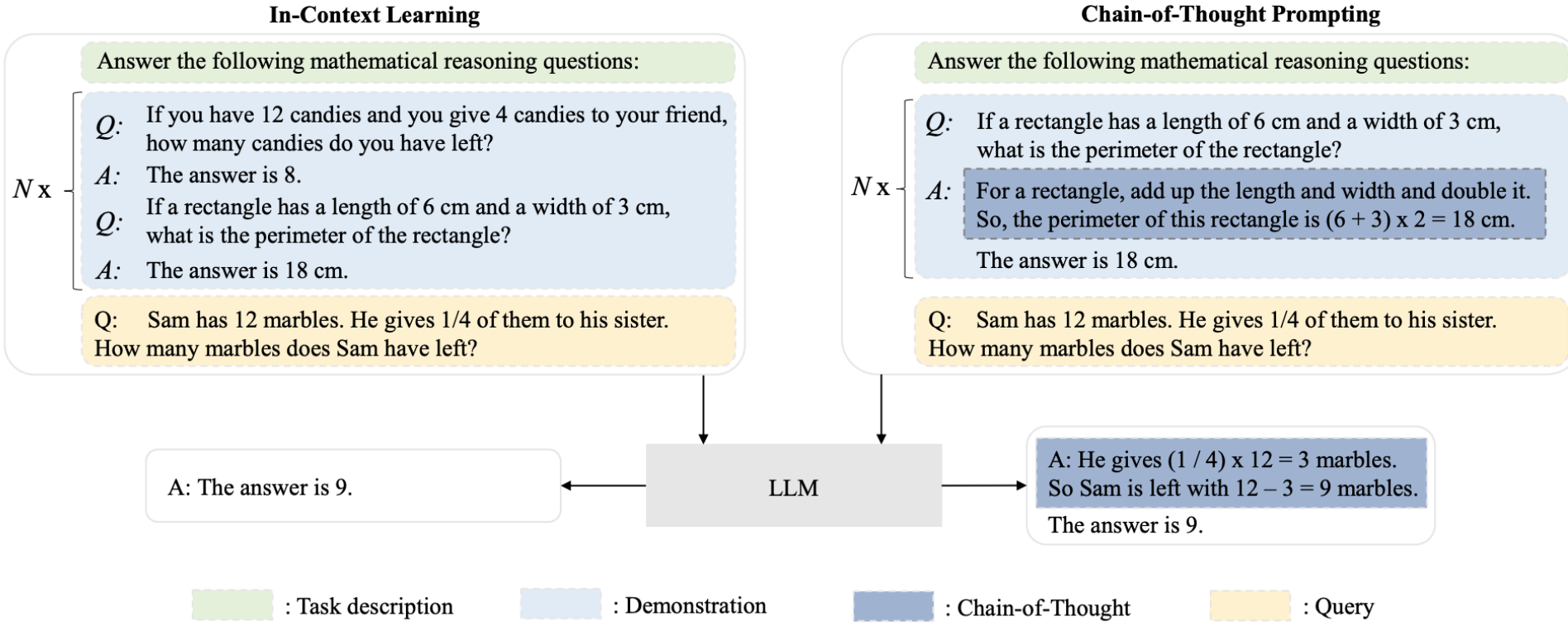


How can they be so LARGE?

