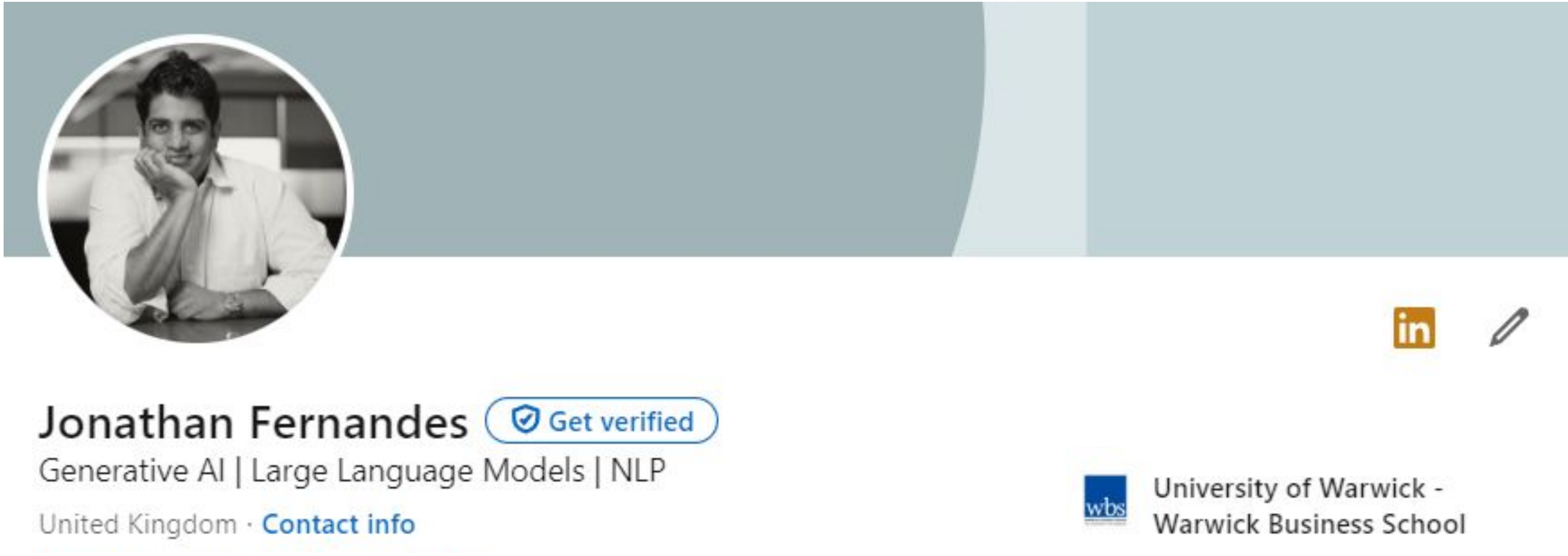
Advanced AI: Text to Image Generation

About me

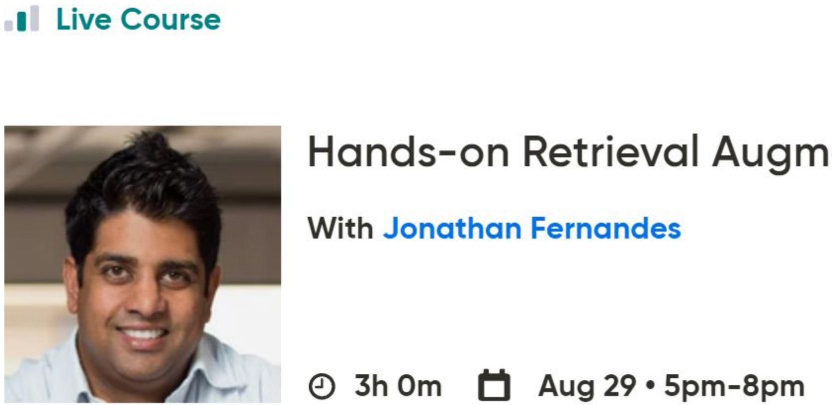


Jonathan Fernandes

Generative AI | Large Language Models | NLP

United Kingdom · Contact info

University of Warwick - Warwick Business School



Live Course

Hands-on Retrieval Augmented Generation (RAG)

With Jonathan Fernandes

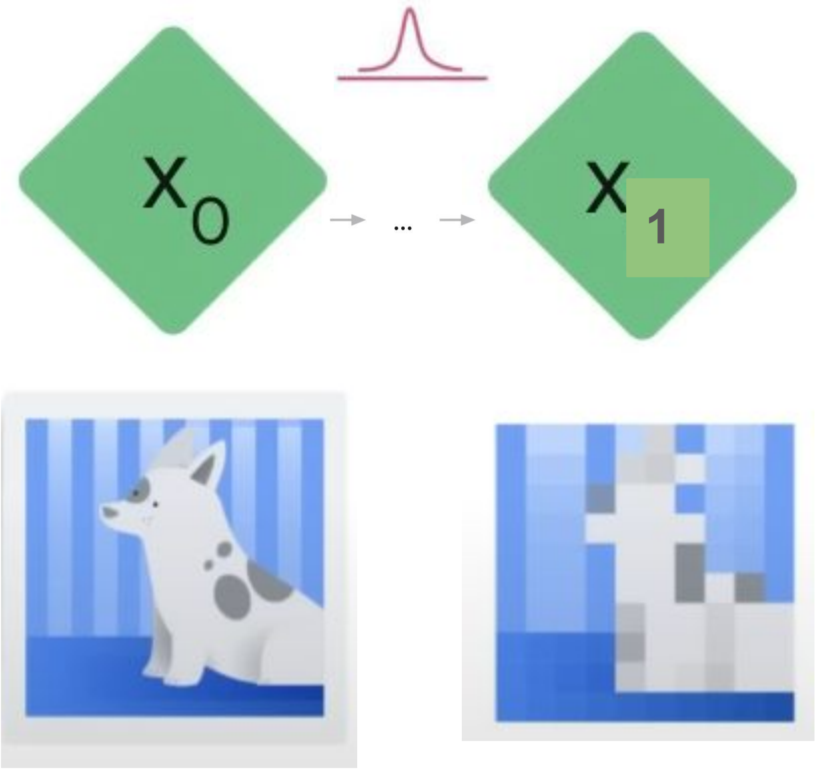
3h 0m

Aug 29 • 5pm-8pm

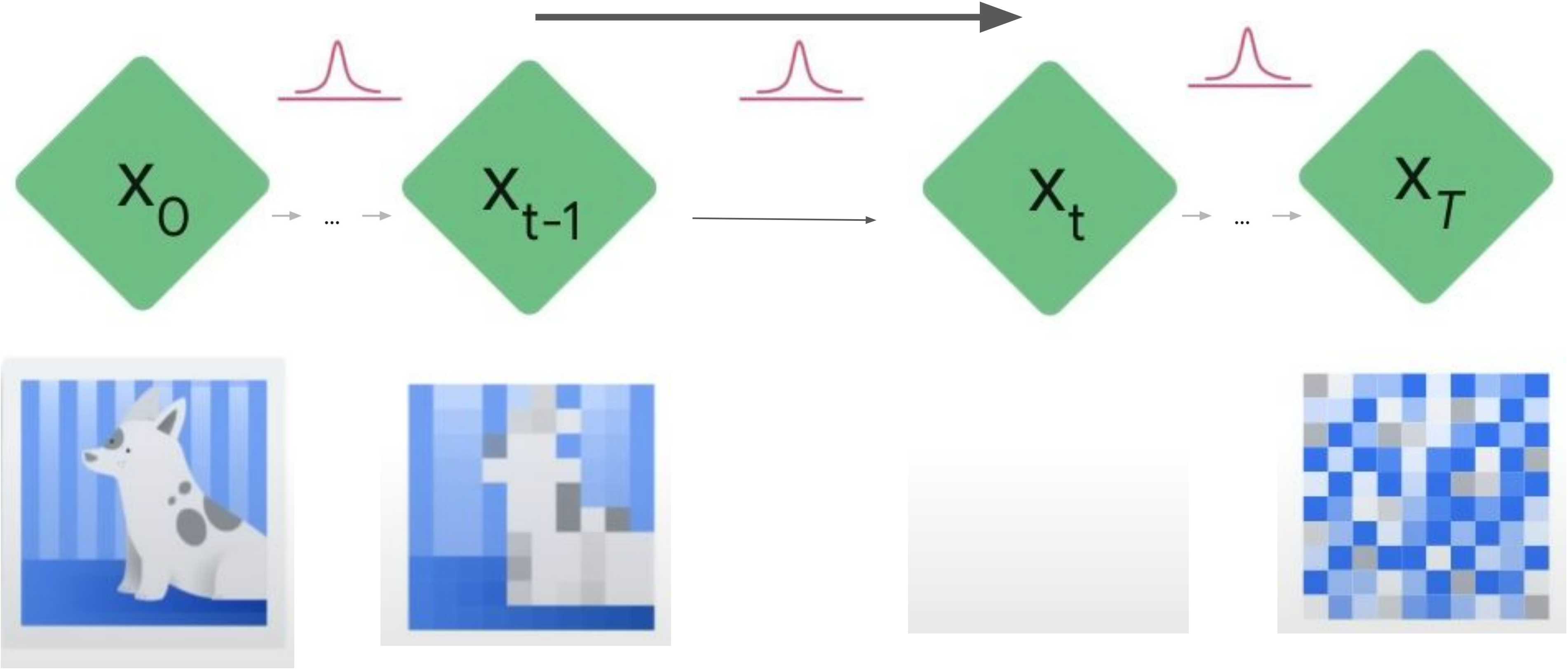
What is diffusion?

What are diffusion models?

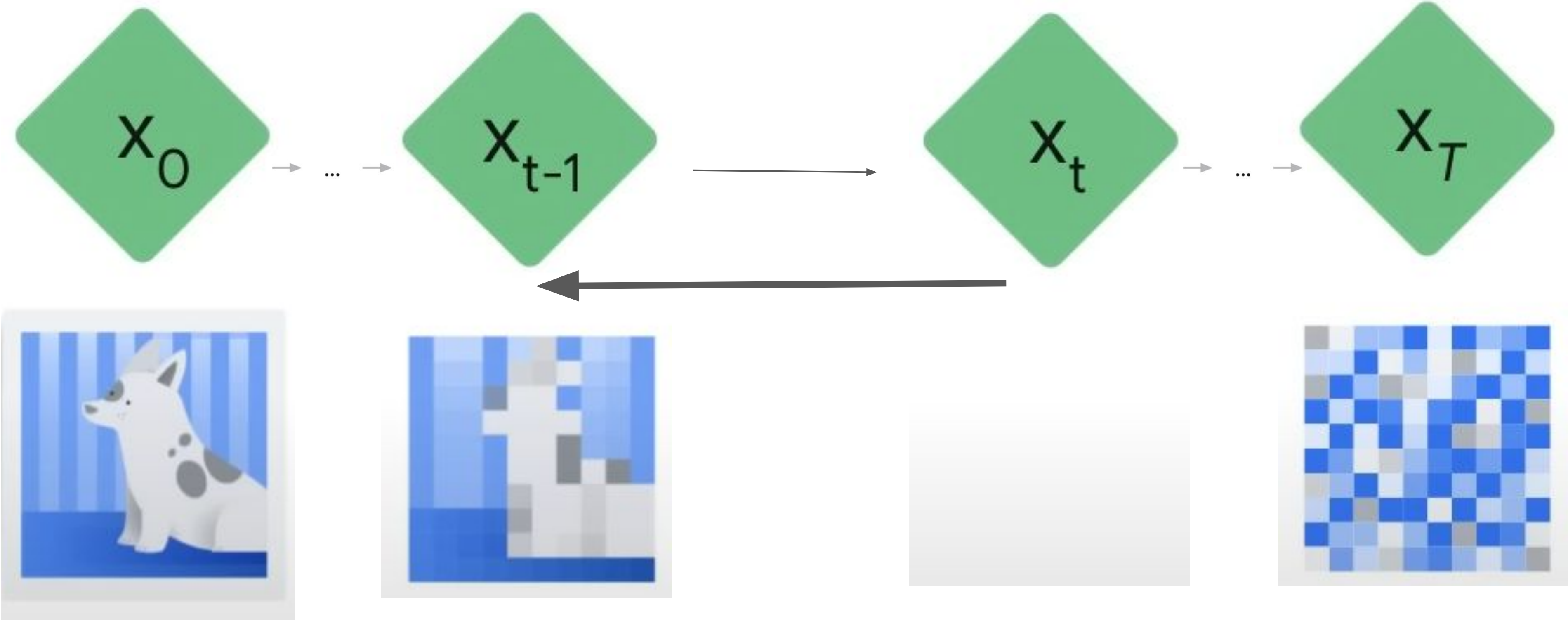
What are diffusion models?



