Advanced Generative AI: Models, Tools and Applications

Stable Diffusion, Denoising, and Autoencoders

Quick Recap

Lorem ipsum dolor sit amet, adipiscing elit, sed do incididunt ut labore magna aliqua.?

- How do the tasks in the SuperGLUE benchmark differ from those in GLUE, and what implications do these differences have for the evaluation of language models' understanding capabilities?
- Considering that BIG-bench is a collaborative benchmark, how does its approach to evaluating large language models (LLMs) compare to the more standardized evaluations like HELM, and what are the potential benefits and challenges of this collaborative model?

Engage and Think

StyleFusion, a leading fashion company, is undertaking a revolutionary project to transform fashion design by leveraging advanced AI technologies. The objective is to create an AI system that can design unique, trend-setting clothing by merging historical fashion trends with modern styles. This system employs stable diffusion techniques to craft a variety of fashion designs, utilizing a comprehensive database of both historical and modern fashion elements. Additionally, it incorporates denoising algorithms to improve the sharpness and detail of each design produced.

How might the use of stable diffusion techniques in the AI system affect the creative process of fashion design at StyleFusion?

Learning Objectives

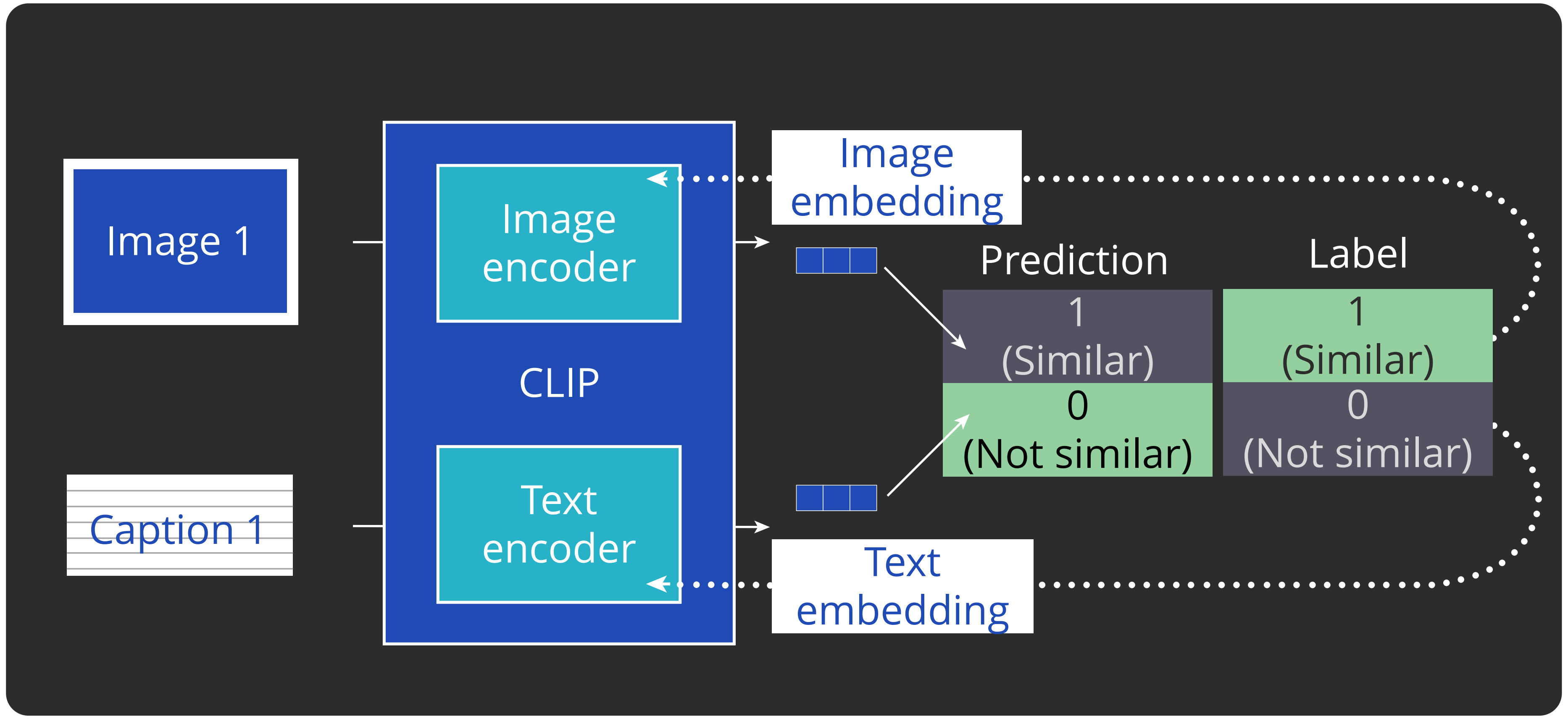
By the end of this lesson, you will be able to:

- Explain the principles of stable diffusion and its role in denoising to gain insights into advanced image generation and refinement techniques
- Analyze how autoencoders are used in contrastive learning to enhance the model's ability to understand and differentiate between complex data patterns
- Discover the concept of shared embedding spaces and their significance in multimodal AI to enable more coherent and effective interpretation of diverse data types

Stable Diffusion and Denoising

Stable Diffusion

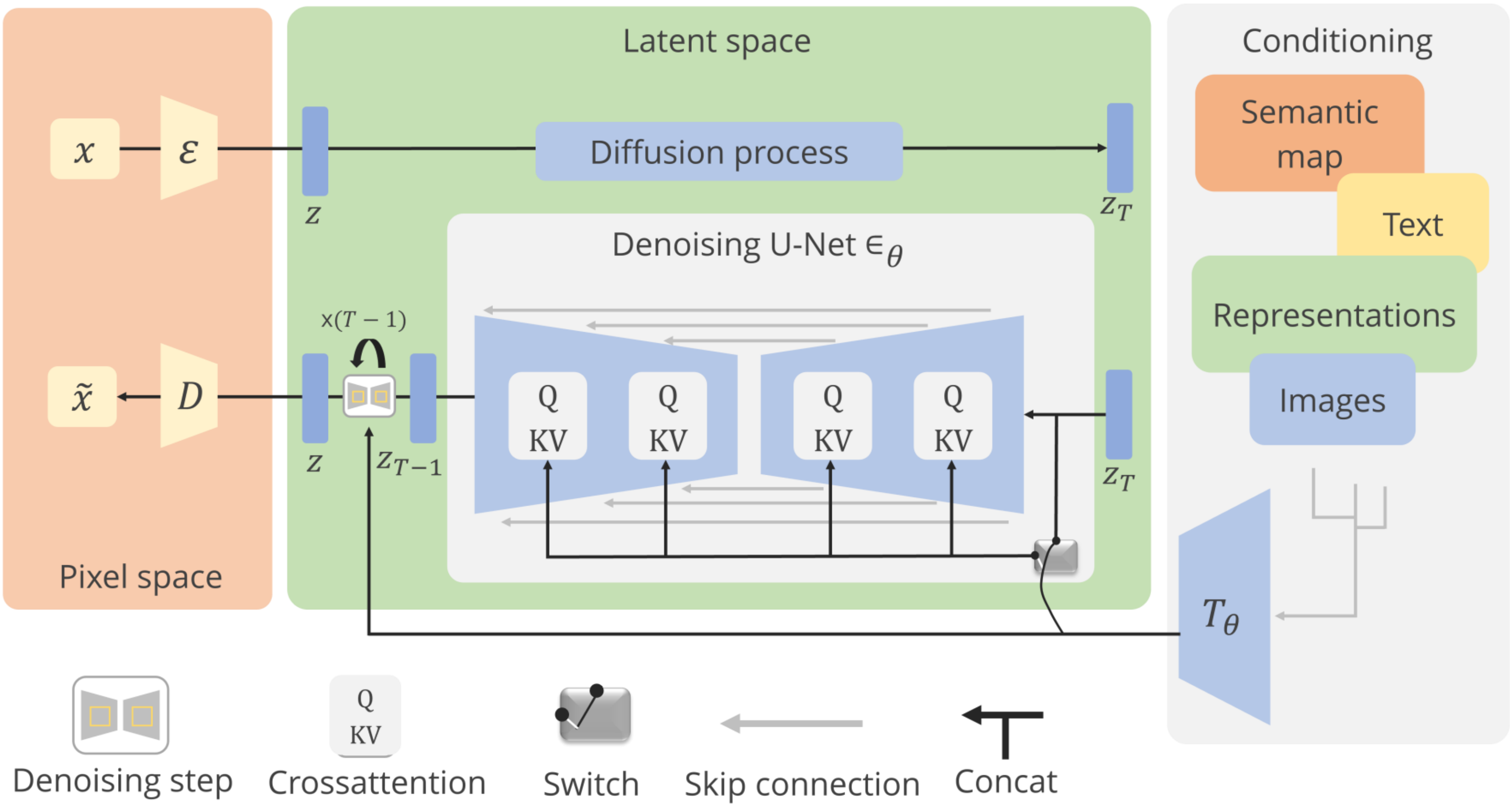
Stable diffusion is a type of deep generative artificial neural network that uses a diffusion model to generate detailed images conditioned on text descriptions.