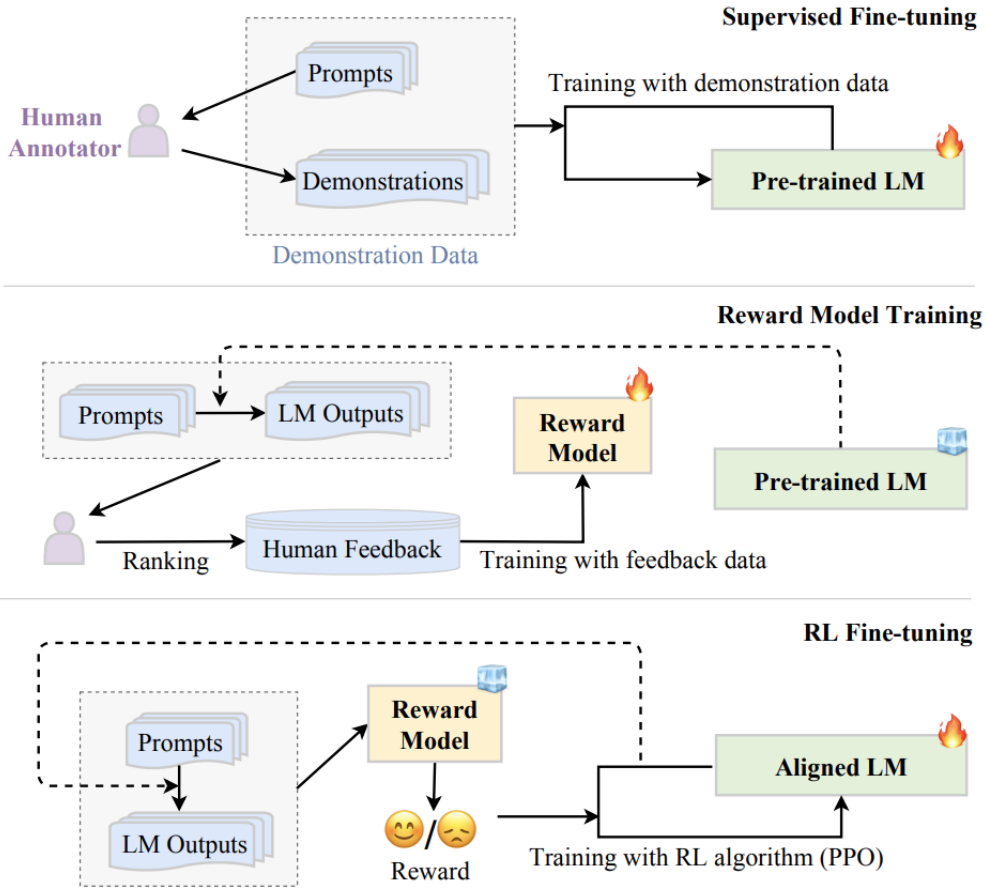
> **Pre-training: Language modeling, Denoising autoencoding, …**



How can they chat like us?



How can they chat like us?

