

From Text to Image

Multimodality in Deep Learning

Antoine Simoulin

Friday November 4th

QuantumBlack Data Science Coffee
External Speaker Series

1.0. About Me



Antoine SIMOULIN

I am an **engineer by education and experience**

I have held **academic and industry roles**, giving me a fresh perspective on translating technical expertise into solutions for complex problems.

For the last five years, I worked as a **Sr. Data Scientist at Quantmetry within the NLP research and innovation team**

As a researcher, I conducted cutting-edge research and **completed my Ph.D. within the laboratory of formal linguistics at Paris University,**

Introduction

1.1. Deep Learning for multimodal data

Image generation conditioned by text

Multimodal deep learning architectures can handle **multiple data modalities**, including **text, image, audio, speech, video, and tabular data**. We can distinguish **various architectures**, each combining modalities with a different setup.

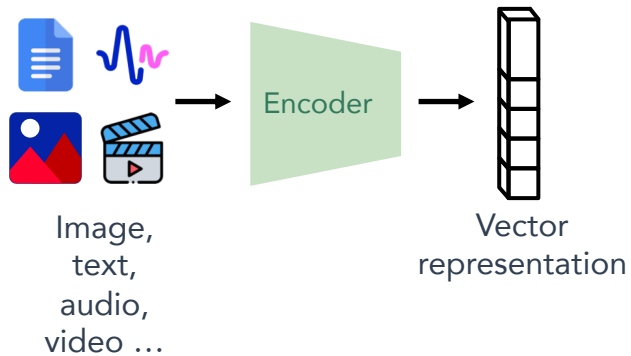


Figure 1. The encoder takes input from one or multiple modalities and outputs a vector representation. This is a fusion-based architecture.

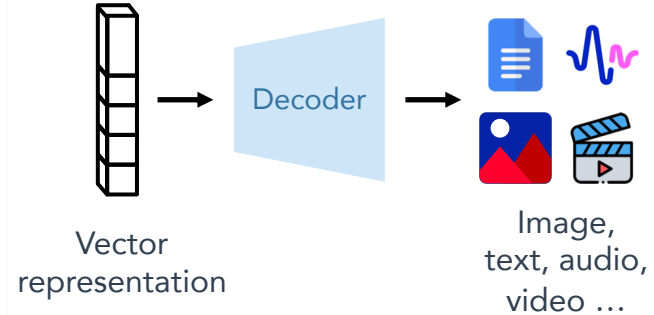


Figure 2. The decoder takes a vector representation input and outputs one or multiple modalities. This is a fusion-based architecture.

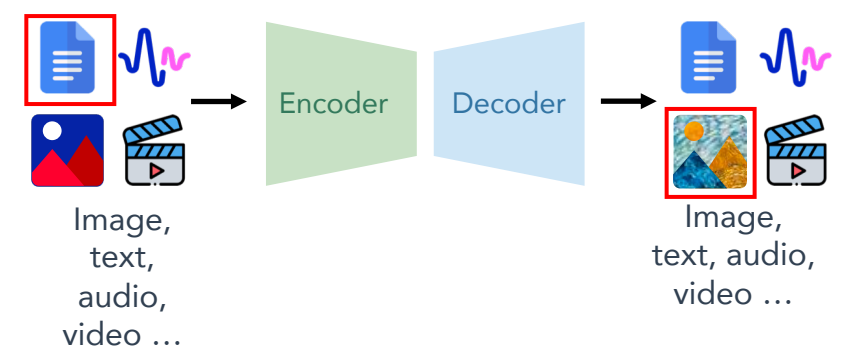


Figure 3. Encoder-decoder takes one or multiple modalities as input and outputs one or multiple different or identical modalities.

1.1. Deep Learning for multimodal data

Technology maturity



Multiple applications, use-cases, startups, open-source contributions, and resources exist.



There has been **steady progress in technology**.



Unstructured data represent a **large portion of available company data**.

Hype Cycle for Artificial Intelligence, 2022



Figure 1. Position of the generative AI technology on the Gartner hype cycle for artificial intelligence 2022.

1.1. Deep Learning for multimodal data

Presentation goal



Give an overview of already **existing applications** and use cases.



Give an overview of the current **underlying technology** and its limits.



Enumerate the **available resources**.



Discuss the potential use cases and **future work direction**.

1.1. Deep Learning for multimodal data

Presentation schedule

Introduction 5'

1 Applications of multimodal text/image neural networks 15'

2 Technical background 15'

3 Available resources and tools 5'

Conclusion and future work 5'

#1

Applications of multimodal text/image
neural networks

1.1. Multimodal search engines

Google Search

At the 2021 IO conference, Google introduced **MUM**, a new model based on the T5 text-to-text architecture [2]. According to them, the model is 1000 times more powerful than BERT [1]. MUM allows **encoding and generating language** and is trained in 75 languages and several tasks simultaneously. Finally, MUM is multimodal, for the moment, through **text, image**, and, in the future, possibly video and audio.



Figure 1. The features of MUM summarize well the current trends in NLP: (1) encoding and generation, (2) multilingual, (3) multi-task trained, and (4) multimodal.

Caption: *A skier is skiing down the snow wearing a white shirt and black shorts.*

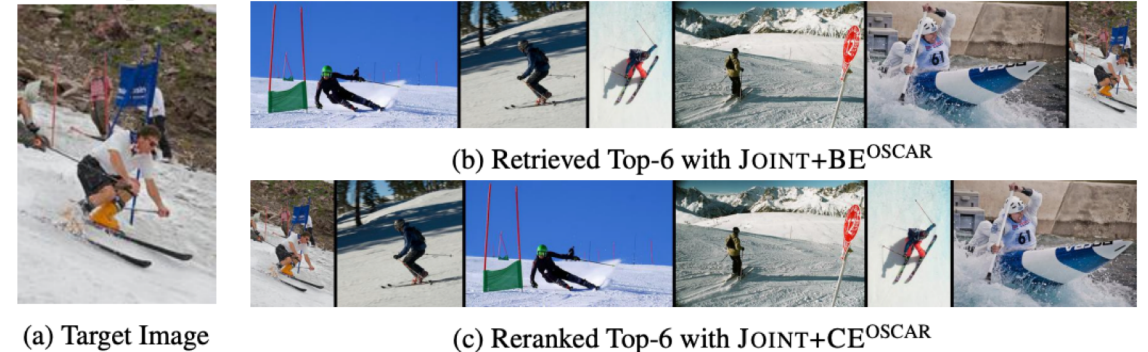


Figure 2. Multimodal search engine [3].

[1] <https://blog.google/products/search/introducing-mum/>

[2] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu: [Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer](#). J. Mach. Learn. Res. 21: 140:1-140:67 (2020)

[3] Gregor Geigle, Jonas Pfeiffer, Nils Reimers, Ivan Vulic, Iryna Gurevych: [Retrieve Fast, Rerank Smart: Cooperative and Joint Approaches for Improved Cross Modal Retrieval](#). CoRR abs/2103.11920 (2021) 9

1.1. Image edition

Google Magic Eraser

In 2021, Google introduced the **Magic Eraser**, which can automatically **figure out what you're trying to get rid of** by analyzing what you circle, so you don't have to worry about precise brushing. Once you decide what to erase, Magic Eraser uses machine learning again to predict what pixels would look like if the distraction were not there. The tool can run **directly on the device** and enable **subtle image editing without technical knowledge**.