It is also capable of tasks such as inpainting, outpainting, and generating image-to-image translations guided by a text prompt.

Understanding the Stable Diffusion Model

Stable diffusion uses a kind of diffusion model called a latent diffusion model (LDM), developed by the CompVis group at LMU Munich.



The autoencoder within the LDM captures the perceptual structure of the data by projecting the data into latent space.

The model consists of a variational autoencoder (VAE), U-Net, and an optional text encoder. The code and model weights of Stable Diffusion are open-sourced.

Stable Diffusion Models: Applications

Generative art

- Stable diffusion models can be used in the creation of generative art, enabling the generation of complex and unique visuals that would be challenging to create by hand.
- This application extends to various visual art forms, including paintings and videos.

Deepfake video generation

- The model can be utilized in the creation of deepfake videos, where it can generate realistic videos of people doing or saying things they never did.
- This has implications for various applications, including entertainment and visual effects.

Stable Diffusion Models: Applications

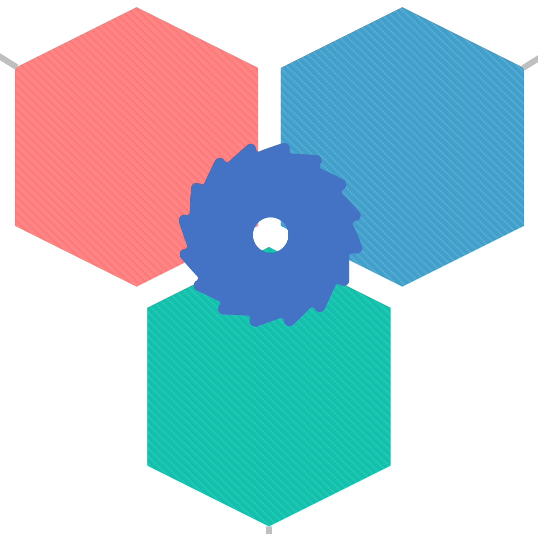
Image generation via prompting

- Stable diffusion models can generate images based on textual prompts or conditioning inputs, allowing for controlled and customizable image synthesis.

Image super-resolution

- These models can enhance the resolution and quality of low-resolution images, generating high-resolution outputs with improved details and clarity.

Challenges in Training Stable Diffusion Models



Optimizing tile-based pipelines

It has a native resolution of 512×512 pixels, which can be a limitation in handling non-square or higher-resolution images.

Addressing issues with human limbs and physics

These models often struggle with generating realistic human limbs and maintaining consistent physics, leading to artifacts like limbs passing through solid objects or anatomically incorrect structures.

Customization

Integrating content outside the pretrained sphere and training them on a large volume of novel images for an existing, mature, and capable model is a challenge.