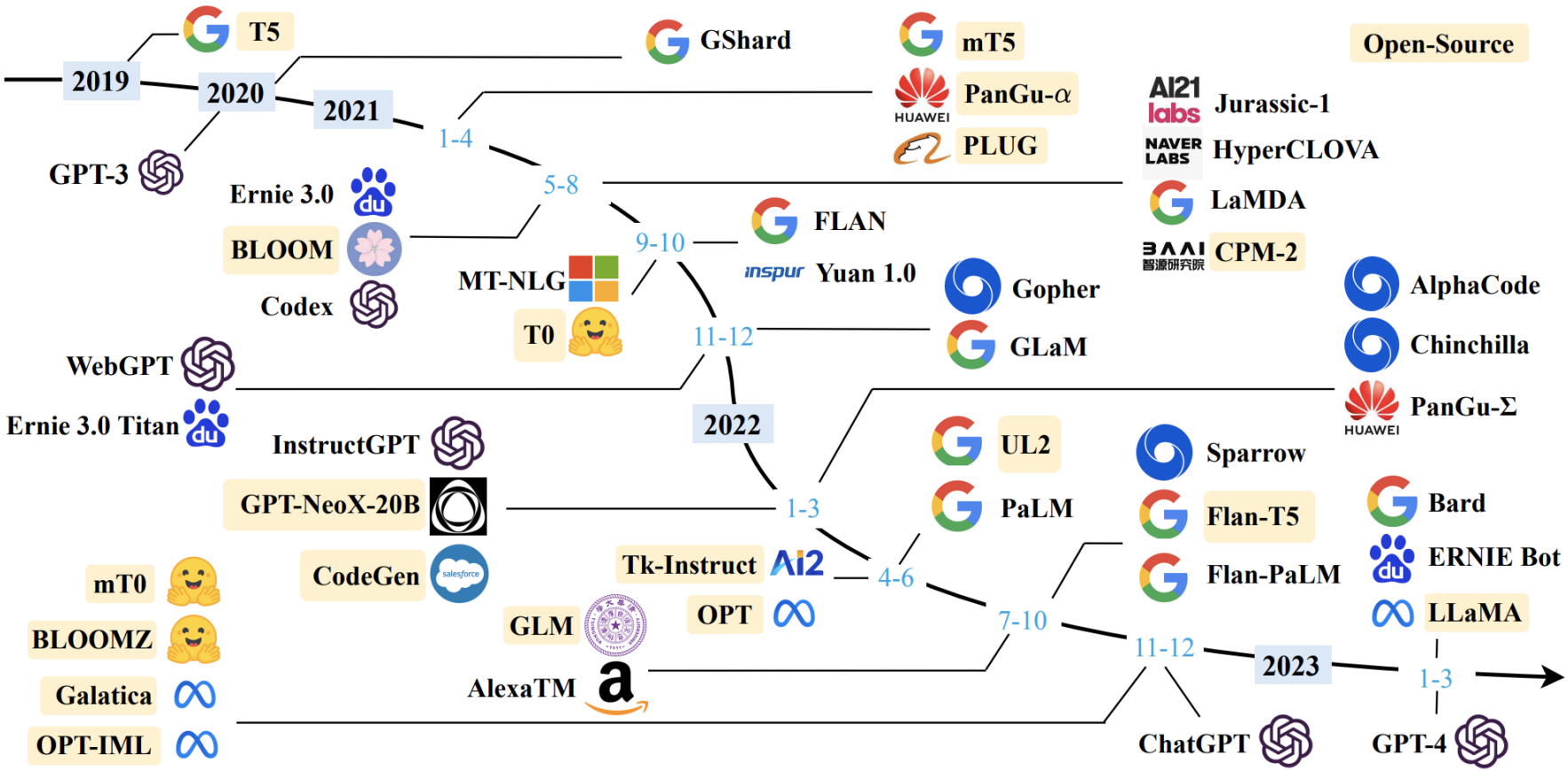


Are they just chatbots?

- **Text generation.** The ability to generate text on any topic that the LLM has been trained on is a primary use case.
- **Translation.** For LLMs trained on multiple languages, the ability to translate from one language to another is a common feature.
- **Content summary.** Summarizing blocks or multiple pages of text is a useful function of LLMs.

Are they just chatbots?

- **Rewriting content.** Rewriting a section of text is another capability.
- **Classification and categorization.** An LLM is able to classify and categorize content.
- **Sentiment analysis.** Most LLMs can be used for sentiment analysis to help users to better understand the intent of a piece of content or a particular response.
- **Conversational AI and chatbots.** LLMs can enable a conversation with a user in a way that is typically more natural than older generations of AI technologies.



https://arxiv.org/pdf/2303.18223.pdf

What can WE do with them?

GPT-4

GPT-4 can accept images as inputs and generate captions, classifications, and analyses.