Figure 1. Illustration of the google magic eraser tool.

[1] <https://blog.google/products/photos/magic-eraser/>

1.1. Video Edition using text

Runway editor

Runway [2] is a company developing **software to generate content** such as images and videos **using natural language text input**.

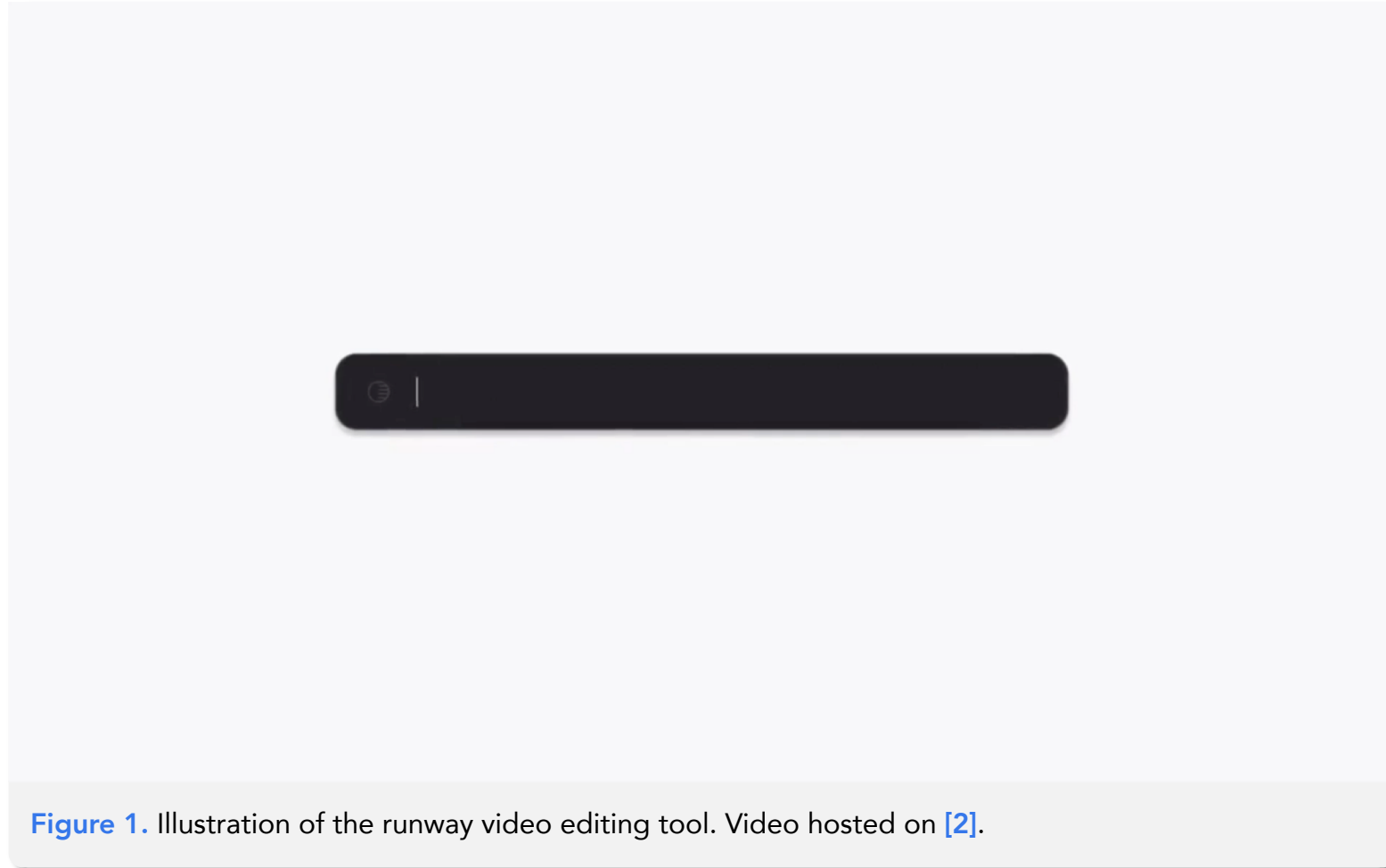


Figure 1. Illustration of the runway video editing tool. Video hosted on [2].

[1] <https://runwayml.com/>

[2] <https://twitter.com/runwayml/status/1568220303808991232?s=20&t=xVzWrPMm9jDz6YDFRTPkTw>

1.1. General Artificial Intelligence Research

DeepMind Gato model

DeepMind recently released a new **multimodal AI system: "Gato"** [1]. The algorithm is trained on many tasks (several hundred): from controlling a robot arm to writing poems. DeepMind called it a "general-purpose" system.

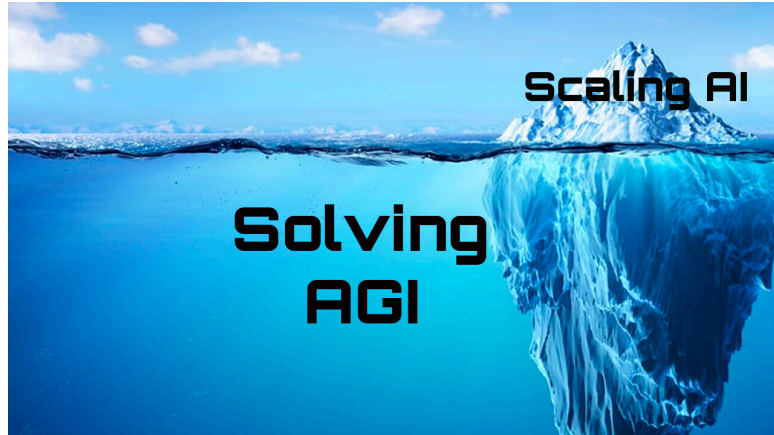


Figure 1. DeepMind researcher claims new 'Gato' AI could lead to AGI, says 'the game is over!' (thenextweb.com)



Figure 1 | A generalist agent. Gato can sense and act with different embodiments across a wide range of environments using a single neural network with the same set of weights. Gato was trained on 604 distinct tasks with varying modalities, observations and action specifications.

Figure 2. Gato model pre-training paradigm [1].

[1] Scott E. Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, Nando de Freitas: [A Generalist Agent](#). CoRR abs/2205.06175 (2022)

Microsoft LayoutLMv3

Buckley ACCOUNTING SERVICES 95 Hart Street Sydney NSW 2174 Bill to: Javlan Romero 88 Green Passage Sydney NSW 2355	Ship to: Cristina Browning 12 Beacon Avenue Sydney NSW 2095	Invoice Invoice #: 37220207 Invoice Date: 12/27/2021 ABN: 92670378 Due Date: 12/27/2021
---	--	--

Qty	Description	Unit Price	Amount
5	toe ring	77	385
4	eye liner	87	348
5	plastic fork	28	140
4	packing peanuts	72	288
2	checkbook	12	24
4	packing peanuts	72	288
5	puddle	86	430
3	desk	75	225
5	spoon	26	130
5	shoes	29	145
Subtotal			2,093.00
GST			240.30
Total			2,333.30

[illegible]

[1] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, Furu Wei: [LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking](#). ACM Multimedia 2022: 4083-4091