What are diffusion models?



What are diffusion models?



Go to notebook

https://playground.com/

https://playground.com/

- (Set the number of images generated to 1 or 2 - otherwise, you will run out of credits quickly)
- Shared google doc (image and prompt)
- 3-minute exercise
- Our end goal - Use text to generate images

The latest and greatest in text to image generation

FLUX.1

- Released this month
- Black forest labs
- 3 flavours
  - FLUX.1 dev
  - FLUX.1 schnell
  - FLUX.1 pro
- No NSFW filter
- https://huggingface.co/black-forest-labs/FLUX.1-dev
- https://huggingface.co/spaces/black-forest-labs/FLUX.1-dev

What are some of the problems here?

BURGER KING
BURGER KING

How would you ensure there is no nudity/violence/gore in the images that are generated? [3 minutes]

Can you provide images and associated prompts of bias [5 minutes]

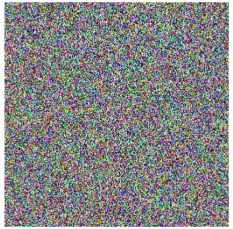
Image inpainting:

https://huggingface.co/spaces/SkalskiP/FLUX.1-inpaint

[3 minutes]

High level overview

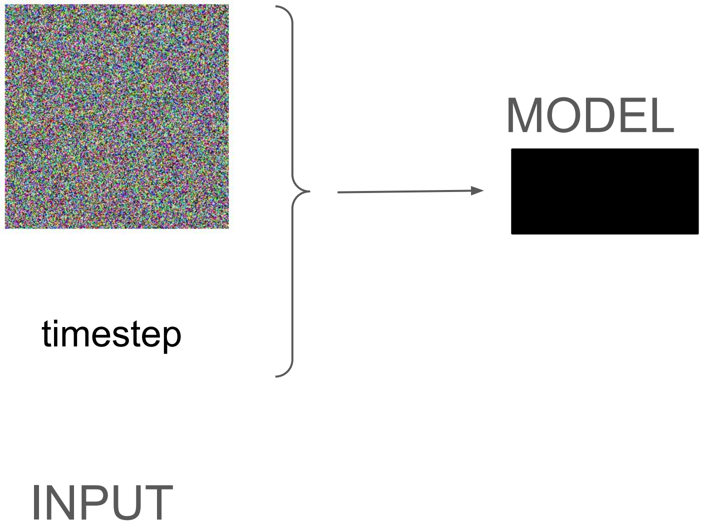
Diffusion models (inference)



timestep

INPUT

Diffusion models (inference)

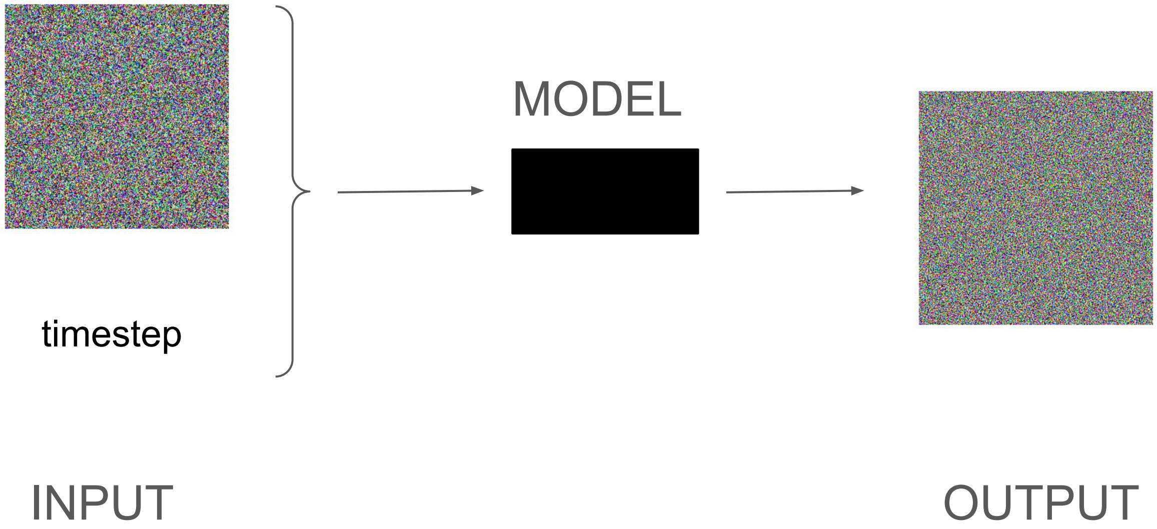


timestep

MODEL

INPUT

Diffusion models (inference)



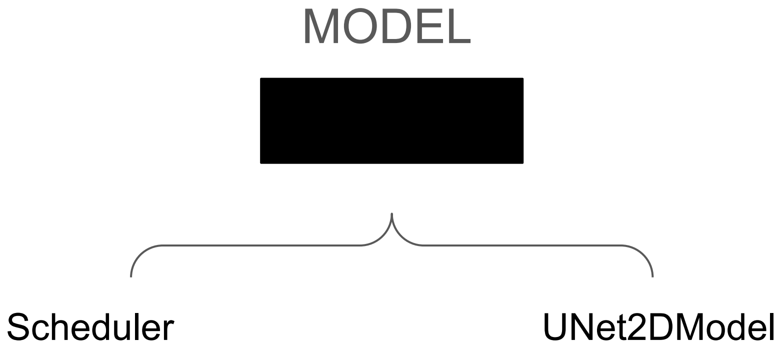
timestep

MODEL

INPUT

OUTPUT

Diffusion models (inference)



MODEL

Scheduler

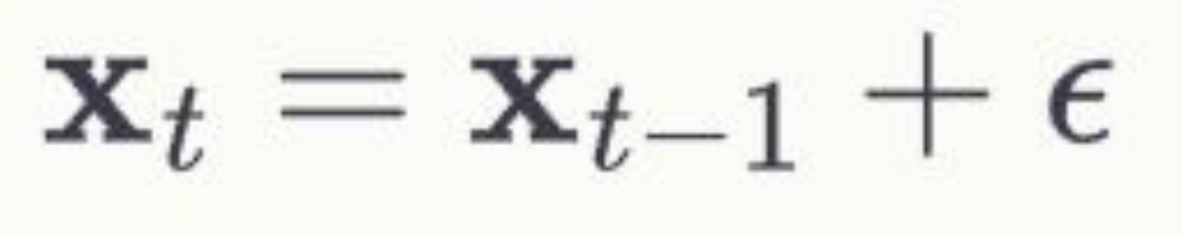
UNet2DModel

Go to notebook

Types of Diffusion Models

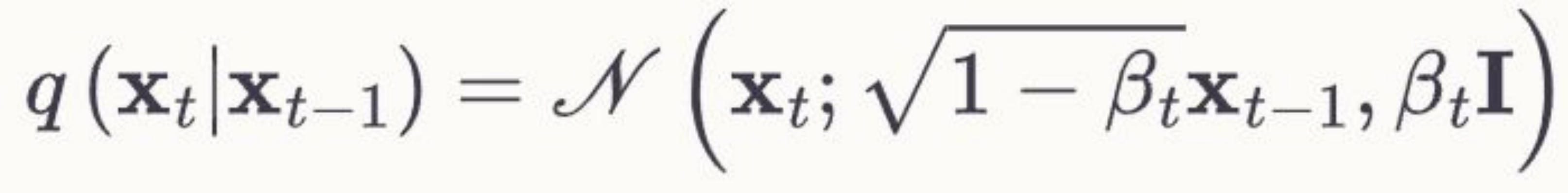
- Unconditional
- Conditional

Scheduler



$$\mathbf{x}_t = \mathbf{x}_{t-1} + \epsilon$$

$$\mathbf{x}_t = \sqrt{1-\beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\epsilon$$

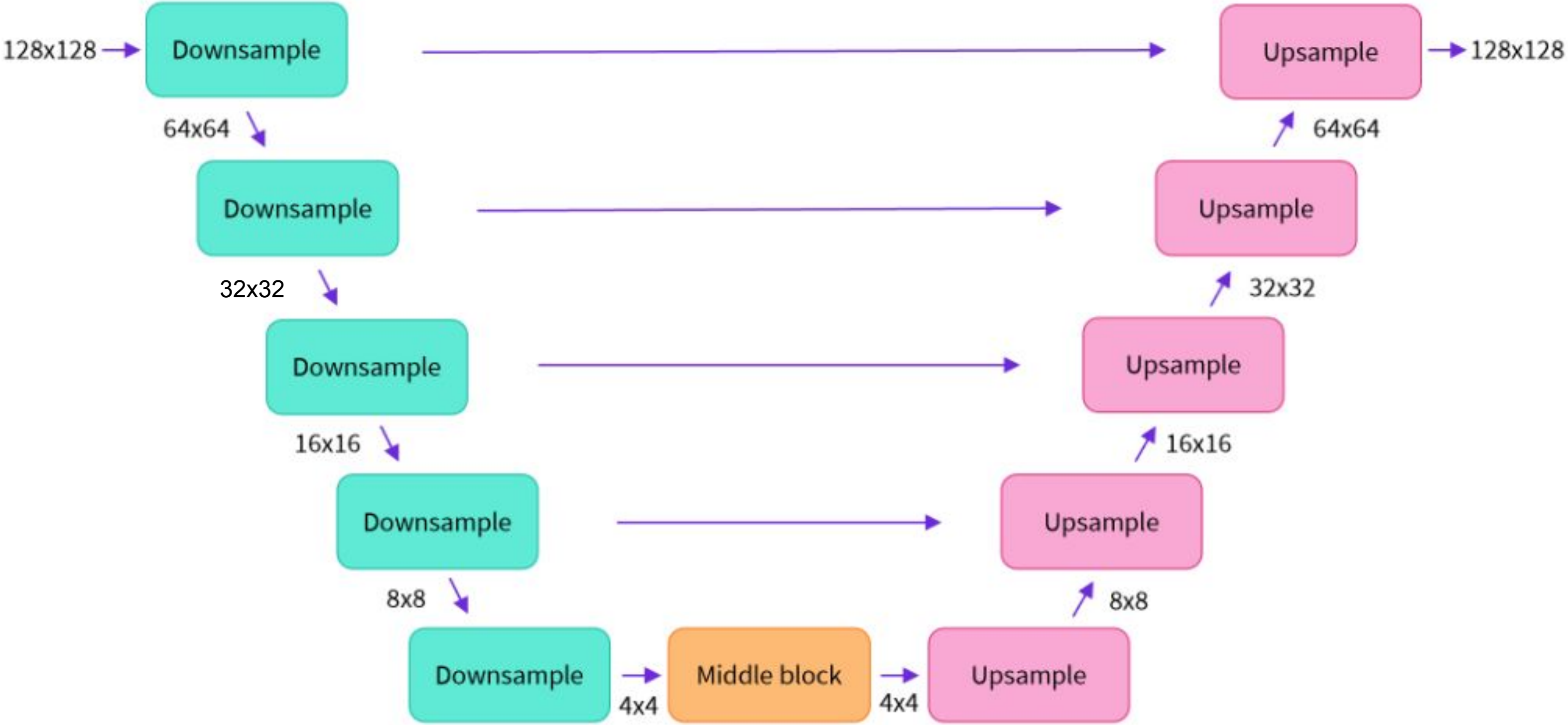


$$q\left(\mathbf{x}_t|\mathbf{x}_{t-1}\right) = \mathscr{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}\right)$$

[Notebook: Scheduler]

U-Net model

U-Net model


128x128
Downsample
64x64
Downsample
32x32
Downsample
16x16
Downsample
8x8
Downsample
4x4
Middle block
4x4
Upsample
8x8
Upsample
16x16
Upsample
32x32
Upsample
64x64
Upsample
128x128

[Notebook: U-Net Model]

Train a model

Train a model

Load training image from dataset

Train a model

Load training image from dataset

Add varying noise levels for diverse denoising tasks.

Train a model

Load training image from dataset

Add varying noise levels for diverse denoising tasks.

Input noisy images to the model.

Train a model

Load training image from dataset

Add varying noise levels for diverse denoising tasks.