Input

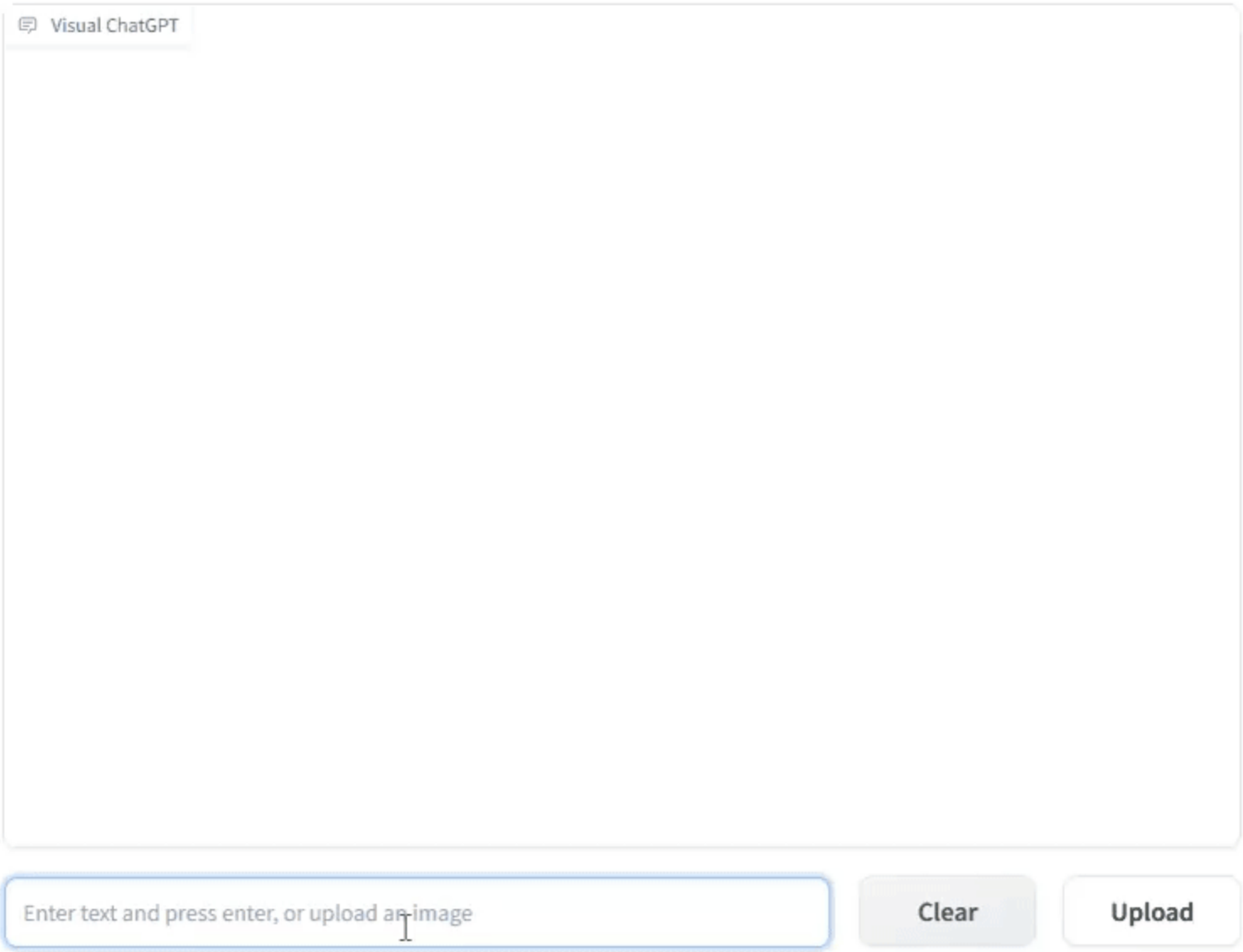
What can I make with these ingredients?

Output

There are many options for what you can make with these ingredients! Some possibilities include:

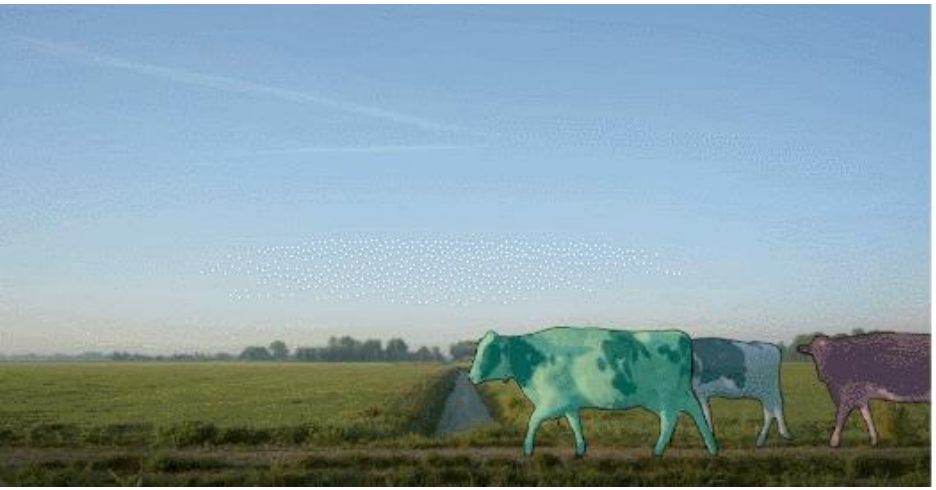
- Pancakes or waffles
- Crepes
- French toast
- Omelette or frittata
- Quiche
- Custard or pudding
- Cake or cupcakes
- Muffins or bread- Cookies or biscuits