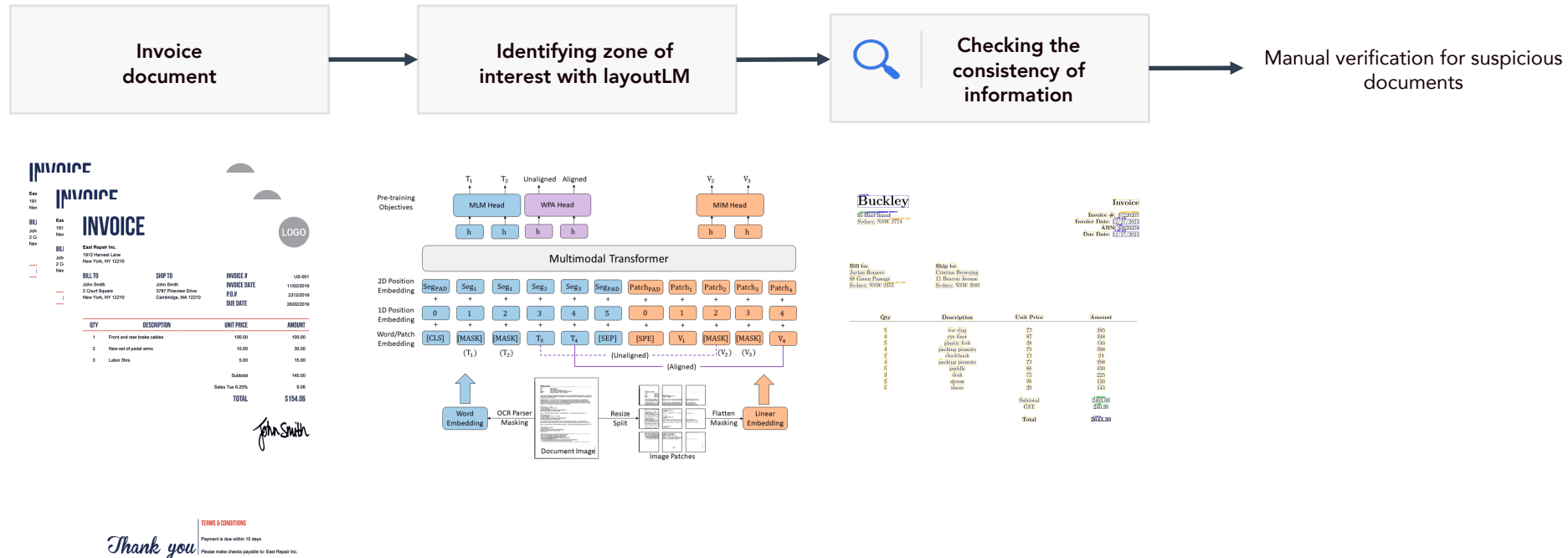
[2] <https://twitter.com/osanseviero/status/1536374736590057474?s=20&t=xVzWrPMm9jDz6YDFRTPkTw>

[3] <https://twitter.com/philipvollet/status/1565334436702347266?s=20&t=xVzWrPMm9jDz6YDFRTPkTw>

1.1. Document processing

Microsoft LayoutLMv3 uses cases and application

I have led a project in **identifying forged documents for an insurance company**. Part of the pipeline involved **automatically extracting information from the document and verifying its internal consistency**. For example, I confirmed the sum of the amount matched the detailed listing.



1.1. Multimodal representation

Facebook CommerceMM

Facebook released a new multimodal model, CommerceMM [1], which creates rich representations of commerce data. With state-of-the-art performance in product recognition & retrieval, CommerceMM helps tailor suggestions & search results to connect shoppers with the items they want.

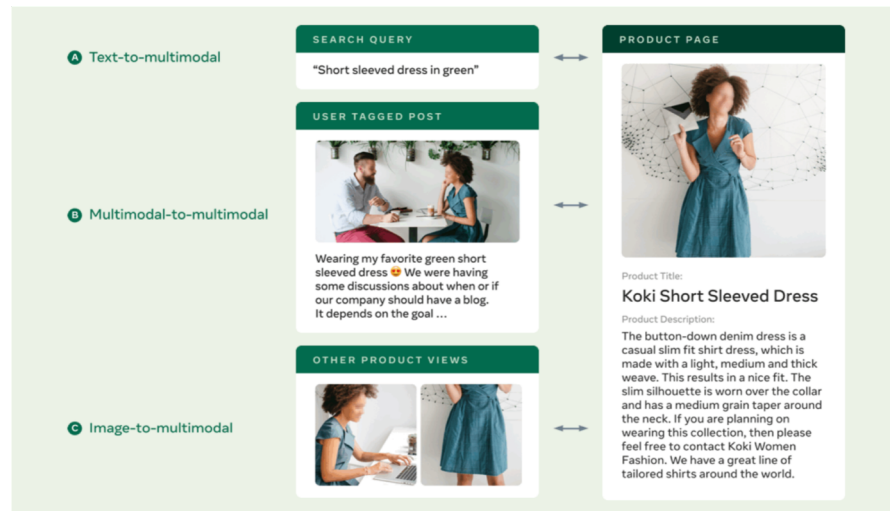


Figure 1. Use Microsoft CommerceMM to “understand” multimodal data by integrating its characterizations of a post’s text and image.

Model Architecture

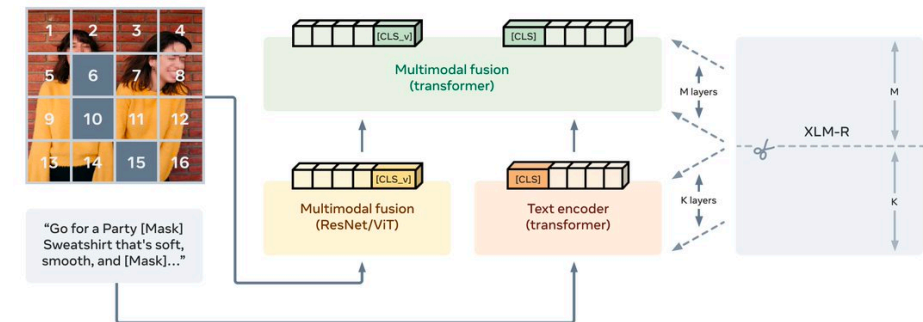


Figure 2. CommerceMM model architecture.

[1] Licheng Yu, Jun Chen, Animesh Sinha, Mengjiao Wang, Yu Chen, Tamara L. Berg, Ning Zhang: [CommerceMM: Large-Scale Commerce MultiModal Representation Learning with Omni Retrieval](#). KDD 2022: 4433-4442

1.1. Inspirational content generation

Architecture design

Using latent text-to-image diffusion, This House Does Not Exist [1] generates realistic images of modern architecture homes.



thishousedoesnotexist.org/11493414

Figure 1. House design generated using the query “A modernist home in Thailand that is eco-friendly and made from sustainable materials”



Figure 2. AI-generated Future Cities by Manas Bhatia using and Midjourney AI, Adobe Photoshop [2].

[1] <https://thishousedoesnotexist.org/>

[2] <https://amazingarchitecture.com/futuristic/ai-generated-future-cities-by-manas-bhatia>

1.1. Open sources release

Stable diffusion

In May 2022, Hugging Face released the **open-source library** 🤗 **Diffusers**, which provides **pre-trained diffusion models** across multiple modalities, such as vision and audio, and serves as a **modular toolbox** for inference and training diffusion models.



Figure 1. Image generation for the prompt « a high tech solarpunk utopia in the Amazon rainforest » using the Hugging Face stable diffusion demo.