Denoising

The denoising process in stable diffusion is achieved through a reverse diffusion process done gradually through multiple steps to remove noise, resulting in clearer images.

- The denoising function used in stable diffusion is learned implicitly by minimizing the difference between the predicted and ground-truth noise-free images.
- This has gained attention due to its ability to generate high-quality images even in the presence of high noise levels.
- The diffusion process in stable diffusion models works by destroying training data by adding noise and then learning to recover the data, allowing it to generate coherent images from noise.

Denoising Diffusion Implicit Models (DDIM)

DDIM is a class of diffusion-based generative models primarily used for image synthesis and data denoising, operating by modeling the joint probability distribution of noisy data.

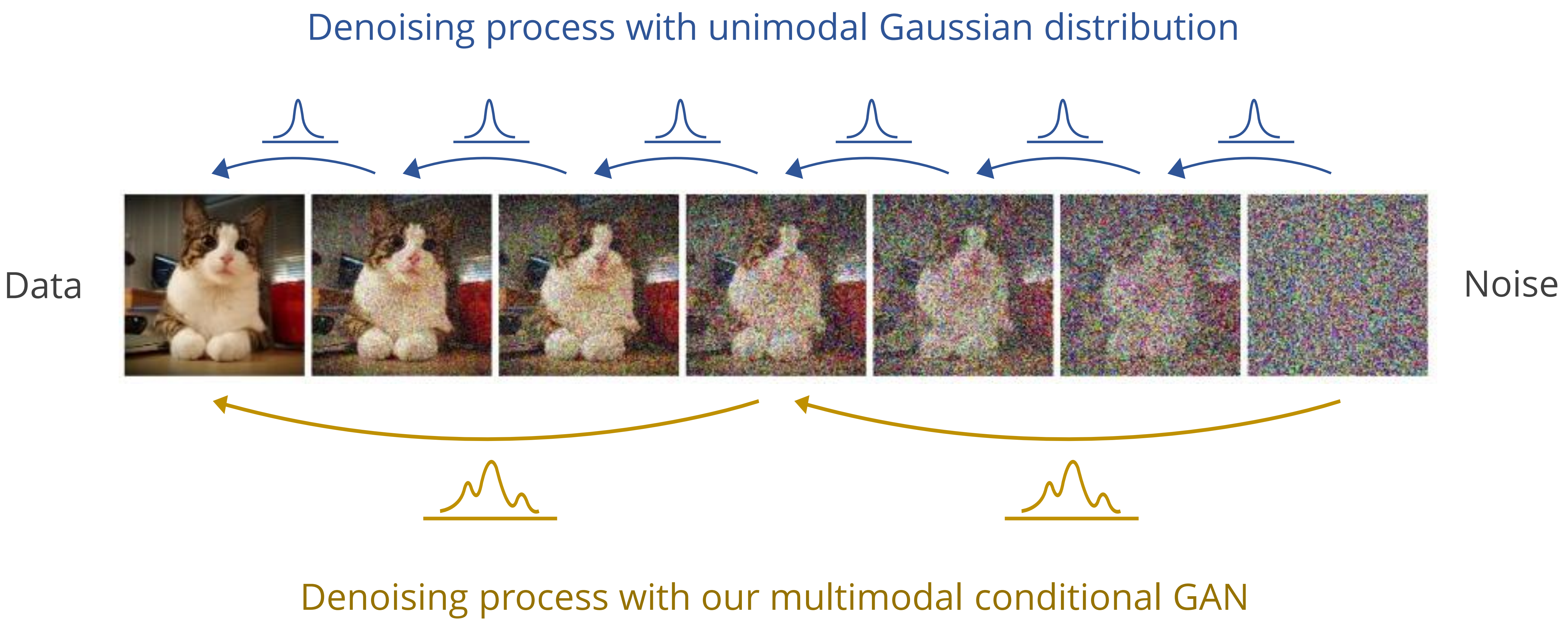
The denoising function used in DDIM is learned implicitly by minimizing the difference between the predicted and ground-truth noise-free images.