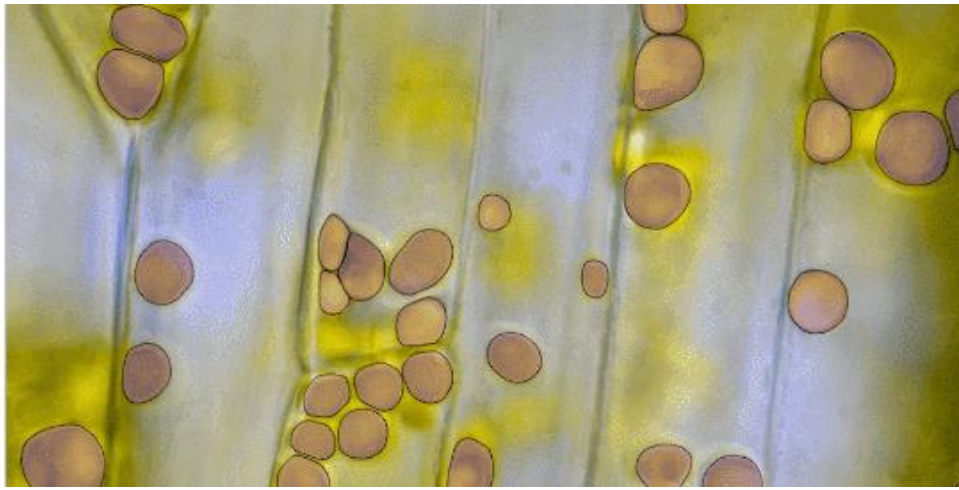
These are just a few examples, but the possibilities are endless!