Figure 2. Image generation for the prompt « A mecha robot in a favela in expressionist style » using the Hugging Face stable diffusion demo.

[1] <https://github.com/huggingface/diffusers/>

#2

Technical background

1.1. Generative Adversarial Networks (GAN)

Training paradigm

GANs [1] consist of **two parts** trained in an end-to-end fashion. The **generator generates images** out of random noise, while the **discriminator aims to distinguish generated images from authentic images** from a training set. Consequently, the generators need to improve the quality of their generated images as the discriminator improves at discriminating images.

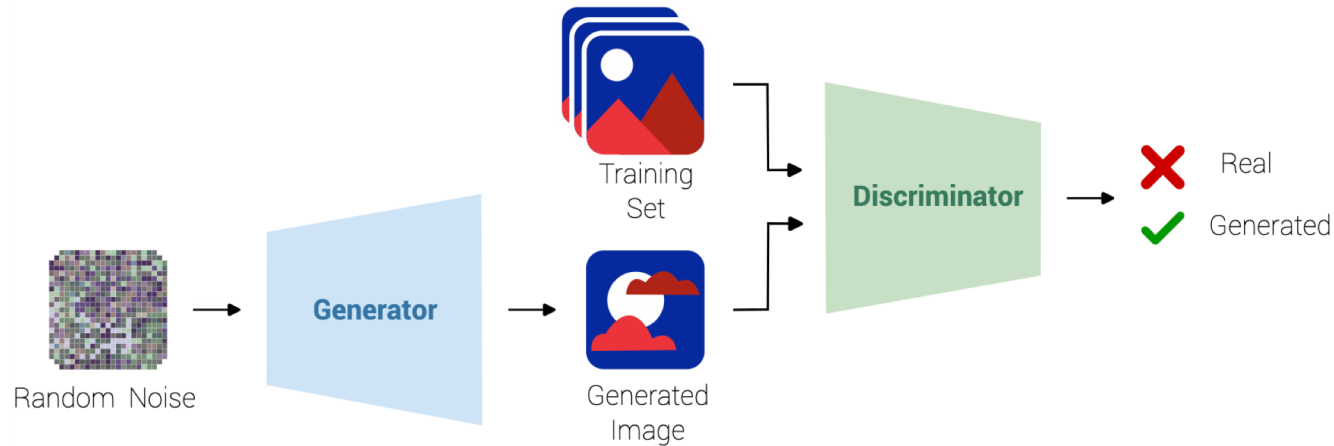


Figure 1. Illustration of the GAN training process.

[1] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, Yoshua Bengio: [Generative Adversarial Nets](#). NIPS 2014: 2672-2680

1.1. Generative Adversarial Networks (GAN)

Example of applications



[1] <https://thispersondoesnotexist.com/>

1.1. Neural Style Transfer

Training paradigm

Neural Style Transfer [1] **transfers the style from one image to another while keeping the content unchanged**. The architecture requires **two networks**, a **neural transfer network** and a **feature extractor**. The transfer model is trained to generate images given the input image. Next, the feature extractor compares the internal representations of the input, generated, and style images and attempts to reduce the distance between internal layer representations encoding styles or content.

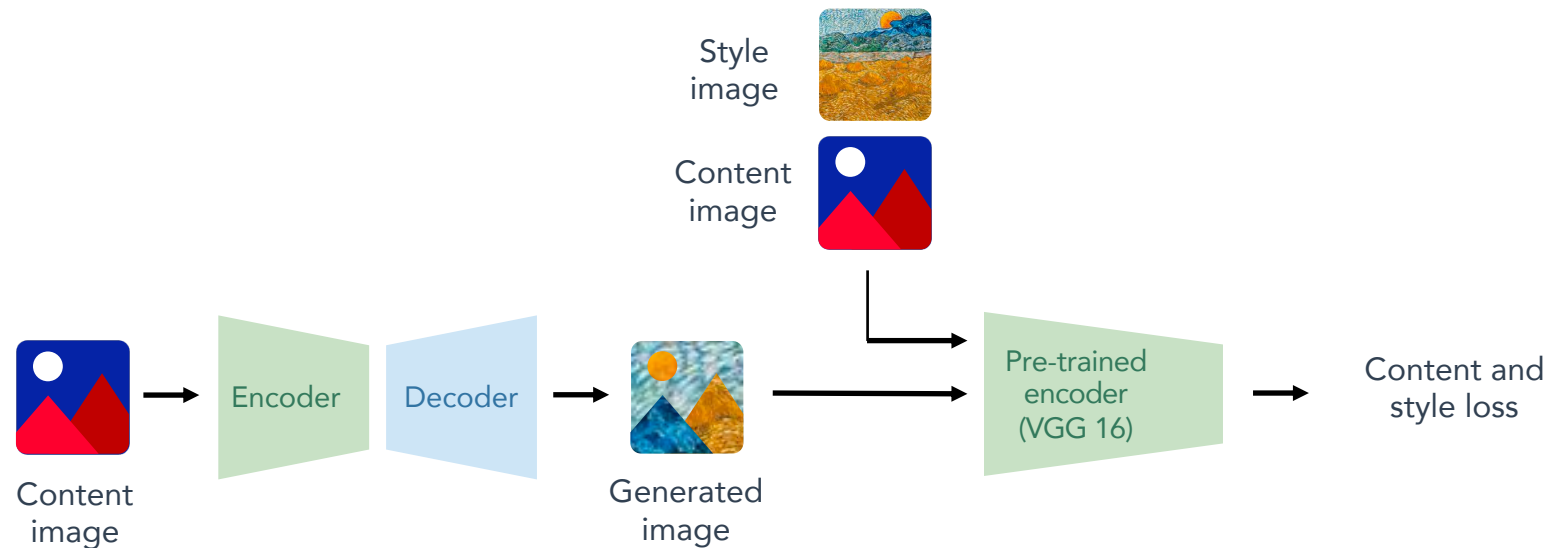


Figure 1. Illustration of the architecture used to train neural style transfer models.

[1] Alexander Mordvintsev, Christopher Olah and Mike Tyka: Inceptionism: [Going Deeper into Neural Networks](#), Google blog post, June 2015

1.1. Neural Style Transfer

Application using Deep Dream Generator

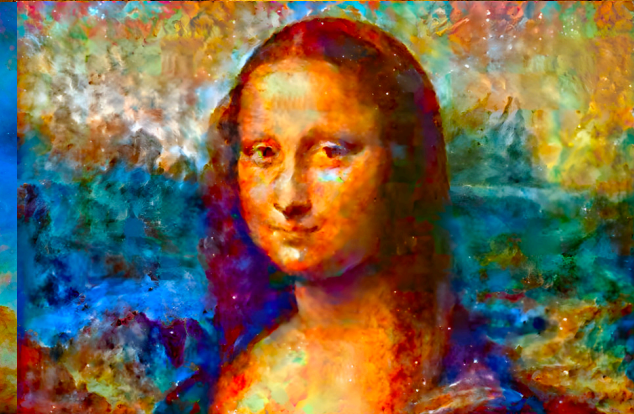
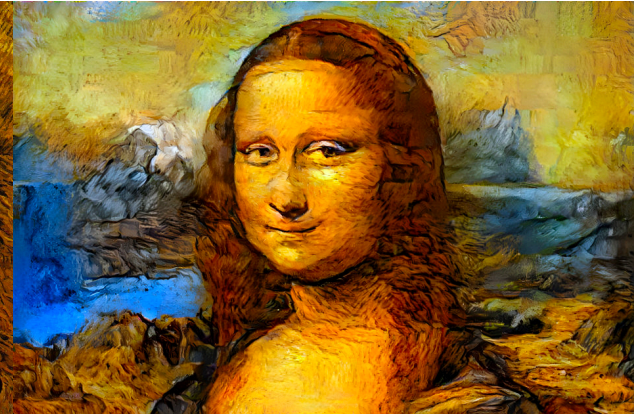
Content
image



Style
image



Generated
image



[1] <https://deepdreamgenerator.com/>

1.1. Diffusion process

General principle

Diffusion models [1, 2, 3, 4] process iteratively by progressively adding noise to the input image over T steps until it is entirely random noise. Instead of generating the image from random noise in a single step, we decompose the problem by training the model to **generate a slightly less noisy image at each step**.