Input noisy images to the model.

Evaluate model's denoising performance.

Train a model

Load training image from dataset

Add varying noise levels for diverse denoising tasks.

Input noisy images to the model.

Evaluate model's denoising performance.

Update model weights based on evaluation.

Notebook: Train a model

Evaluating Generated Images

Fréchet Inception Distance

- Creating Artwork (Generating Images)

Fréchet Inception Distance

- Creating Artwork (Generating Images)
- Art Inspector (Inception Model)
-

Fréchet Inception Distance

- Creating Artwork (Generating Images)
- Art Inspector (Inception Model)
- Gallery Walk (Feature Extraction)

Fréchet Inception Distance

- Creating Artwork (Generating Images)
- Art Inspector (Inception Model)
- Gallery Walk (Feature Extraction)
- Comparing Notes (FID score)

Limitations of FID

- Compare distributions, not for a single image
- Not good for low or high-resolution images
- Sensitive to many factors

Limitations of FID

- Compare distributions, not for a single image
- Not good for low or high-resolution images
- Sensitive to many factors

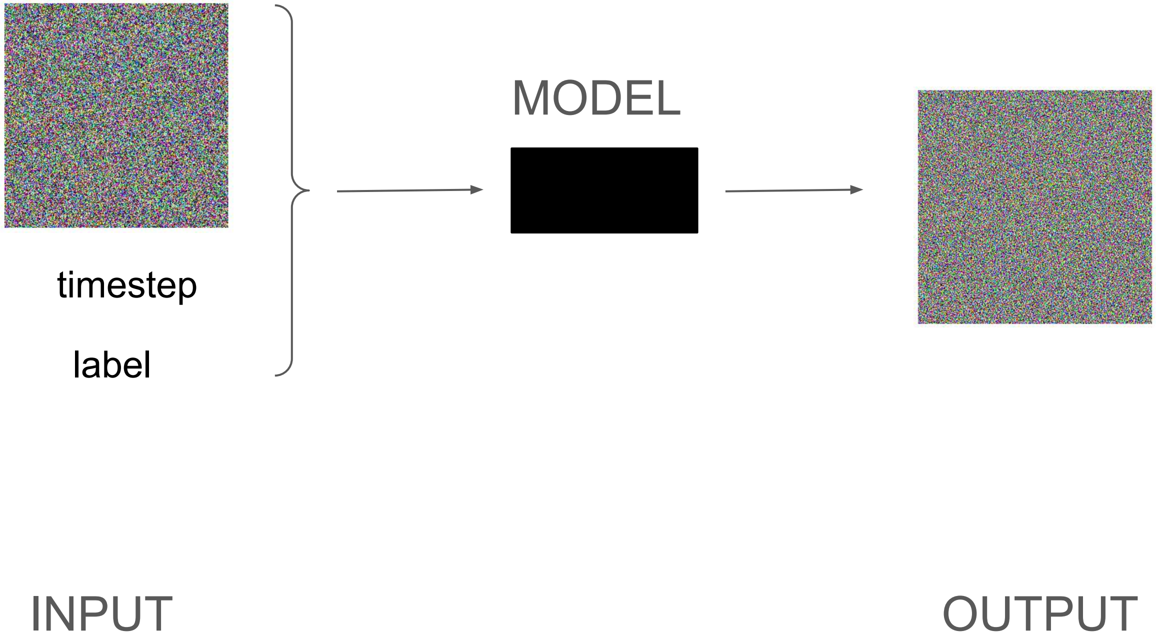
Limitations of FID

- Compare distributions, not for a single image
- Not good for low or high-resolution images
- Sensitive to many factors

Conditioned Diffusion models

https://github.com/zalandoresearch/fashion-mnist

Diffusion models (inference)



MODEL

timestep

label

INPUT

OUTPUT

[Notebook: Conditioned Diffusion Models]

Exercise - CIFAR-10

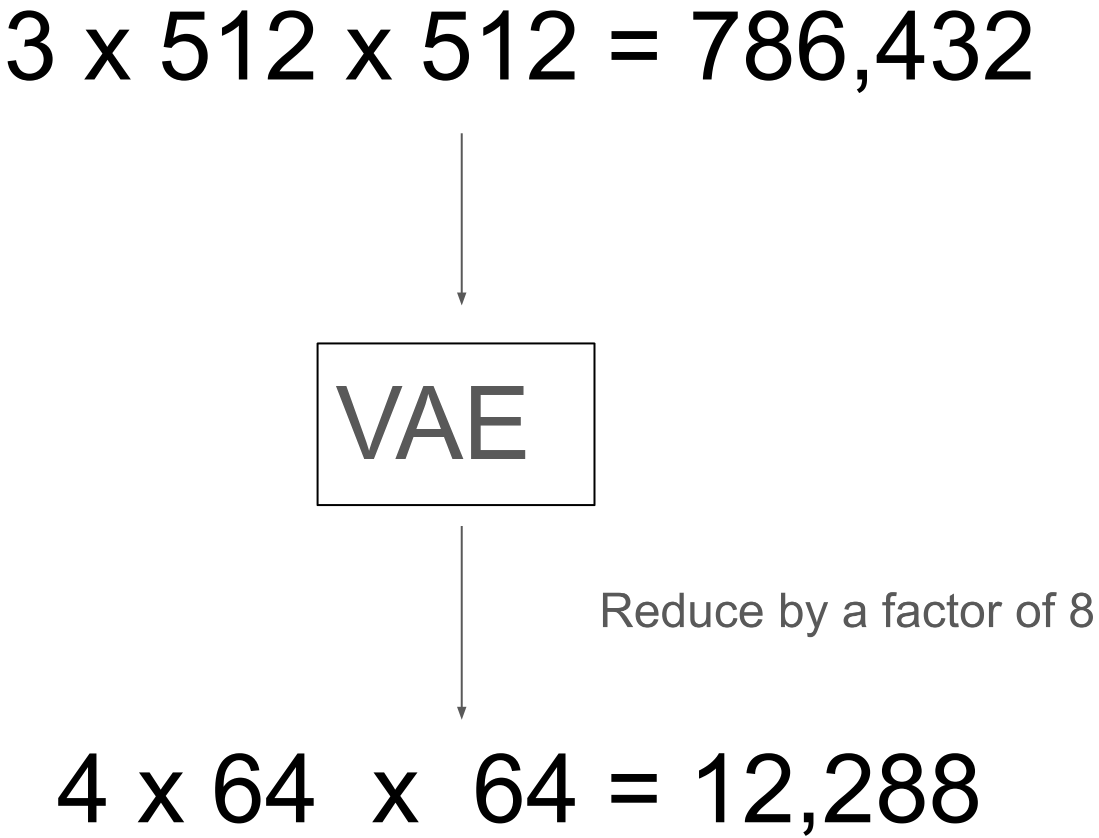
Solution

Making improvements - latent diffusion

Latent diffusion

- Scaling
- Images so far have only been 32x32

3 x 512 x 512 = 786,432



3 x 512 x 512 = 786,432

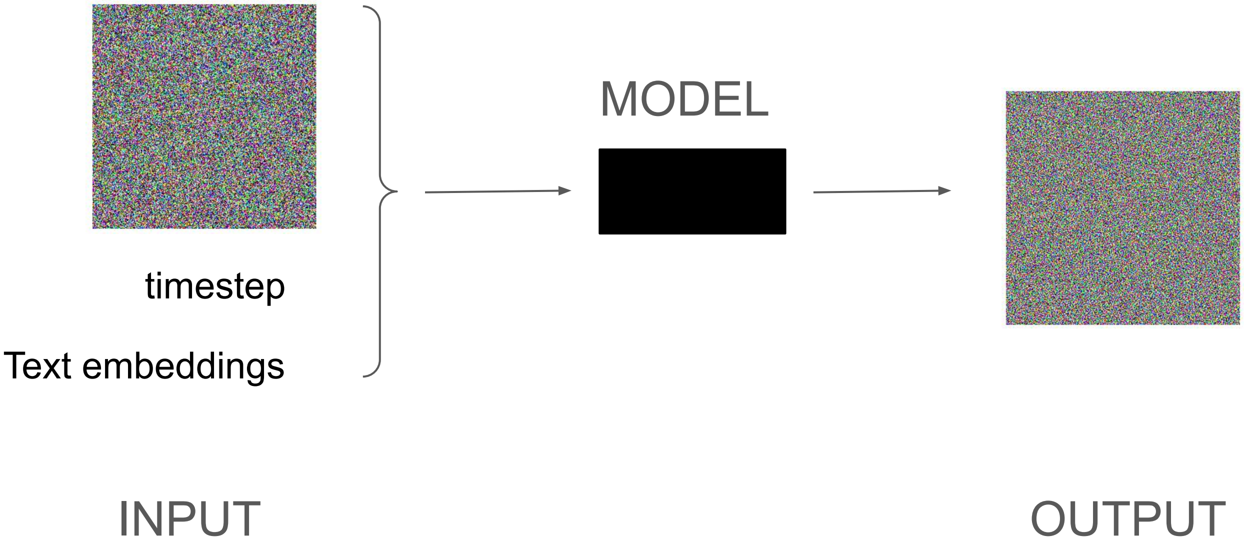
VAE

Reduce by a factor of 8

4 x 64 x 64 = 12,288

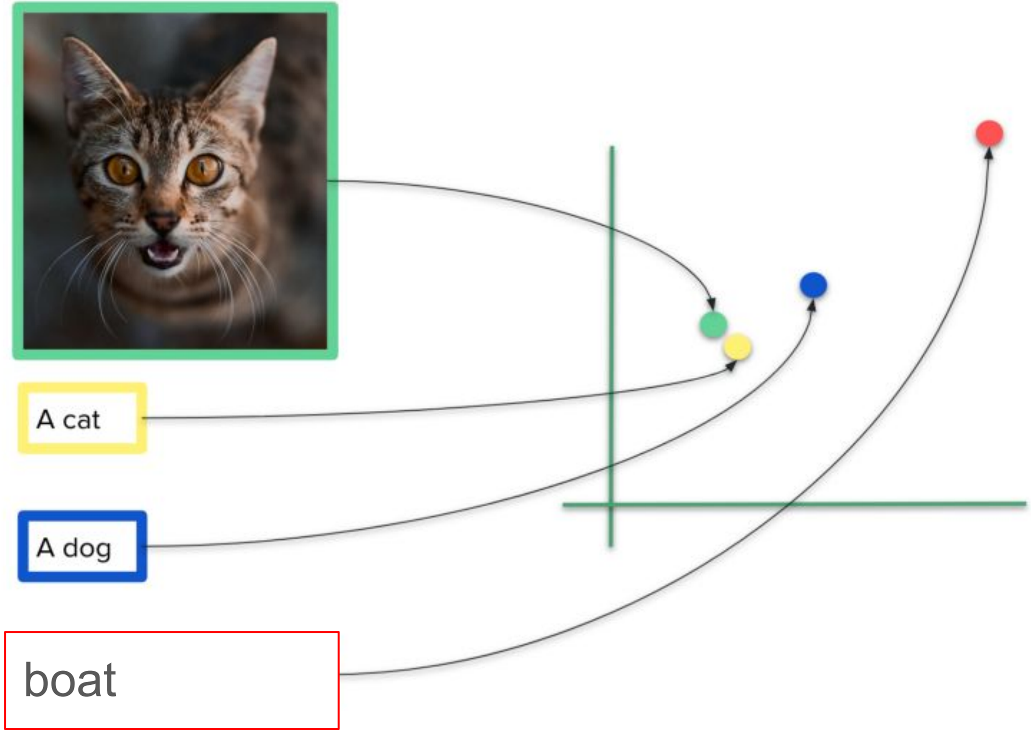
Text encoder - CLIP

Diffusion models (inference)