Visual ChatGPT
Enter text and press enter, or upload an image
Clear
Upload

Meta - Segment Anything

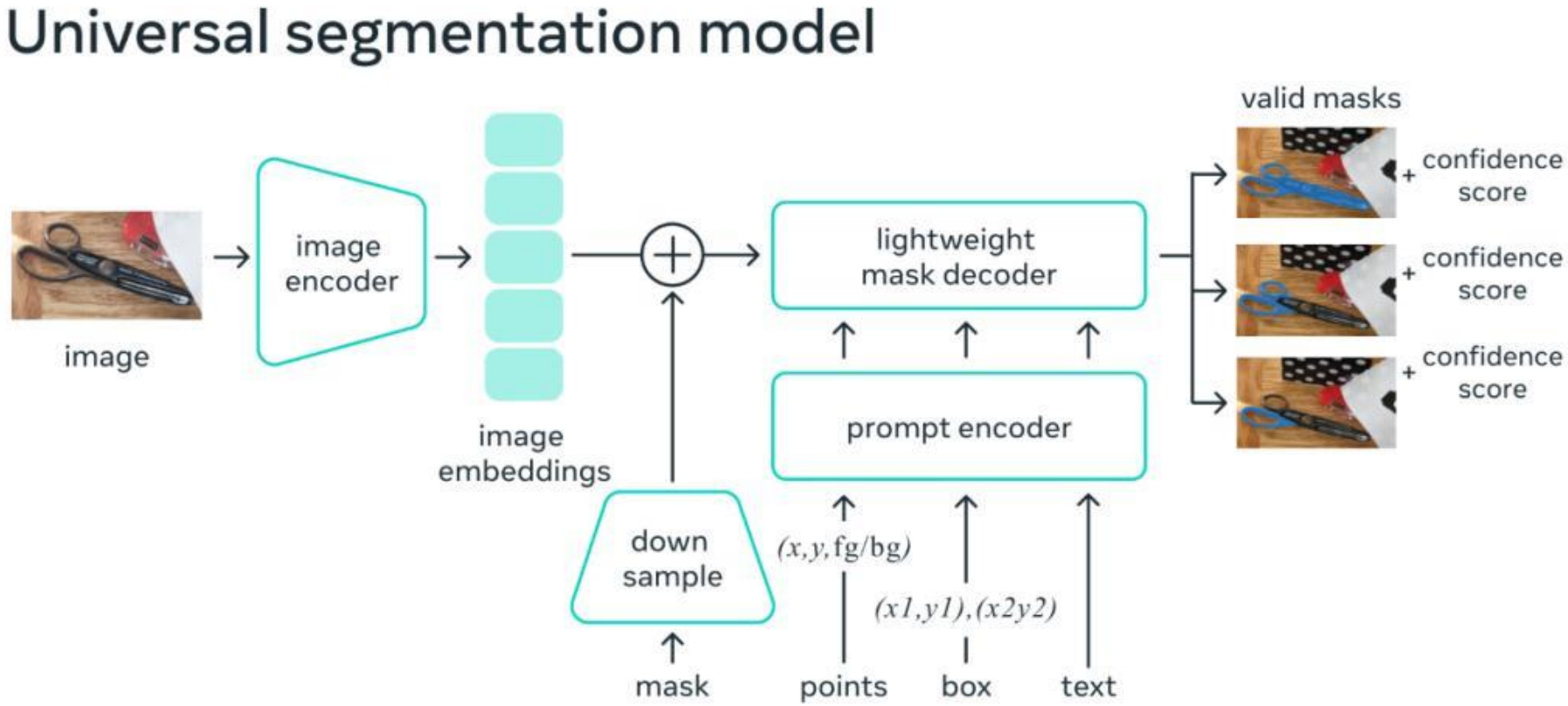
Meta - Segment Anything





Meta - Segment Anything

Meta - Segment Anything



Meta SAM

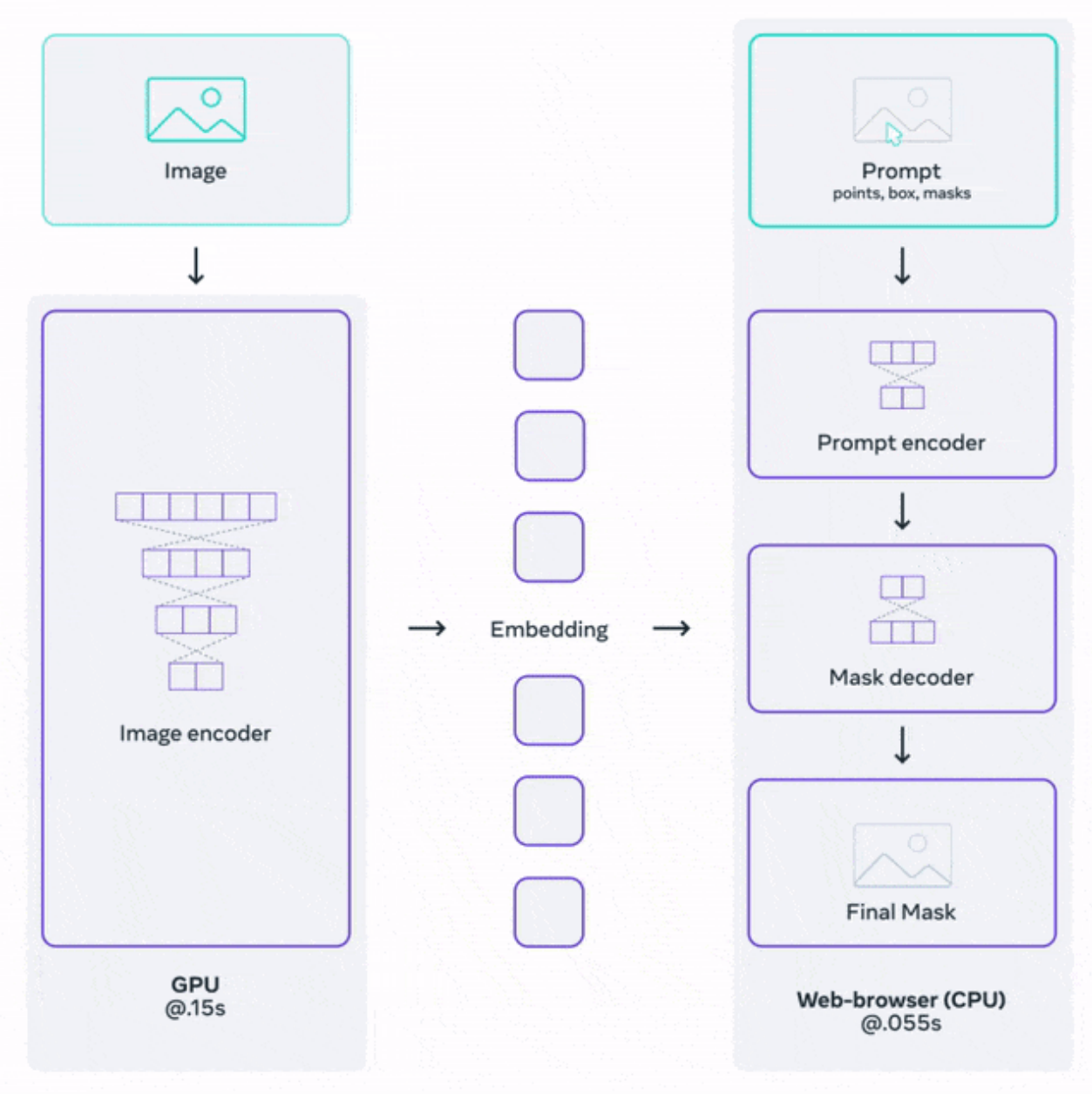
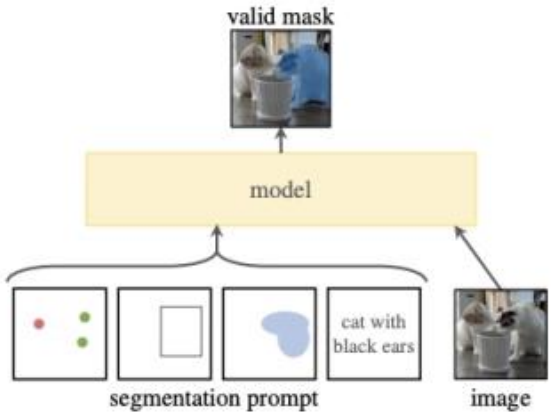
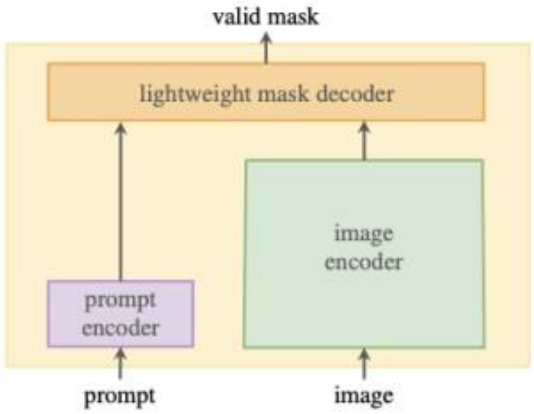


Image
Image encoder
GPU
@.15s
Embedding
Prompt
points, box, masks
Prompt encoder
Mask decoder
Final Mask
Web-browser (CPU)
@.055s

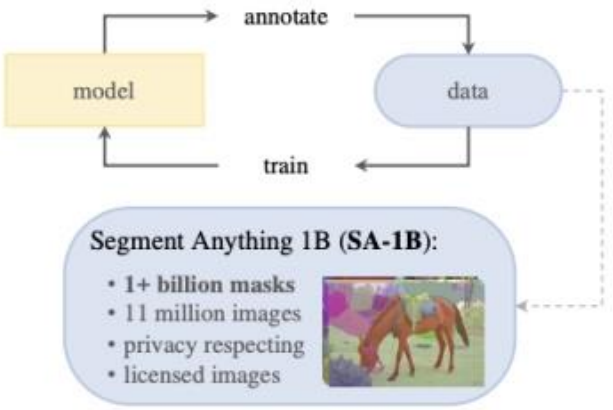
Meta - Segment Anything



(a) **Task**: promptable segmentation



(b) **Model**: Segment Anything Model (**SAM**)



(c) **Data**: data engine (top) & dataset (bottom)

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學
Department of Computing
電子計算學系

Thank you!