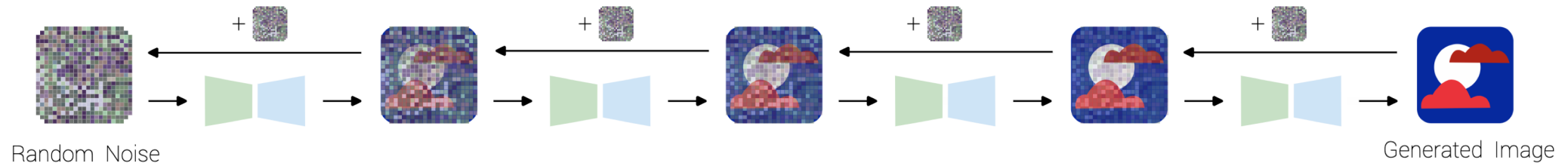


Figure 1. Illustration of the forward and backward diffusion process.

[1] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, Surya Ganguli: [Deep Unsupervised Learning using Nonequilibrium Thermodynamics](#). ICML 2015: 2256-226

[2] Jonathan Ho, Ajay Jain, Pieter Abbeel: [Denosing Diffusion Probabilistic Models](#). NeurIPS 2020

[3] Alexander Quinn Nichol, Prafulla Dhariwal: [Improved Denosing Diffusion Probabilistic Models](#). ICML 2021: 8162-8171

[4] Prafulla Dhariwal, Alexander Quinn Nichol: [Diffusion Models Beat GANs on Image Synthesis](#). NeurIPS 2021: 8780-8794

1.1. Diffusion process

Training

In practice, we generate images with **various degrees of additional noise** and train the encoder-decoder model to **predict the noise from the image** such that we can reconstruct the original image by removing the noise.

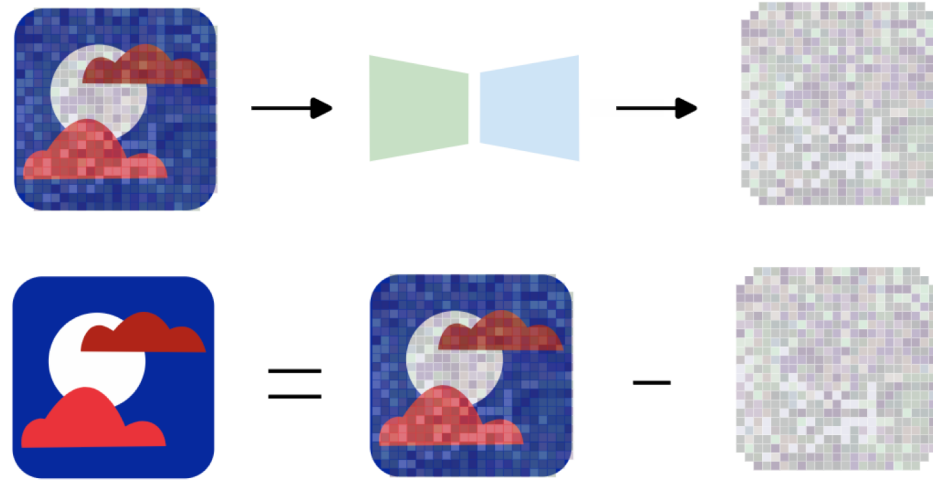
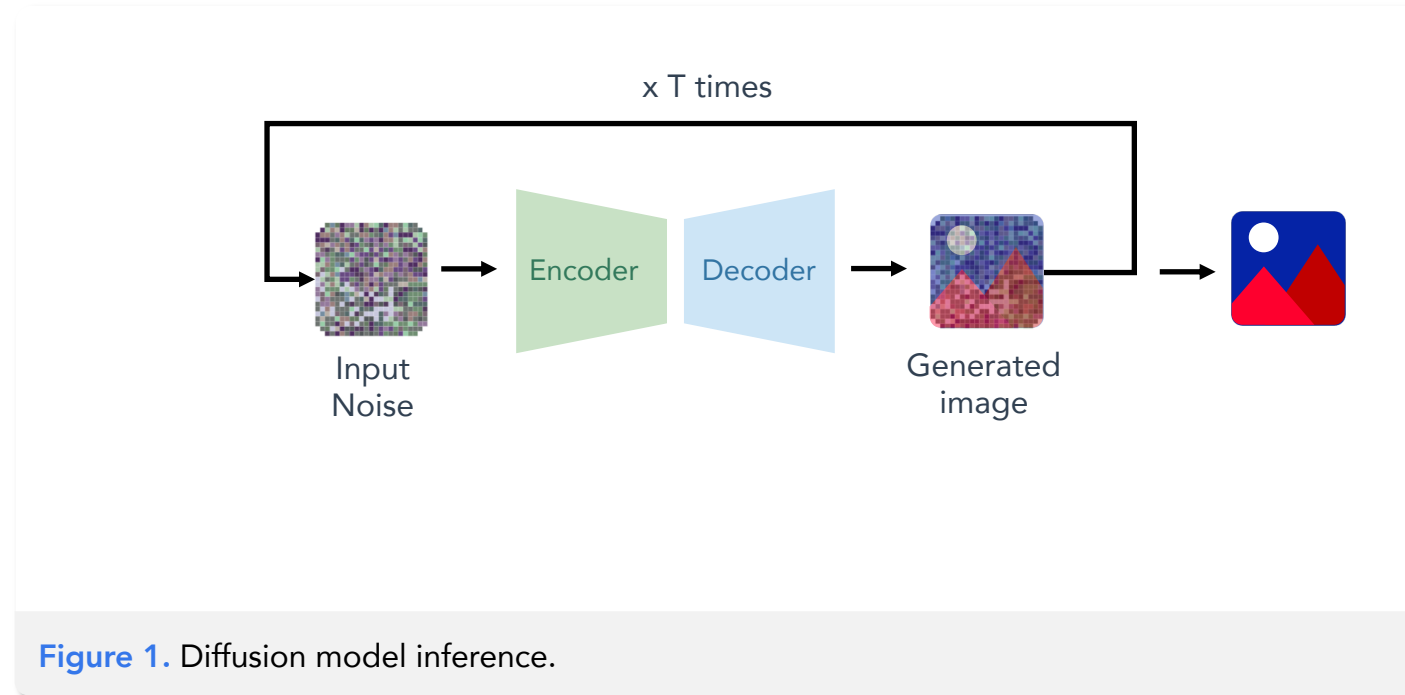


Figure 1. Diffusion model training.