The denoising function is parameterized using a deep neural network, which inputs the noisy image and noise level and outputs the noise-free image.

DDIM has gained attention due to its state-of-the-art performance in image-denoising.

Denoising Diffusion Implicit Models (DDIM)

These models have demonstrated impressive results in high-fidelity image generation and offer strong sample diversity and faithful mode coverage of the learned data distribution.


Denoising process with unimodal Gaussian distribution
Data
Noise
Denoising process with our multimodal conditional GAN

Denoising Diffusion Models

Some key aspects of denoising diffusion models include:

- Denoising diffusion models define a forward diffusion process that maps data to noise by gradually perturbing the input data.
- These are also known as score-based generative models, which have emerged as a powerful class of generative models.
- Data generation is achieved using a learned, parametrized reverse process that performs iterative denoising, starting from pure random noise.

Denoising Diffusion Models

These models have been applied to various generative AI tasks, such as text-to-image generation, image super-resolution, and generative art.

One of the main challenges in denoising diffusion models is slow sampling, which has been addressed by developing promising techniques to overcome this issue.

Denoising deep generative models address the problem of dimensionality mismatch in likelihood-based deep generative models by adding Gaussian noise to the data, providing a denoising mechanism that aims to sample from the model as though no noise had been added.

Role of Denoising in Stable Diffusion Models

The role of denoising in stable diffusion models is to determine the balance between noise reduction and the preservation of image details.

Denoising strength is a crucial parameter in stable diffusion, as it affects the equilibrium between reducing noise and maintaining image details.

An optimal denoising strength is essential to avoid overly aggressive denoising, which can cause loss of details, or too mild denoising, which may result in obscured valuable data.

Demo: Stable Diffusion Text to Image

Duration: 15 minutes

Overview:

Create a system using the Stable Diffusion model from the diffusers library to transform images based on textual prompts, including both positive and negative descriptors.

The system should be capable of loading images from URLs or local paths and applying AI-driven modifications to these images, with a focus on adjusting various diffusion strengths to achieve different visual effects.

Note

Please download the solution document from the Reference Material Section and follow the Jupyter Notebook for step-by-step execution.

DEMONSTRATION

Quick Check

Which of the following is a key feature of stable diffusion in generative AI?

A. Stable diffusion models are incapable of tasks such as inpainting, outpainting, and generating image-to-image translations guided by a text prompt.

B. The denoising process in Stable Diffusion is achieved through a reverse diffusion process done gradually through multiple steps to remove noise, resulting in clearer images.

C. Stable diffusion uses a kind of diffusion model called a latent diffusion model (LDM), developed by the CompVis group at LMU Munich.

D. All of the above

Autoencoders and Contrastive Learning

Autoencoders in Generative AI

Autoencoders, particularly variational autoencoders (VAEs), are widely used in generative modeling.

VAEs enable the generation of complex generative models of data, yielding state-of-the-art performance for tasks such as image generation.