https://openai.com/index/clip/

CLIP



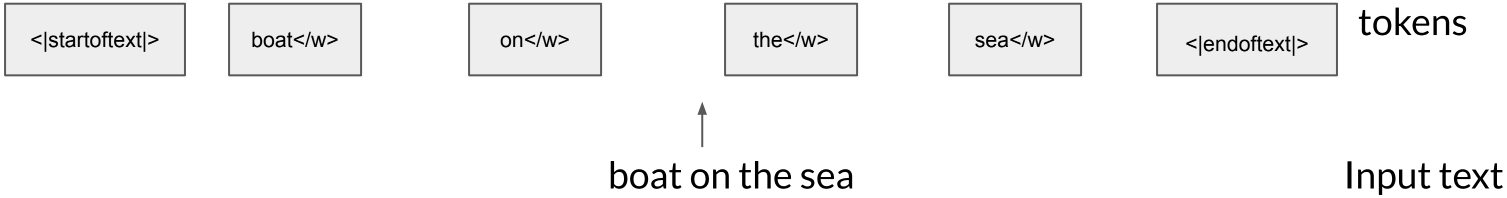
A cat

A dog

boat

boat on the sea

CLIPTextModel



<|startoftext|>

boat</w>

on</w>

the</w>

sea</w>

<|endoftext|>

tokens

boat on the sea

Input text

CLIPTextModel

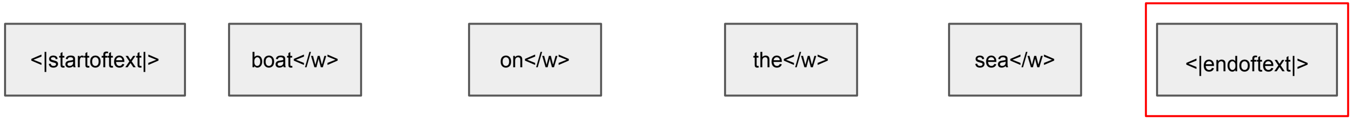
<|startoftext|> boat</w> on</w> the</w> sea</w> <|endoftext|>

tokens

boat on the sea

Input text

CLIPTextModel



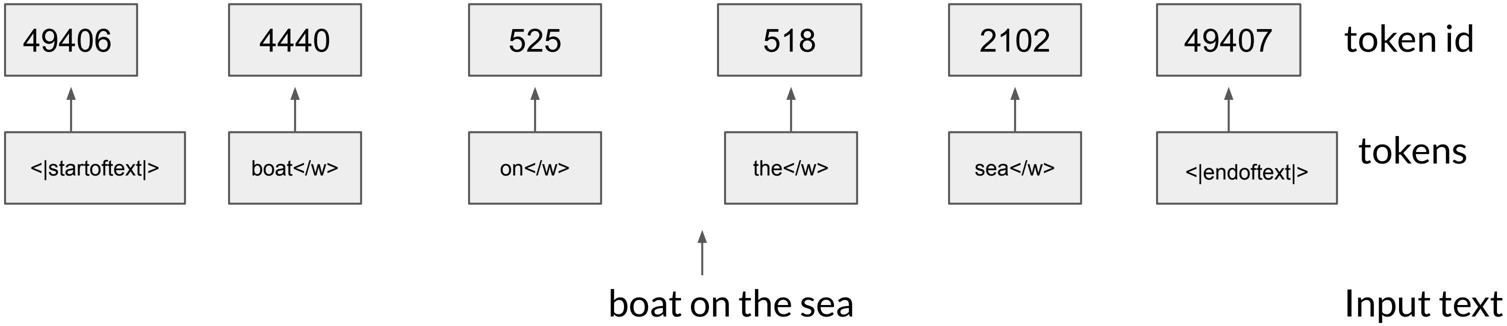
<|startoftext|> boat</w> on</w> the</w> sea</w> <|endoftext|>

tokens

boat on the sea

Input text

CLIPTextModel

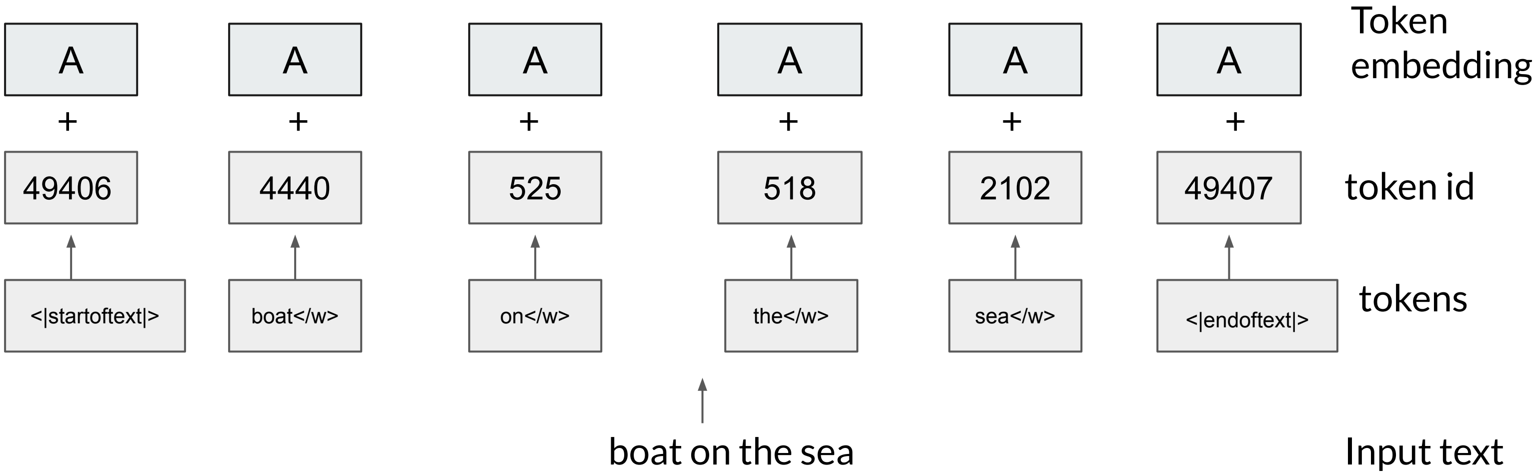


49406	4440	525	518	2102	49407	token id
<|startoftext|>	boat</w>	on</w>	the</w>	sea</w>	<|endoftext|>	tokens

boat on the sea

Input text

CLIPTextModel



A + 49406 <|startoftext|>

A + 4440 boat</w>

A + 525 on</w>

A + 518 the</w>

A + 2102 sea</w>

A + 49407 <|endoftext|>

Token embedding

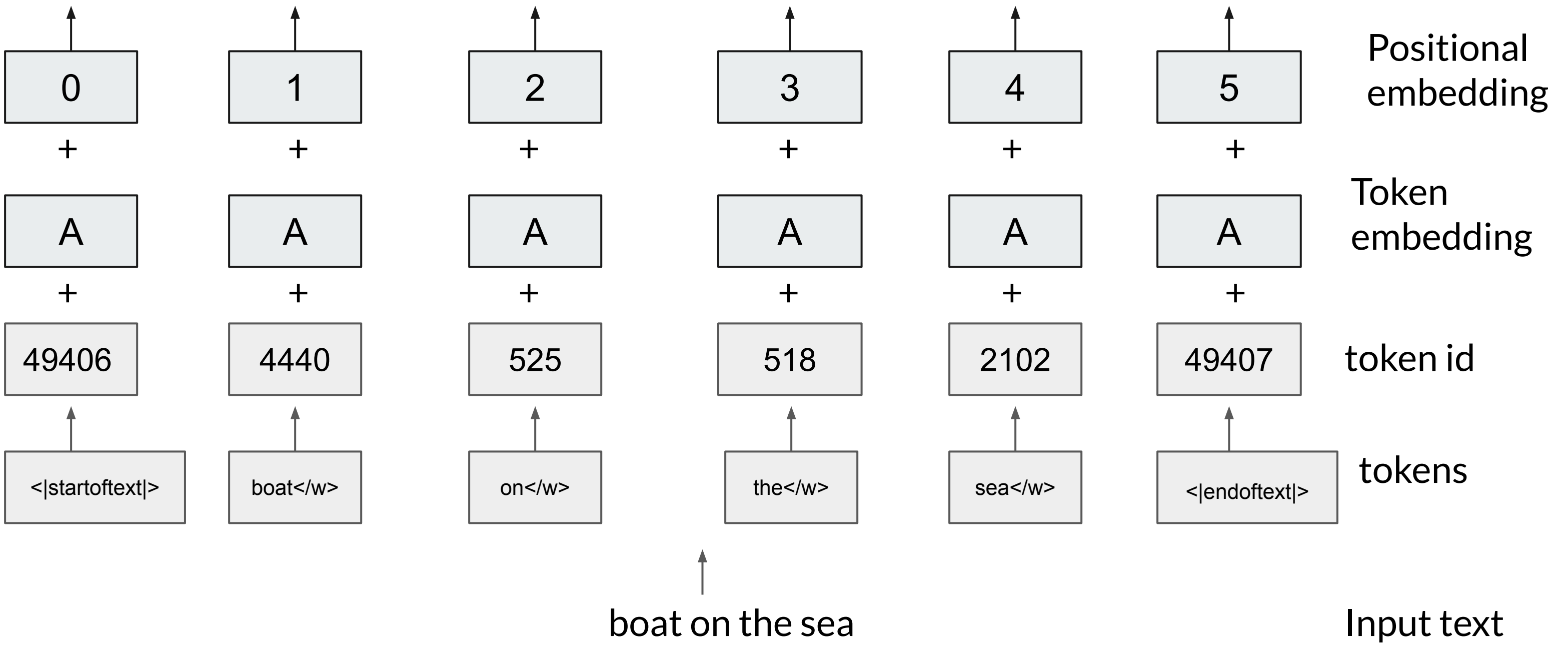
token id

tokens

boat on the sea

Input text

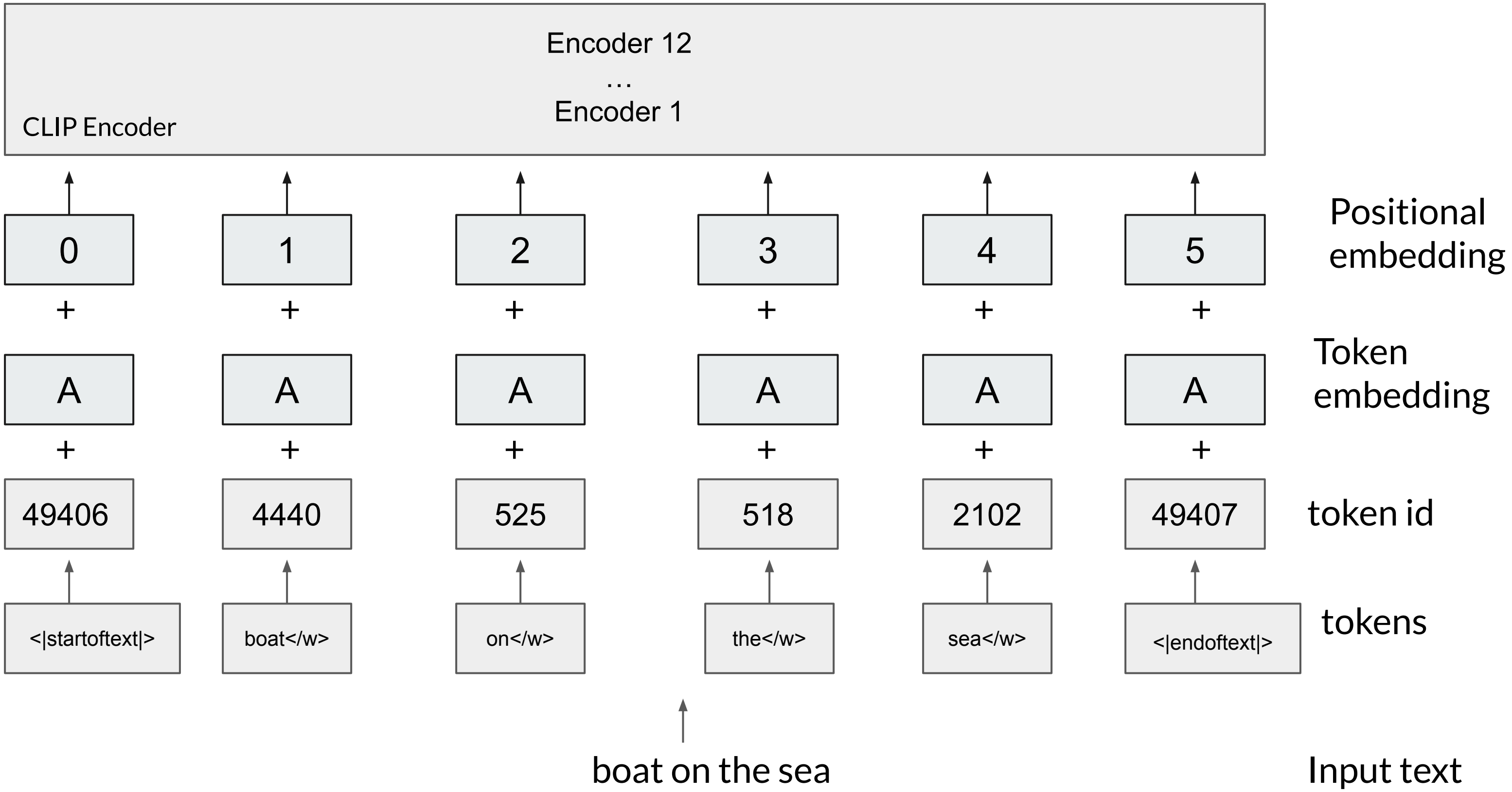
CLIPTextModel



Positional embedding	0	1	2	3	4	5
	+	+	+	+	+	+
Token embedding	A	A	A	A	A	A
	+	+	+	+	+	+
token id	49406	4440	525	518	2102	49407
tokens	<|startoftext|>	boat</w>	on</w>	the</w>	sea</w>	<|endoftext|>