1.1. Diffusion process

Inference

At inference, we feed the network with noise and **generate a slightly noisy image**, which we feed back to the network. We **repeat** the T steps process **until we generate a fully detailed image**.



1.1. Diffusion process

Text conditioning

To guide the process, we can feed the network with a text prompt embedded together with the image from the previous step. We can also select the images to **choose the extent to which we want to fit the text description**.

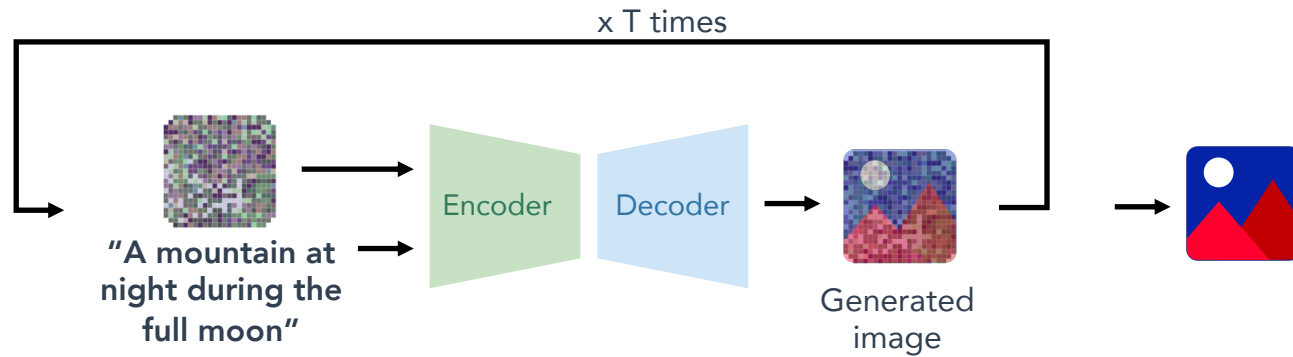


Figure 1. Diffusion model inference.

1.1. Diffusion process

Applications

The diffusion process has started to surpass other computer vision generative approaches on many tasks, including image generation [1] and conditional setting, including converting text descriptions to images [2], inpainting [3], and image modification [4].

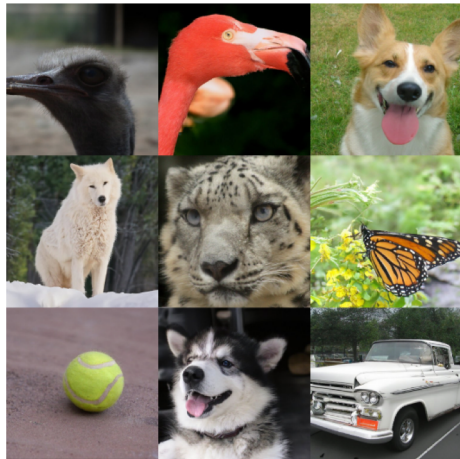


Figure 1. Image generation [1].



Figure 2. Text description to image [2].

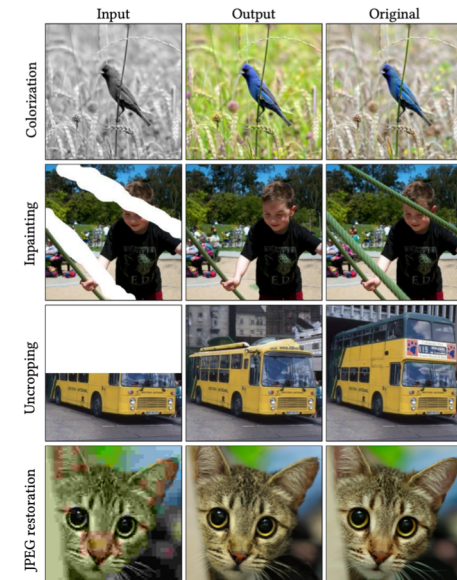


Figure 3. Image Inpainting [3].

[1] Prafulla Dhariwal, Alexander Quinn Nichol: [Diffusion Models Beat GANs on Image Synthesis](#). NeurIPS 2021: 8780-8794

[2] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, Mark Chen: [GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models](#). ICML 2022: 16784-16804

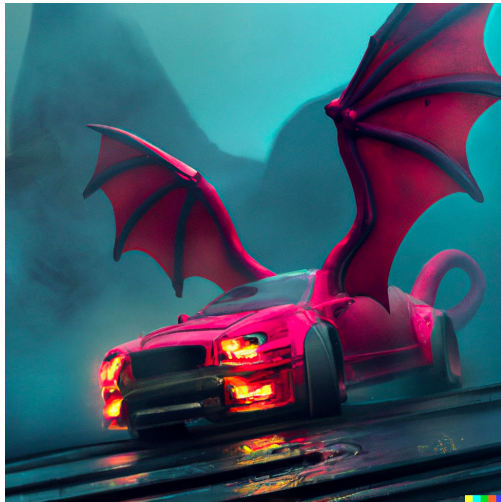
[3] Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, Mohammad Norouzi: [Palette: Image-to-Image Diffusion Models](#). SIGGRAPH (Conference Paper Track) 2022: 15:1-15:10

[4] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, Mark Chen: [Hierarchical Text-Conditional Image Generation with CLIP Latents](#). CoRR abs/2204.06125 (2022)

1.1. Diffusion process

Application using DALL-E-2

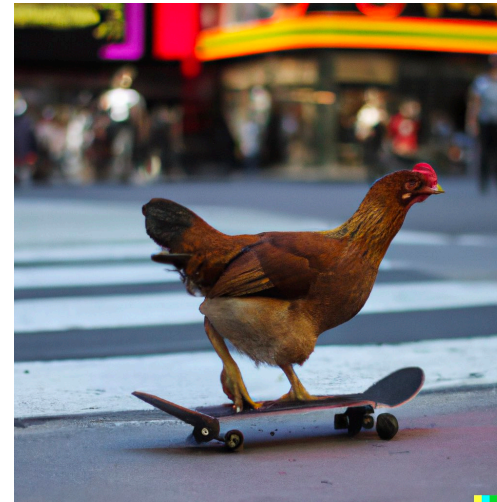
DALL-E 2 [1] is a new AI system capable of **creating realistic images and artworks from a natural language description**. OpenAI develops the model, and it is an iteration of the first version of the model. More recently, Google has also proposed an algorithm for generating realistic images [2].



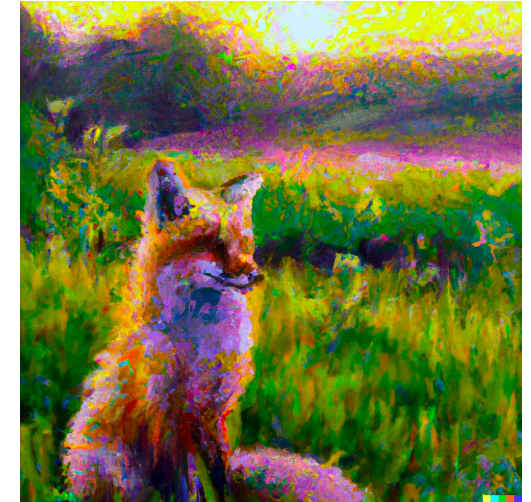
One of my favorite @OpenAI #dalle images today: "car dragon, digital art"



An astronaut riding a horse in a photorealistic style



A photo of a chicken on a skateboard in Times Square. #dalle2



A painting of a fox sitting in a field at sunrise in the style of Claude Monet

Figure 1. Images generated given an input text input using DALL-E-2 [1].