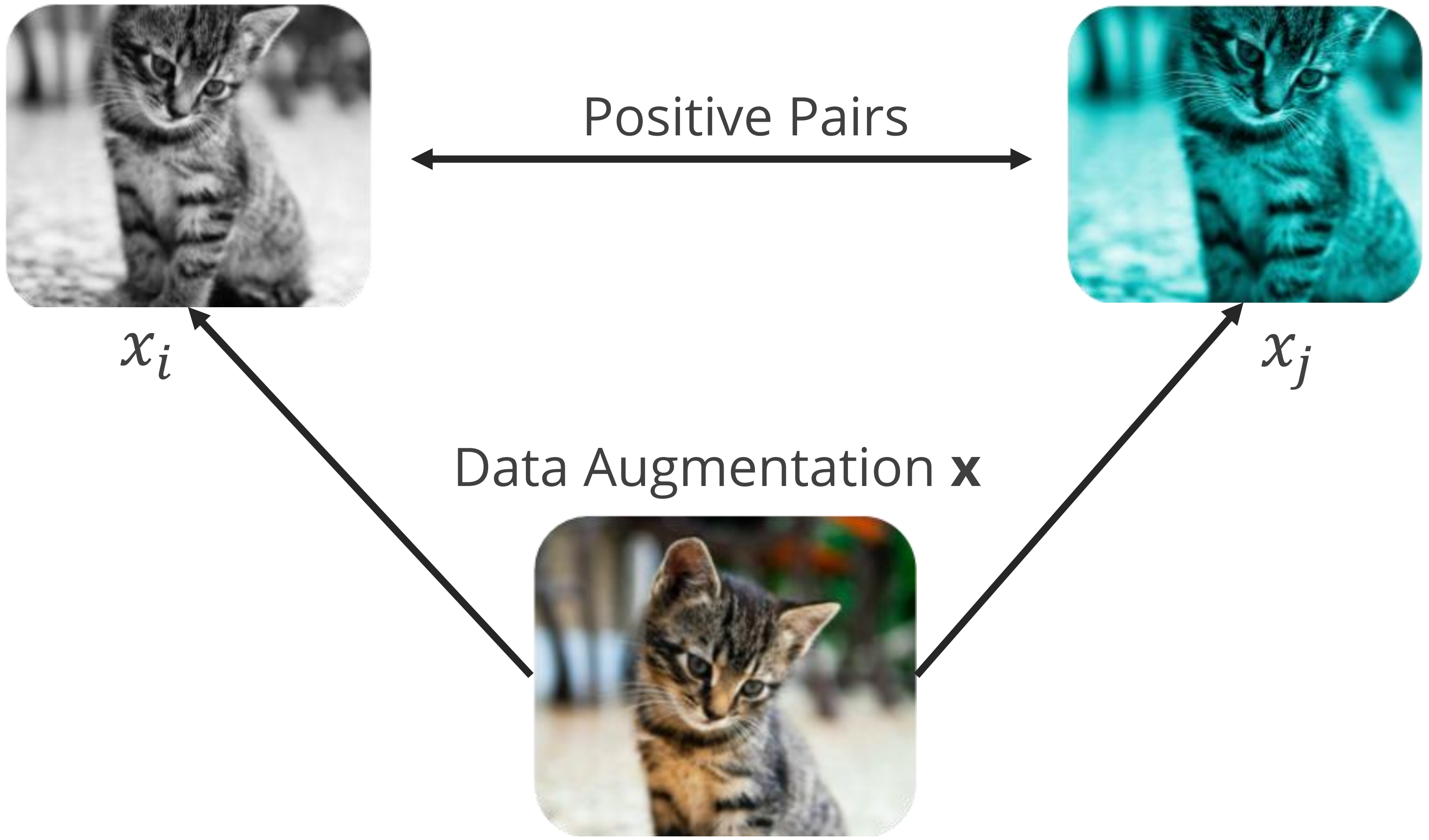
They can be used for various applications, including data generation, image augmentation, image-to-image translation, natural language generation, and de novo molecular design.

Note

The topic of autoencoders is covered in detail in Lesson 03; please refer to it for more detailed information.

Contrastive Learning Techniques

Contrastive learning involves training a model to differentiate between similar and dissimilar pairs of data points by leveraging the principles of contrasting samples against each other.



It is a powerful technique used in various domains, including computer vision, natural language processing, and reinforcement learning.

Autoencoders and Contrastive Learning

The combination of autoencoders and contrastive learning has led to novel approaches for self-supervised learning and representation learning.

For example:

I. Contrastive masked autoencoders (CMAE)

II. ConMH (Contrastive masked autoencoders for self-supervised video hashing)

These methods have shown improved performance in various tasks, such as image generation, image classification, and video hashing.

Autoencoders and Contrastive Learning

The combination of autoencoders and contrastive learning allows models to learn robust representations by leveraging the strengths of both techniques.

Autoencoders can learn compact and informative representations, while contrastive learning helps to learn embeddings that preserve the relationships between data points.

This combination has led to novel approaches that address the challenges faced by traditional autoencoders and contrastive learning methods.