Input text: boat on the sea

CLIPTextModel



Encoder 12

…

Encoder 1

CLIP Encoder

Positional embedding

0 + A + 49406 <|startoftext|>

1 + A + 4440 boat</w>

2 + A + 525 on</w>

3 + A + 518 the</w>

4 + A + 2102 sea</w>

5 + A + 49407 <|endoftext|>

Token embedding

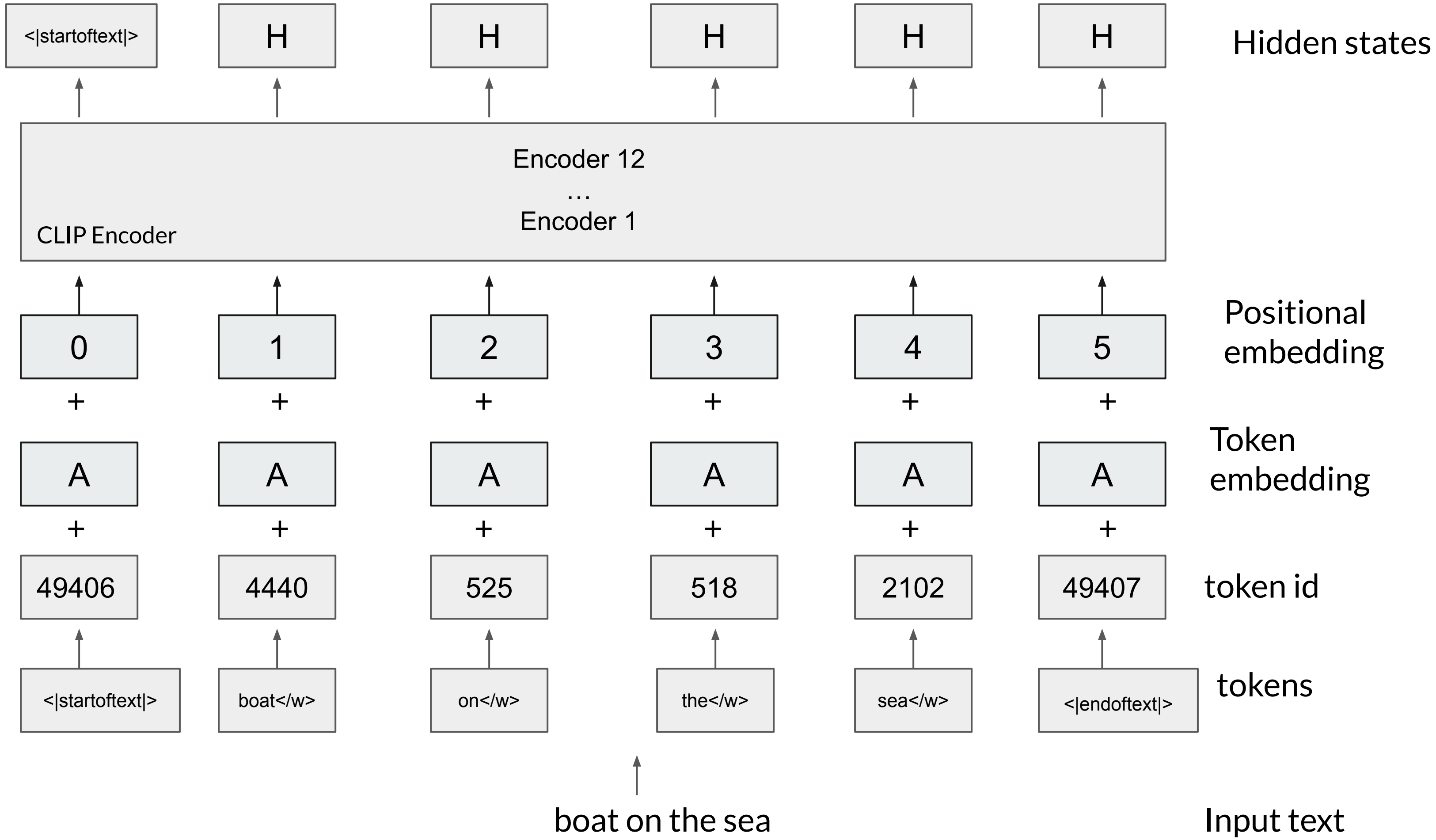
token id

tokens

boat on the sea

Input text

CLIPTextModel



Hidden states	<|startoftext|>	H	H	H	H	H

CLIP Encoder

Encoder 12

. . .

Encoder 1

Positional embedding	0	1	2	3	4	5
	+	+	+	+	+	+
Token embedding	A	A	A	A	A	A
	+	+	+	+	+	+
token id	49406	4440	525	518	2102	49407
tokens	<|startoftext|>	boat</w>	on</w>	the</w>	sea</w>	<|endoftext|>

boat on the sea

Input text

Text encoder in practice

Notebook: CLIP model

Putting it all together using Stable Diffusion

Bias, Limitations and Controversy

- Copyright and Intellectual Property Infringement

- Copyright and Intellectual Property Infringement
- Bias and Discrimination

- Copyright and Intellectual Property Infringement
- Bias and Discrimination
- Deep fakes and miinfornation

- Copyright and Intellectual Property Infringement
- Bias and Discrimination
- Deep fakes and miinfornation
- Privacy and consent

- Copyright and Intellectual Property Infringement
- Bias and Discrimination
- Deep fakes and miinfornation
- Privacy and consent
- Accessibility and control

Next Steps

END

What makes this different to other text to image solutions?

- DALL-E
- DALLE-2
- DALLE-3
- Imagen

Can run on commodity hardware

Model and training details

Model

Text encoder - CLIP ViT-L/14

UNet = 860M parameter model

Autoencoder - downsampling factor of 8.

The model was pretrained on 256x256 images and then finetuned on 512x512 images.

Training time

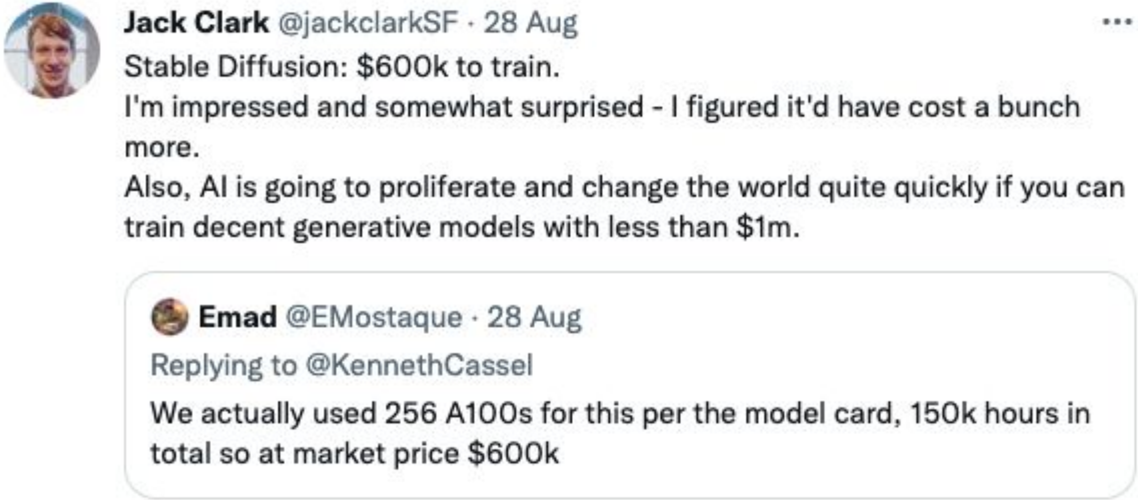
- Hardware Type: A100 PCIe 40GB
- Hours used: 150000

Training data

The core dataset was trained on LAION-Aesthetics, a soon to be released subset of LAION 5B.

LAION-Aesthetics was created with a new CLIP-based model that filtered LAION-5B based on how “beautiful” an image was, building on ratings from the alpha testers of Stable Diffusion.

Cost



Jack Clark @jackclarkSF · 28 Aug

Stable Diffusion: $600k to train.
I'm impressed and somewhat surprised - I figured it'd have cost a bunch more.
Also, AI is going to proliferate and change the world quite quickly if you can train decent generative models with less than $1m.

Emad @EMostaque · 28 Aug
Replying to @KennethCassel

We actually used 256 A100s for this per the model card, 150k hours in total so at market price $600k

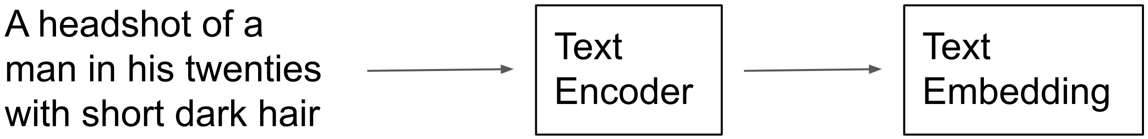
Controversy

- Image regurgitation
- Copying artist styles
    - Getty Images
    - Shutterstock

Applications

https://www.reddit.com/r/StableDiffusion/comments/wyduk1/show_rstablediffusion_integrating_sd_in_photoshop/

What components do we need?

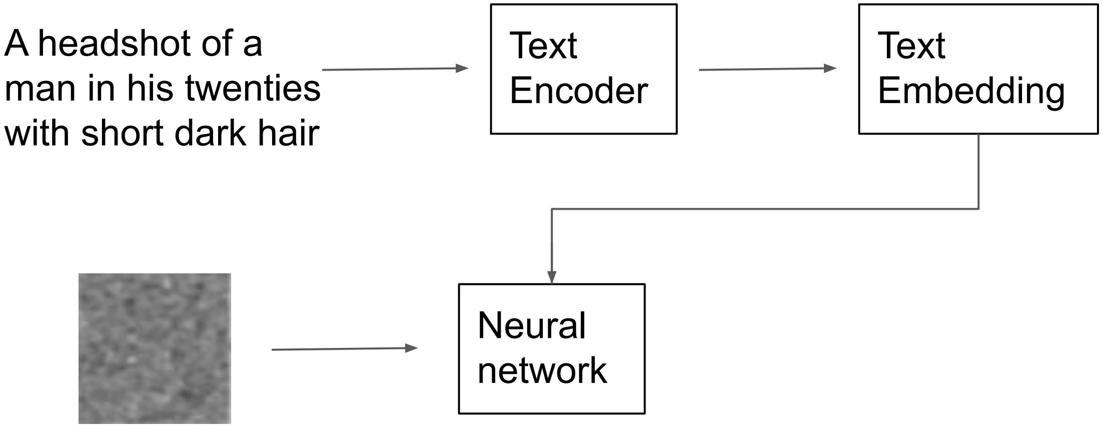


A headshot of a man in his twenties with short dark hair

Text Encoder

Text Embedding

What components do we need?