[1] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, Mark Chen: [Hierarchical Text-Conditional Image Generation with CLIP Latents](#). CoRR abs/2204.06125 (2022)

[2] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, Mohammad Norouzi: [Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding](#). CoRR abs/2205.11487 (2022)

1.1. Diffusion process

Controlling the impact of the text input

We can **adjust the impact of the text input** by generating multiple images at each time. We then rank the images using the model CLIP [1] from OpenAI and select the one that matches the text input best (or not) to **guide the generation process**.

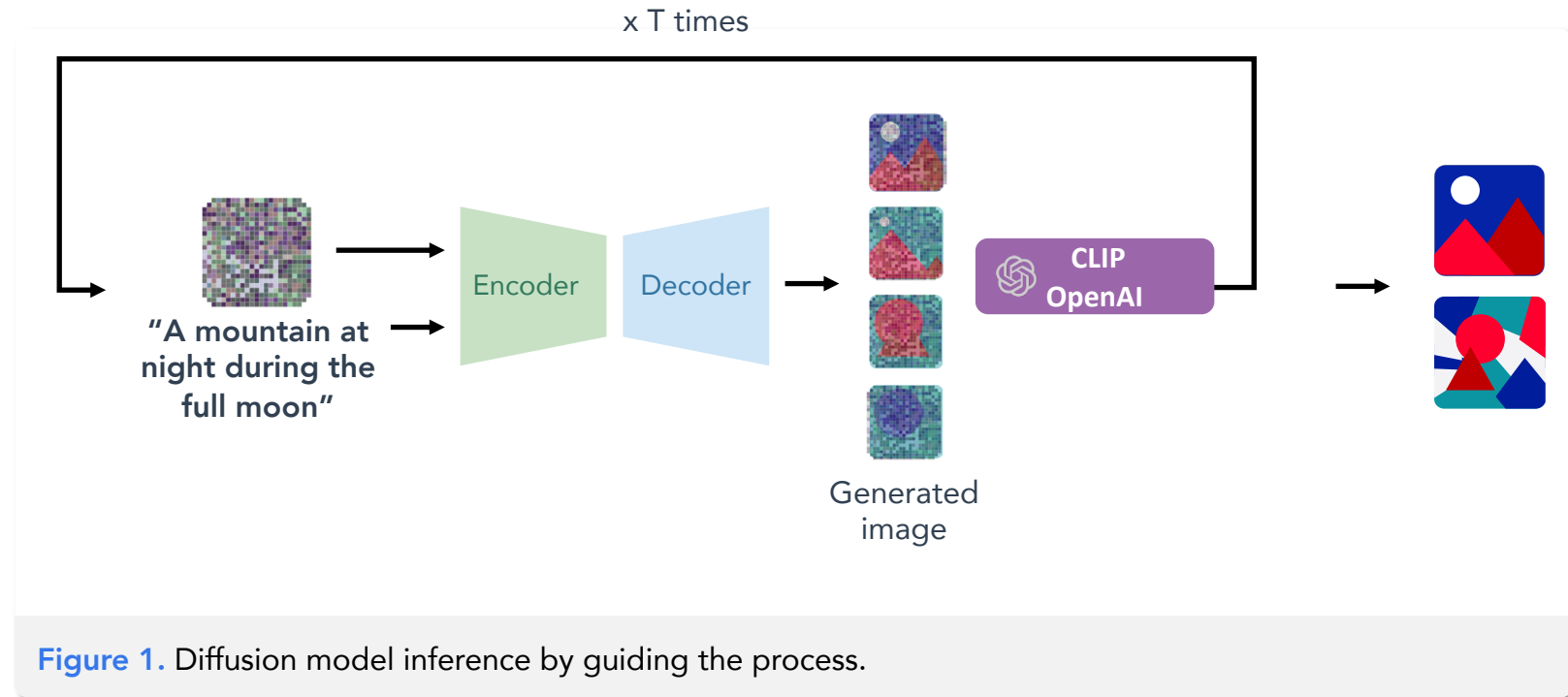


Figure 1. Diffusion model inference by guiding the process.

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever: [Learning Transferable Visual Models From Natural Language Supervision](#). ICML 2021: 8748-8763

1.1. Diffusion process

Reducing the computation requirement by working in the latent space

Finally, generating full-resolution images can be **computationally expensive**. It is possible to embed the images within a **latent space** and perform the process here to reduce the hardware requirement.

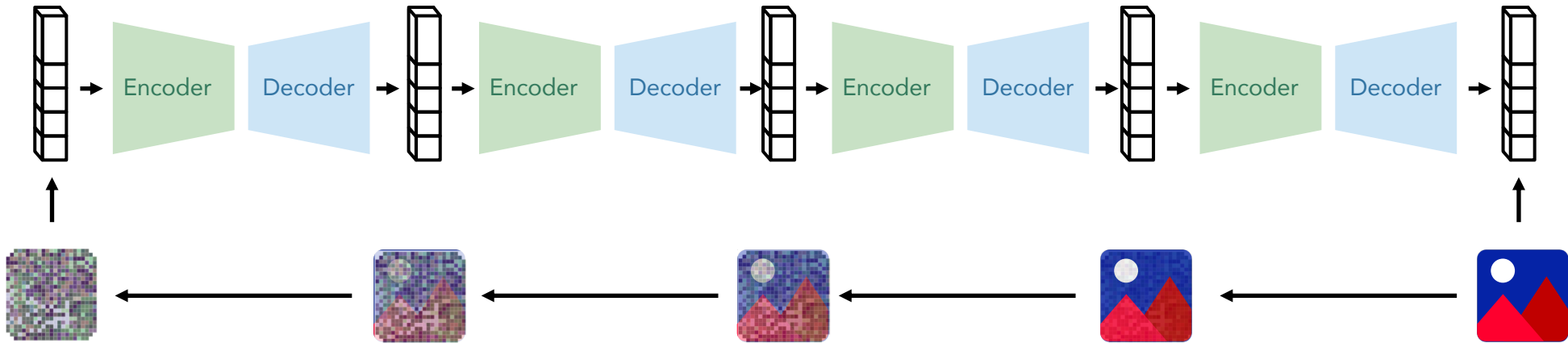


Figure 1. Illustration of the forward and backward diffusion process in latent space.

#3

Available resources and tools

1.1. Resources

Libraries and implementations

Multiple libraries implement diffusion models in **open-source** settings, including stability.ai and Hugging Face.