Autoencoders and Contrastive Learning: Advantages

Improved expressivity

The combination of autoencoders and contrastive learning can improve the expressivity of various models, such as Variational Autoencoders (VAEs).

Enhanced generative capabilities

The combination of autoencoders and contrastive learning can lead to novel approaches that address the challenges faced by traditional autoencoders and contrastive learning methods.

Autoencoders and Contrastive Learning: Advantages

Better generative image quality

The combination of autoencoders and contrastive learning can improve the generative quality of images, especially when samples are obtained from the prior without any tempering.

Applicability to a wide range of tasks

The combination of autoencoders and contrastive learning can be applied to various tasks, such as image generation, image classification, and video hashing

Contrastive Learning Techniques

Some common contrastive learning techniques and concepts include:

Data augmentation

Contrastive learning utilizes data augmentation techniques such as cropping, flipping, rotation, and color transformations to generate meaningful representations from unlabeled data.

Contrastive loss

This is a fundamental training objective used in contrastive learning. It takes a pair of samples that are either similar or dissimilar and brings similar samples closer while pushing dissimilar samples apart in the latent space.

Contrastive Learning Techniques

Some common contrastive learning techniques and concepts include:

SimCLR

SimCLR is a famous self-supervised framework for unsupervised contrastive learning that generates positive image pairs by applying random transformations to the anchor, such as flip and color jitter, to keep the label of the image unchanged.

Representation learning

Contrastive learning enables models to learn about data without labels by training them to recognize similarities and differences between data points, thereby allowing them to extract and learn high-level features of a dataset.

Real-World Applications of Contrastive Learning

Image classification

Object detection

Image segmentation

Semi-supervised learning

Quick Check

Which technique in generative AI utilizes data augmentation techniques such as cropping, flipping, rotation, and color transformations to generate meaningful representations from unlabeled data?