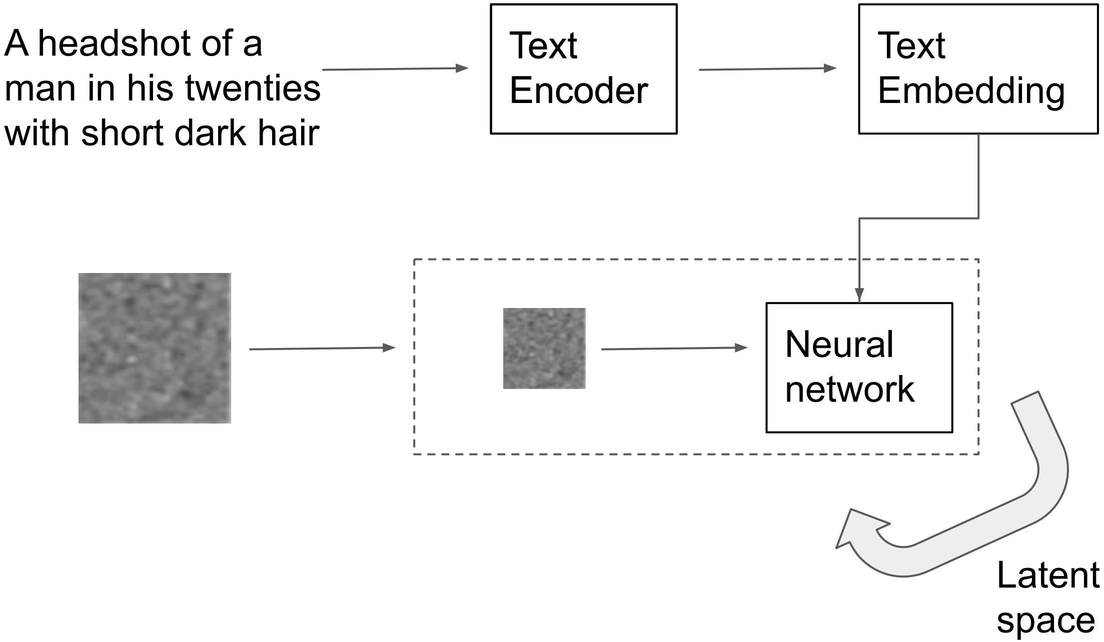
A headshot of a man in his twenties with short dark hair

Text Encoder

Text Embedding

Neural network

What components do we need?



A headshot of a man in his twenties with short dark hair

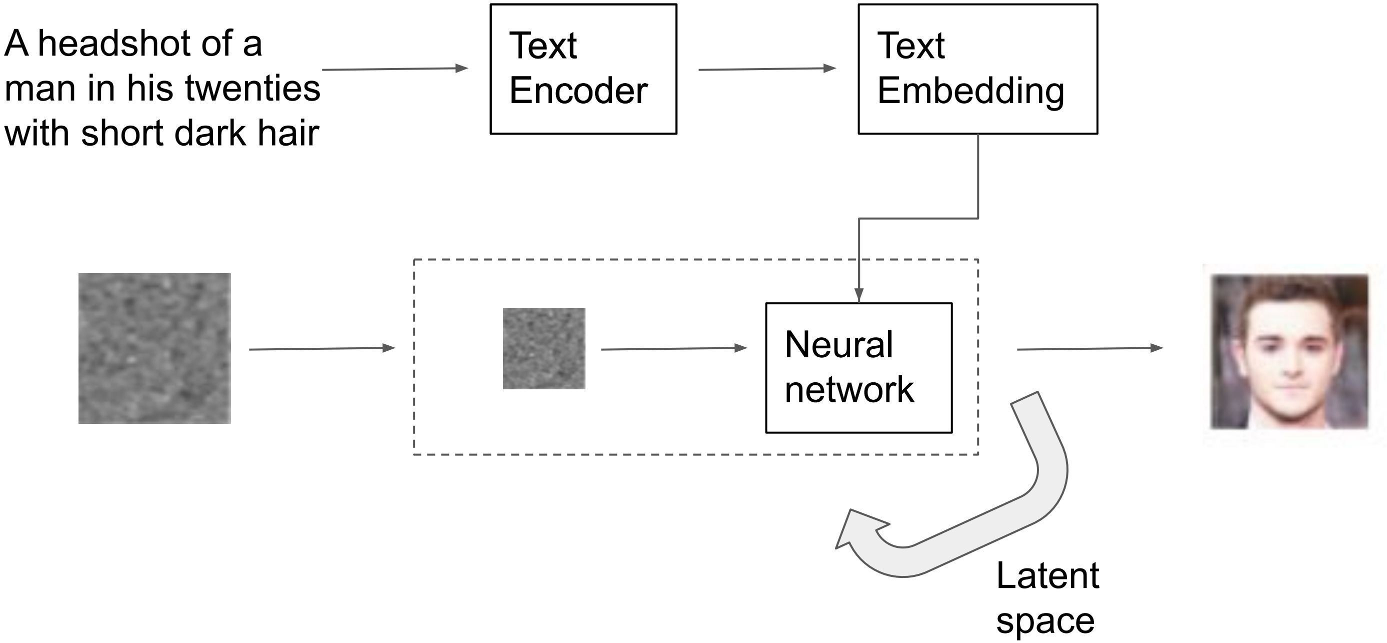
Text Encoder

Text Embedding

Neural network

Latent space

What components do we need?



A headshot of a man in his twenties with short dark hair

Text Encoder

Text Embedding

Neural network

Latent space

The key difference between latent and standard diffusion is that latent diffusion model is trained to generate latent (compressed) representations of the images

What 3 components do we need for latent diffusion?

- A text encoder (CLIP’s Text Encoder)

What 3 components do we need for latent diffusion?

- A text encoder - CLIP’s Text Encoder
- Neural network - UNet

U-Net

+ Time encoding

U-Net

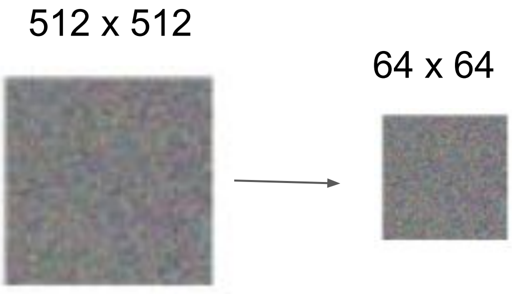
+ Time encoding + Text encoding = Conditioned Image

What 3 components do we need for stable diffusion?