Figure 1. stability.ai



license Apache-2.0 release v0.7.0 Contributor Covenant 2.0

Figure 2. Hugging Face

[1] <https://stability.ai/>

[2] <https://github.com/huggingface/diffusers>

1.1. Resources

Tutorial and documentation

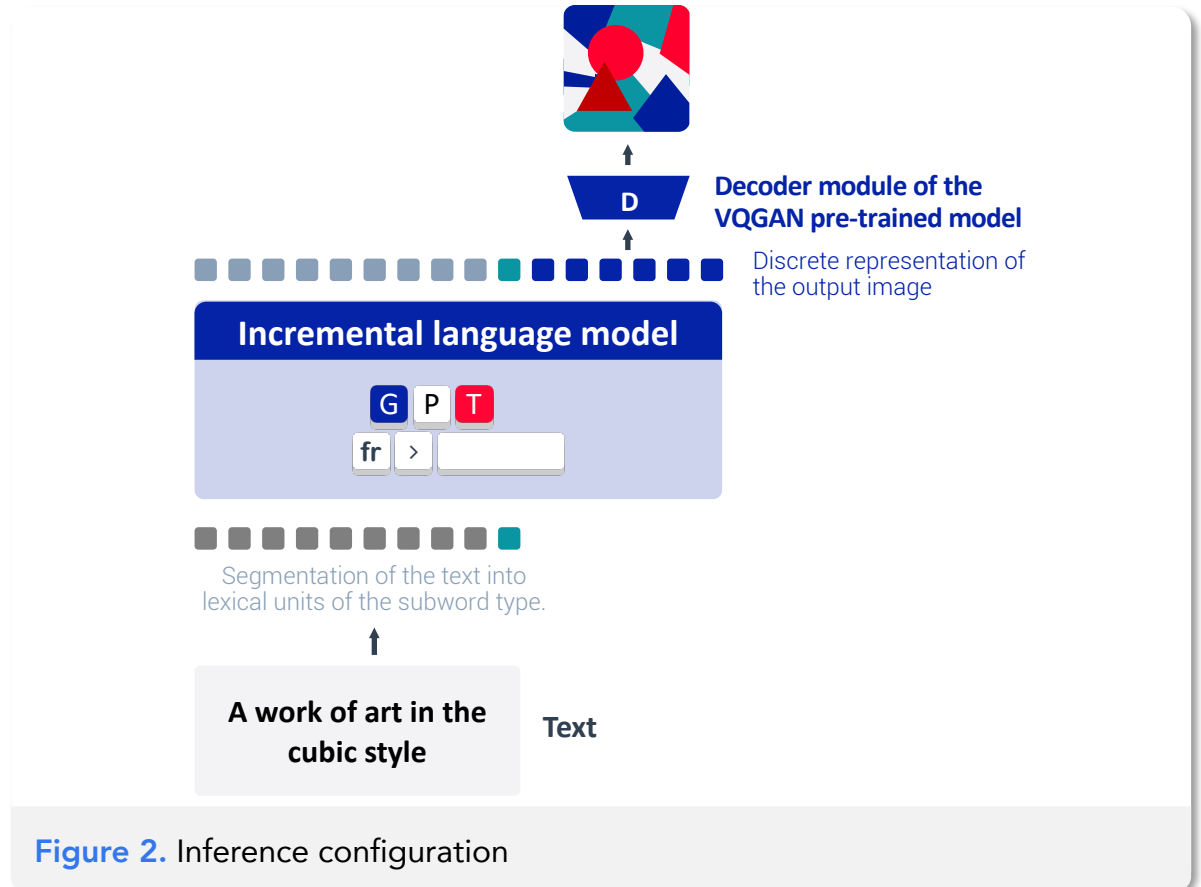
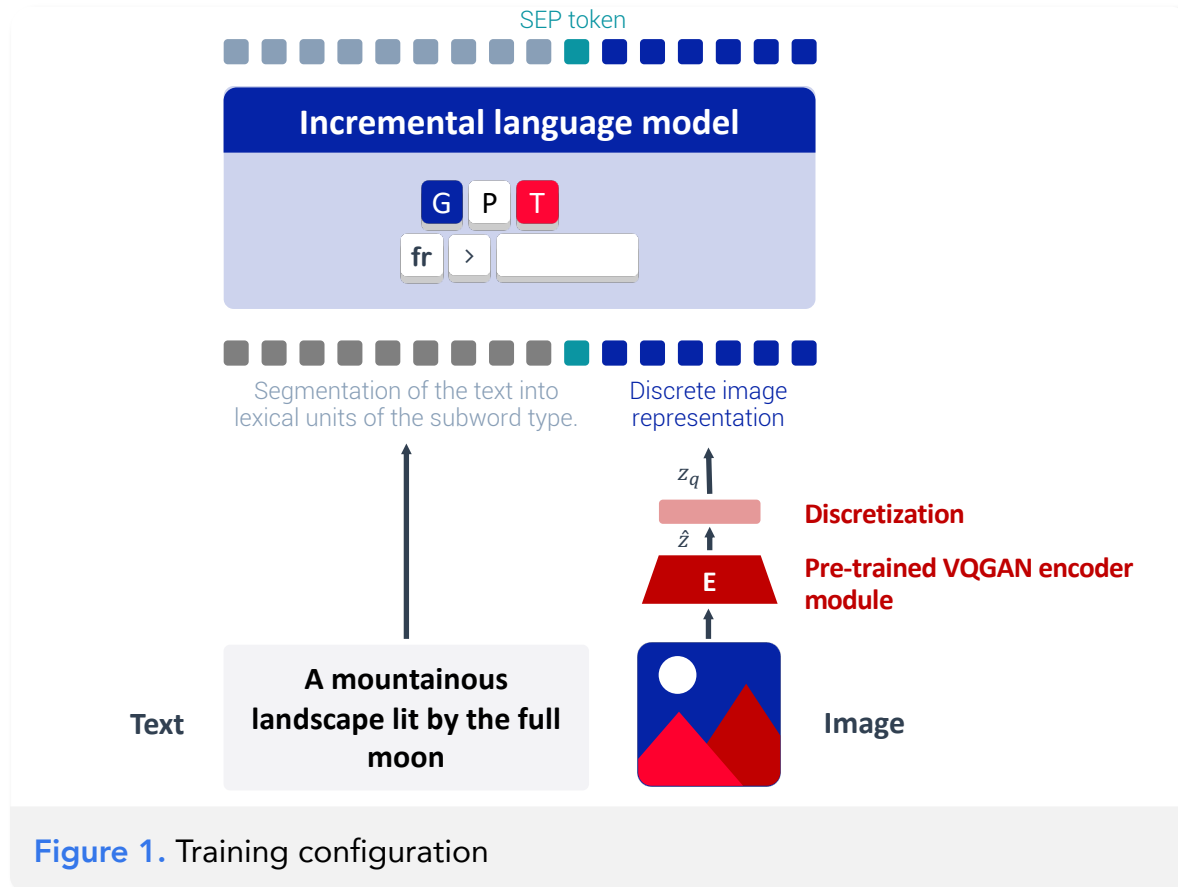
Multiple blog posts also clearly explain the method and models, including:

- <https://www.v7labs.com/blog/ai-generated-art>
- <https://www.assemblyai.com/blog/diffusion-models-for-machine-learning-introduction/>
- <https://lilianweng.github.io/posts/2021-07-11-diffusion-models/>
- <https://jalammar.github.io/illustrated-stable-diffusion/>

1.1. My research

Adaptation of GPT-fr for image generation

My model generates image out of text input in French.



[1] Patrick Esser, Robin Rombach, Björn Ommer: [Taming Transformers for High-Resolution Image Synthesis](#). CVPR 2021: 12873-12883

[2] <https://huggingface.co/asi/igpt-fr-cased-base>

1.1. Conclusion



The technology behind image generation has **enormously evolved** and fused with natural language applications.



Numerous tools have emerged based on these technologies.



Generation content technology still raise **copyright question**.



Some **use cases may be applicable in real-world settings**, including for example, information extraction.

1.1. Contact



<https://www.linkedin.com/in/antoine-simoulin/>