- A text encoder - CLIP’s Text Encoder
- Neural network - UNet
- Autoencoder

Autoencoder

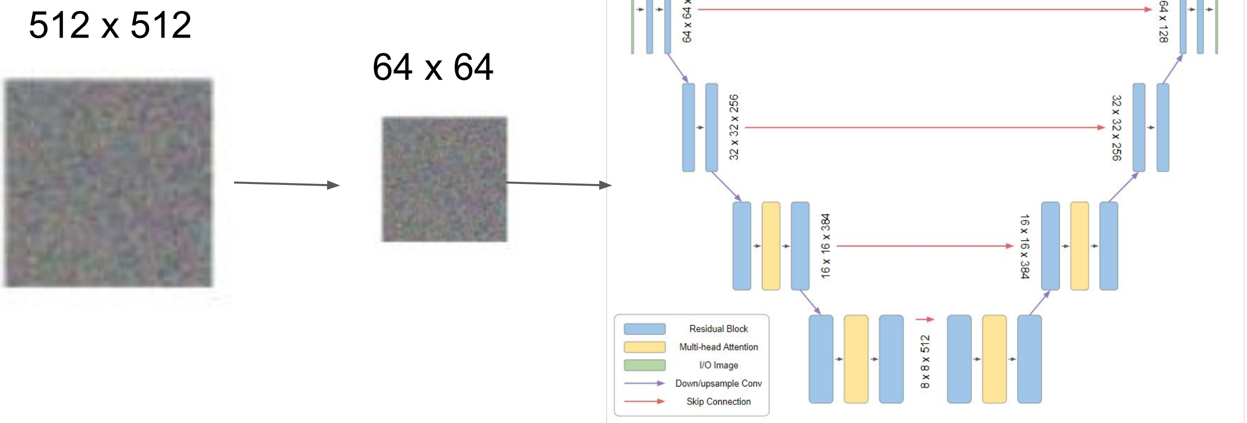
Autoencoder



512 x 512

64 x 64

Autoencoder



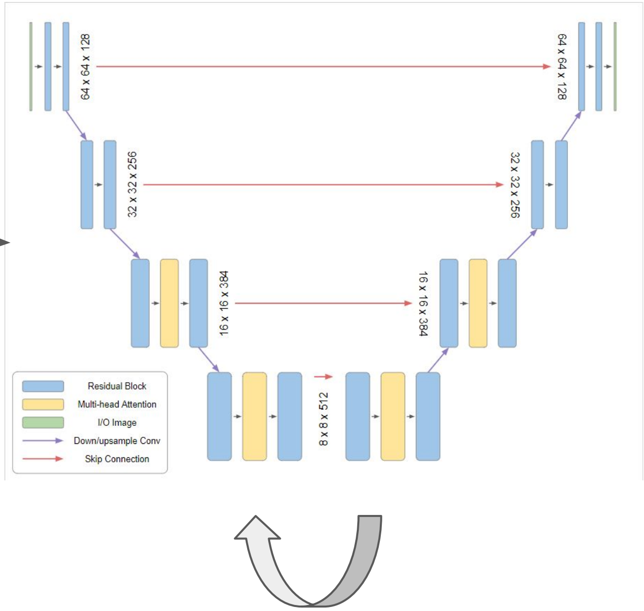
512 x 512

64 x 64

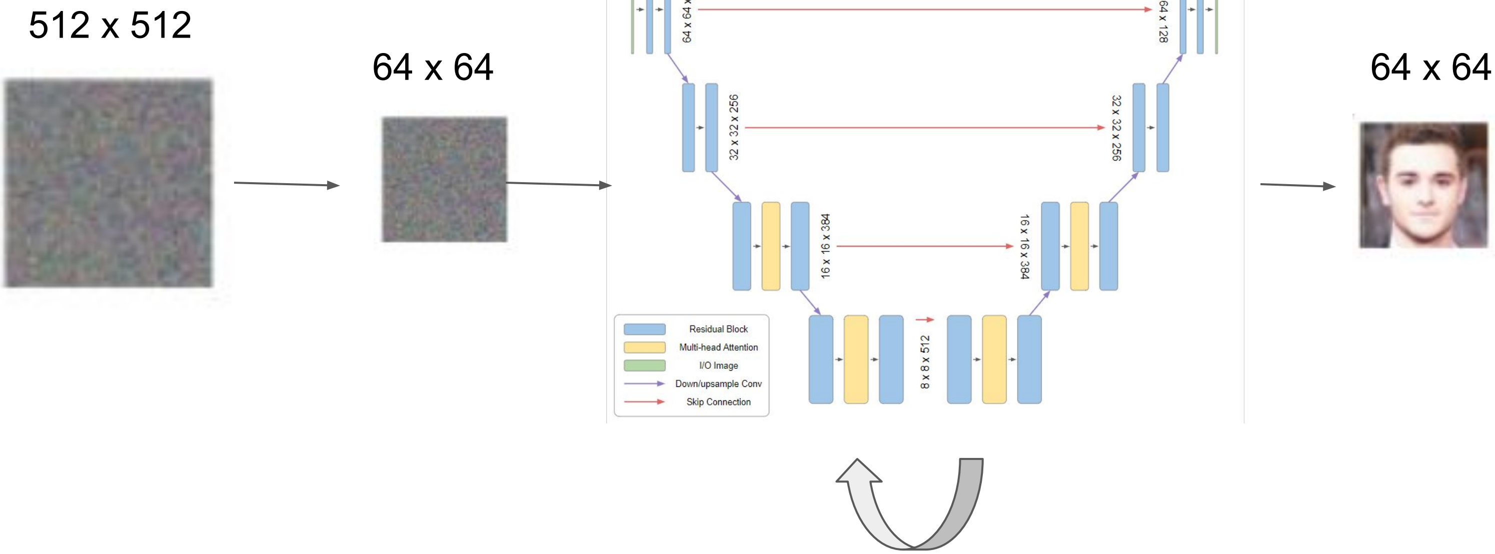
Autoencoder

512 x 512

64 x 64



Autoencoder



512 x 512

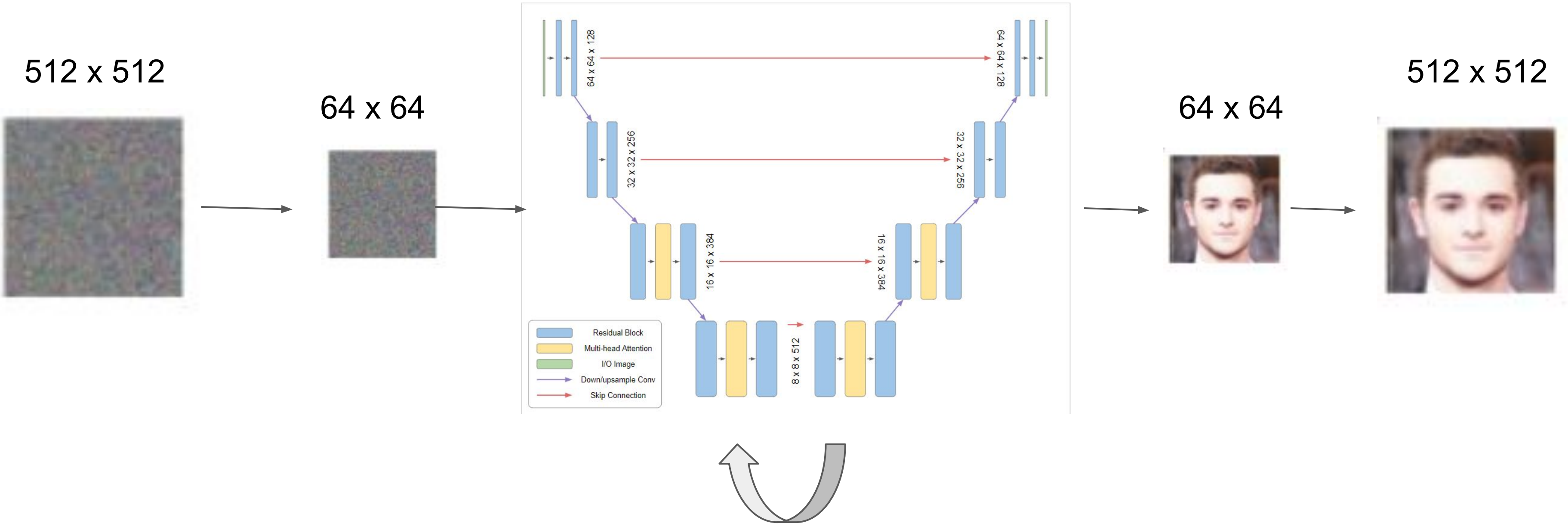
64 x 64

64 x 64

Autoencoder

512 x 512

64 x 64



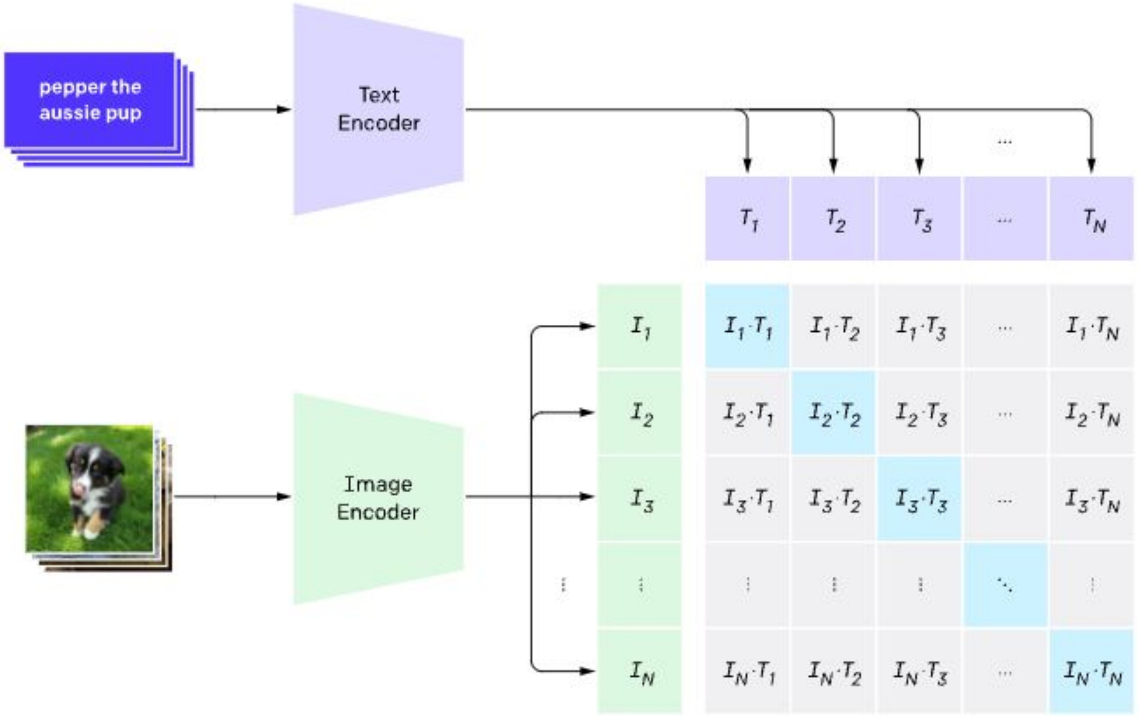
64 x 64

512 x 512

Colab notebook - autoencoder

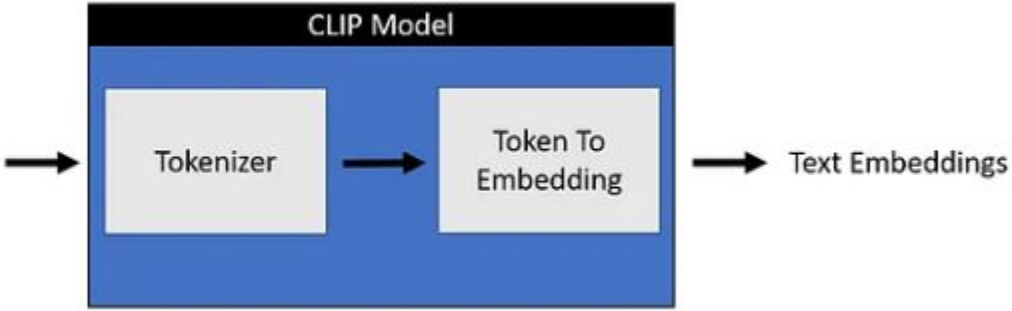
Text encoder - CLIP

CLIP - Contrastive pre-training


pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
⋮ ⋮ ⋮ ⋱ ⋮
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$

Source: https://openai.com/blog/clip/

Prompt: Speedboat in the sea



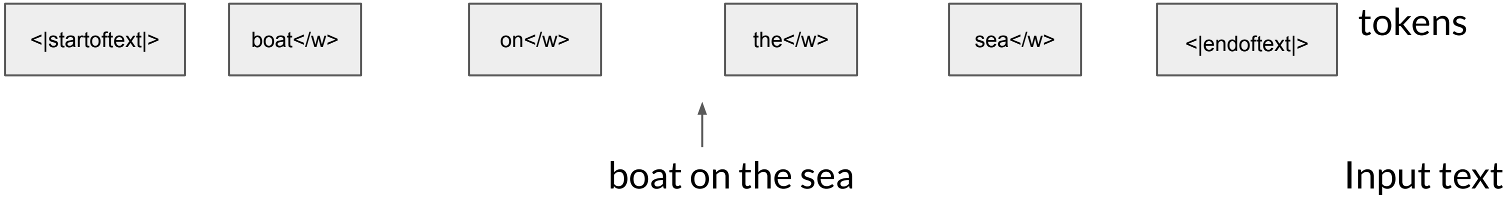
Colab - CLIP Model

Colab - tokenizers

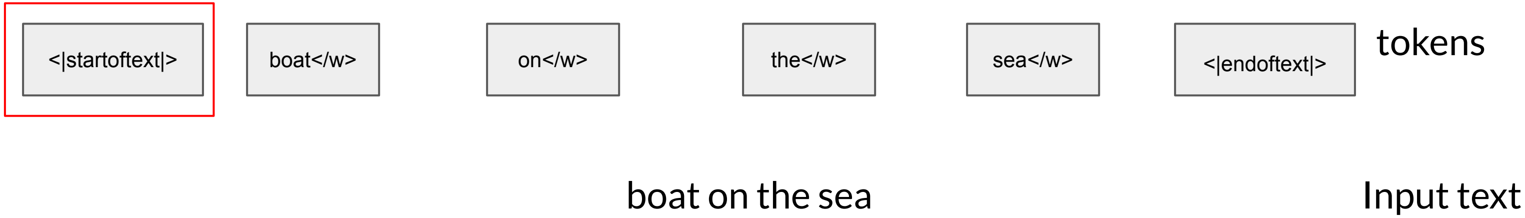
Text encoders

boat on the sea

CLIPTextModel



CLIPTextModel



<|startoftext|>

boat</w>

on</w>

the</w>

sea</w>

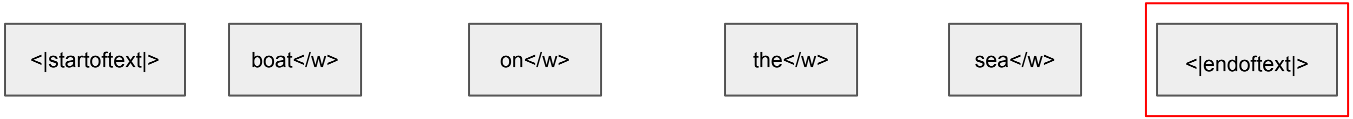
<|endoftext|>

tokens

boat on the sea

Input text

CLIPTextModel



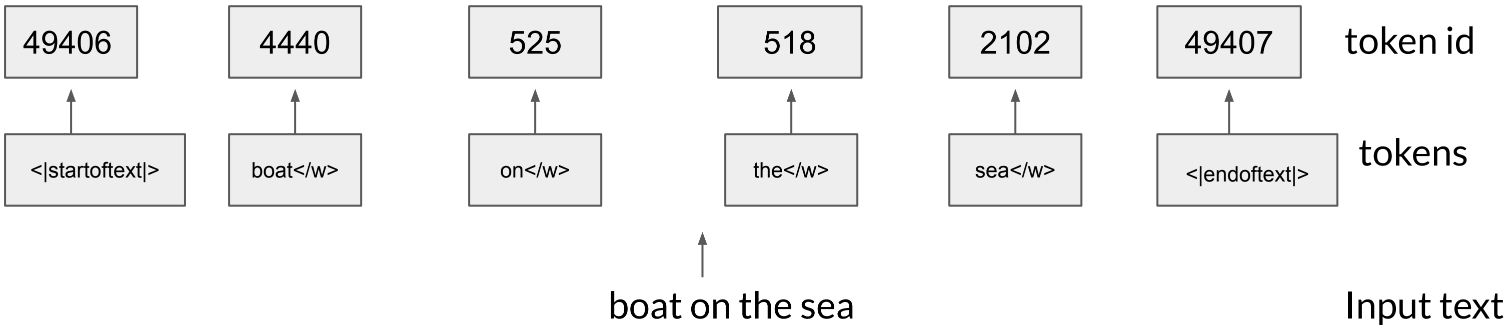
<|startoftext|> boat</w> on</w> the</w> sea</w> <|endoftext|>

tokens

boat on the sea

Input text

CLIPTextModel

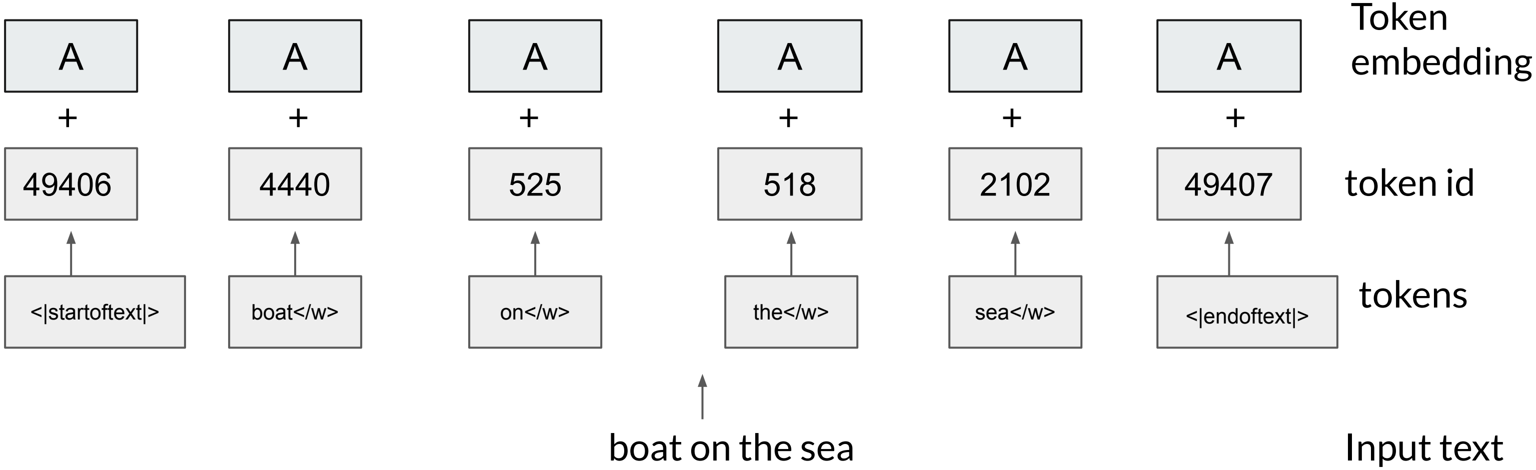


49406	4440	525	518	2102	49407	token id
<|startoftext|>	boat</w>	on</w>	the</w>	sea</w>	<|endoftext|>	tokens

boat on the sea

Input text

CLIPTextModel

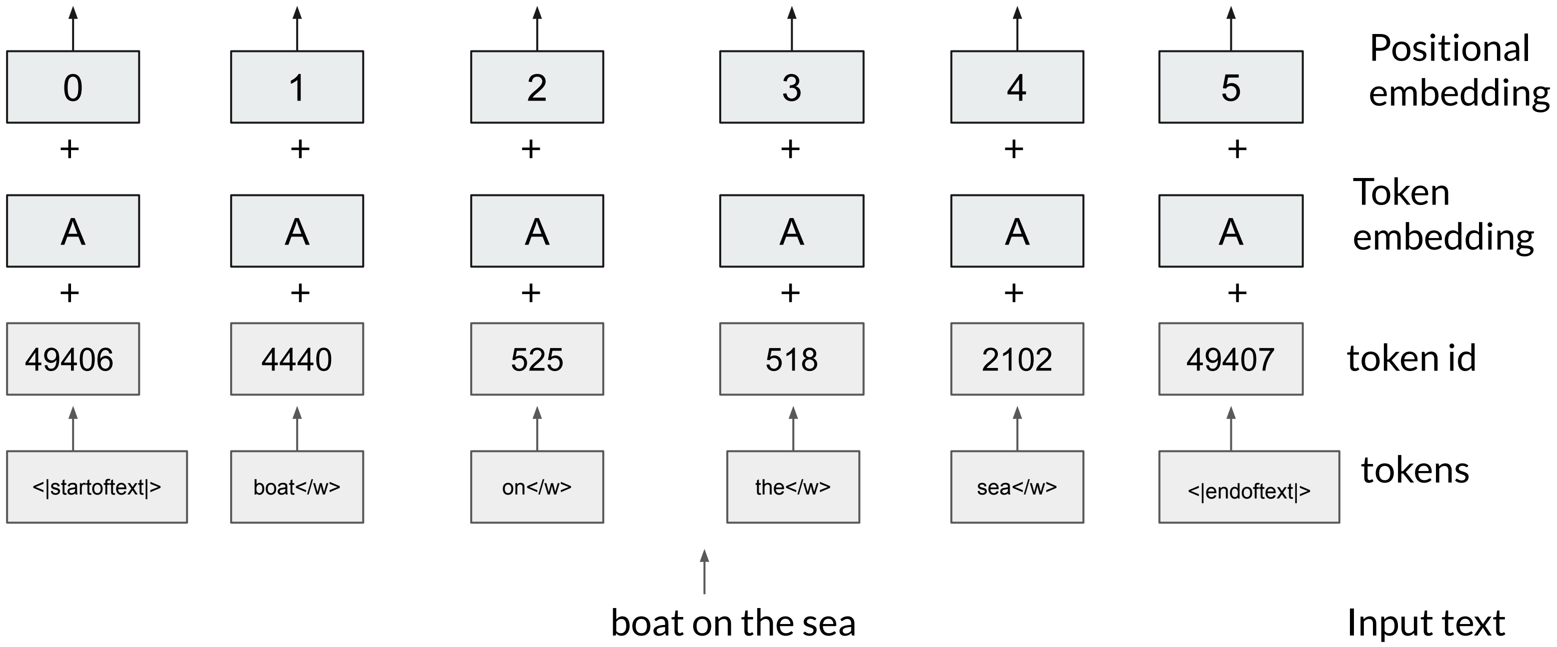


A	A	A	A	A	A	Token embedding
+	+	+	+	+	+	
49406	4440	525	518	2102	49407	token id
<|startoftext|>	boat</w>	on</w>	the</w>	sea</w>	<|endoftext|>	tokens

boat on the sea

Input text

CLIPTextModel

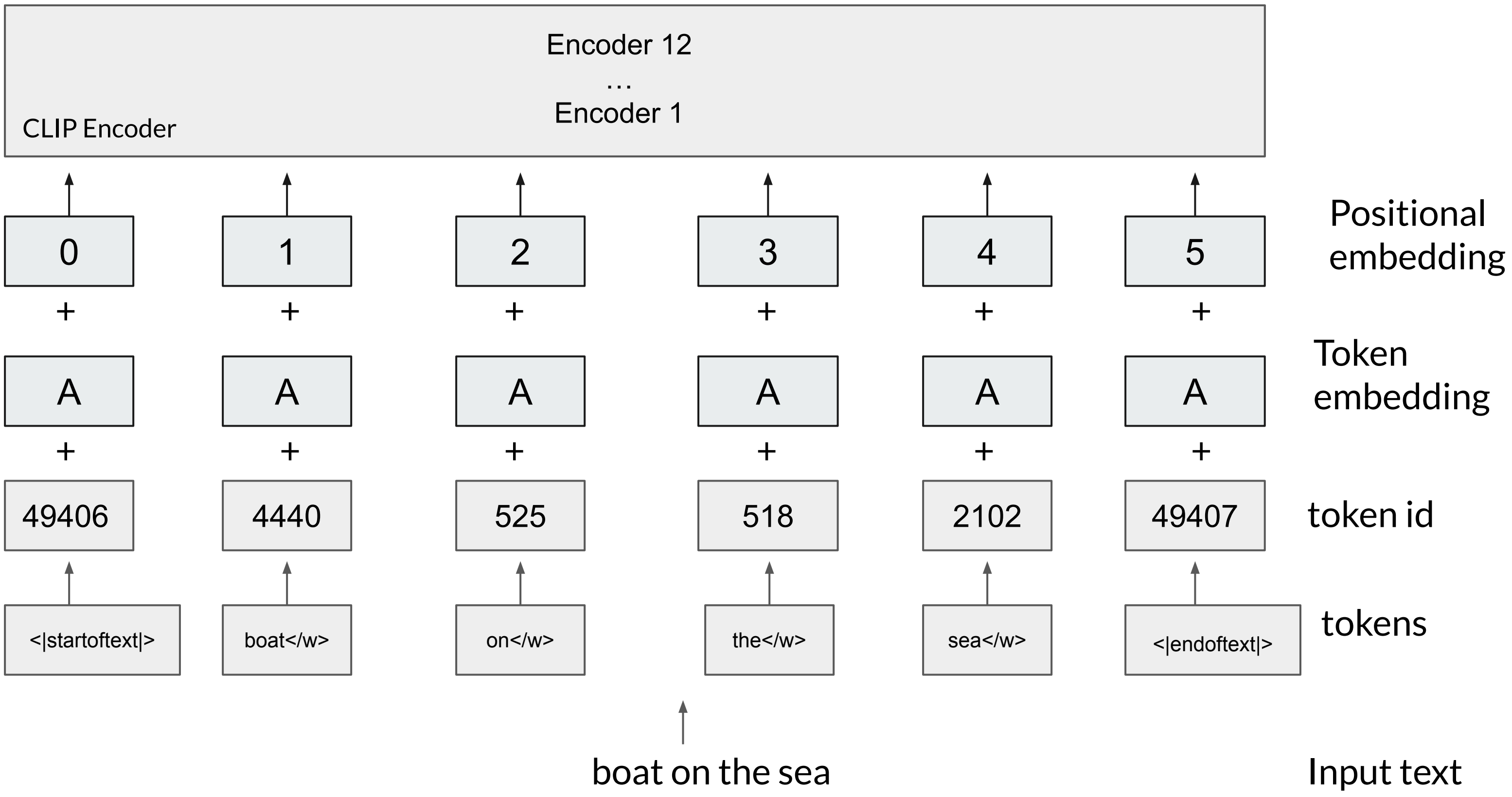


0	1	2	3	4	5	Positional embedding
+	+	+	+	+	+	
A	A	A	A	A	A	Token embedding
+	+	+	+	+	+	
49406	4440	525	518	2102	49407	token id
<|startoftext|>	boat</w>	on</w>	the</w>	sea</w>	<|endoftext|>	tokens

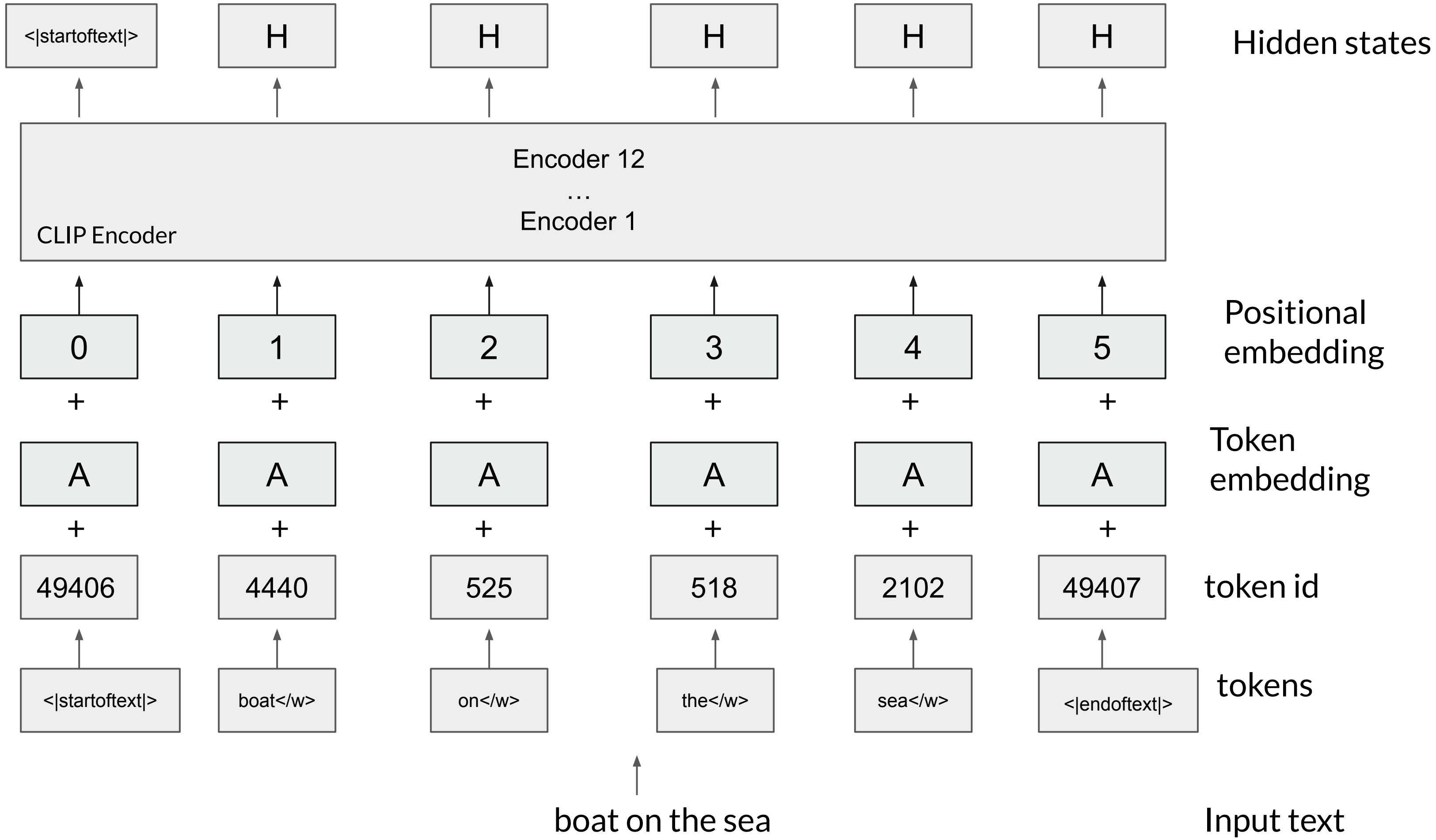
boat on the sea

Input text

CLIPTextModel



CLIPTextModel



Colab - text encoders

END