A. Autoencoders

B. Variational autoencoders (VAEs)

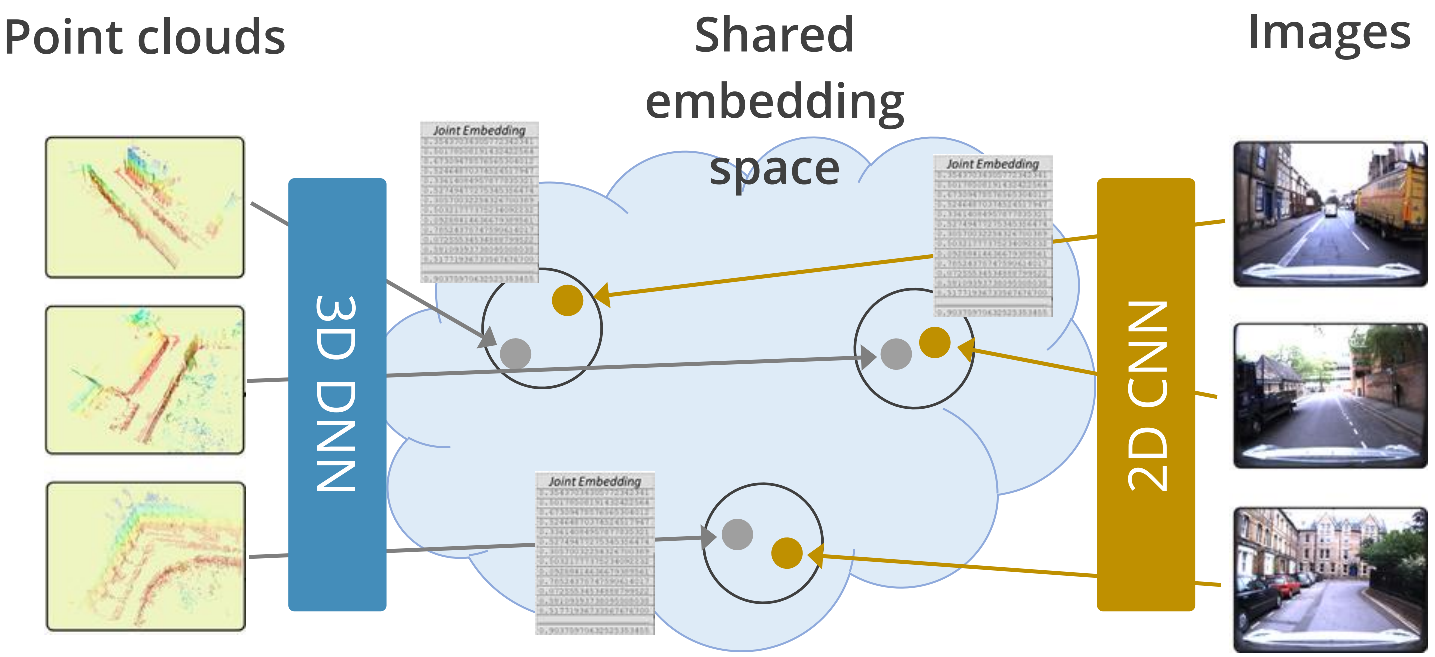
C. Contrastive learning

D. Contrastive masked autoencoders (CMAE)

Shared Embedding Spaces

Shared Embedding Spaces

Shared embedding spaces refer to the process of mapping different types of data into a single, unified embedding space.



Point clouds

Shared embedding space

Images

3D DNN

2D CNN

- This allows for the representation of various modalities, such as images, text, audio, and depth, in a common latent space.
- This can be beneficial for various machine learning tasks, including cross-modal retrieval, composing modalities with arithmetic, and cross-modal detection and generation.

Shared Embedding Spaces: Application

Knowledge graph completion

Shared embedding spaces can project entities and relations in knowledge graphs into continuous vector spaces, allowing for more efficient and accurate knowledge graph completion.

Multi-modal spaces co-embedding

A cost-effective method for co-embedding multi-modal spaces can preprocess embeddings using pretrained models for all without passing gradients through these models, avoiding cost inefficiencies.

Cross-modal retrieval

Shared embedding spaces enable the development of systems that can retrieve related items from different modalities, such as images and text.

Shared Embedding Spaces: Application

Composing modalities with arithmetic

By learning a joint embedding across multiple modalities, models can be composed using arithmetic operations, allowing for novel applications and emergent capabilities.

Cross-modal detection and generation

Shared embedding spaces can detect and generate items across different modalities, improving the performance of vision models for visual and non-visual tasks.

Shared Embedding Spaces vs. Individual Embedding Spaces

Shared embedding spaces	Individual embedding spaces
It maps different types of data into a single, unified embedding space, allowing for the representation of various modalities, such as images, text, audio, and depth, in a common latent space.	It refers to the vector spaces where each coordinate corresponds to a specific feature or attribute of the data.
This approach enables tasks like cross-modal retrieval, composing modalities with arithmetic, cross-modal detection, and generation.	In individual embedding spaces, each modality (e.g., images, text, audio) is represented in its own separate latent space without being combined with other modalities.
Shared embedding spaces can improve the performance of vision models for visual and non-visual tasks and can be used in knowledge graph analysis.	Individual embedding spaces are often used in unsupervised learning tasks, such as dimensionality reduction and data clustering, to capture the intrinsic structure of the data.

Techniques Used to Train Shared Embedding Spaces

A method of generating embeddings is by using an encoder-decoder model.

- This is a type of neural network that can compress input data into a smaller representation, called an embedding.
- The model consists of two parts: an encoder and a decoder.
- The encoder takes in the input data and applies a series of non-linear transformations, such as convolutional layers or self-attention layers, to compress it into a smaller representation.

Techniques Used to Train Shared Embedding Spaces

Below are the techniques to train the shared embedding spaces:

- Data augmentation techniques such as cropping, flipping, rotation, and color transformations can generate meaningful representations from unlabeled data.
- Graphical models: Shared embedding spaces can be used in graphical models to project entities and relations in knowledge graphs into continuous vector spaces, allowing for more efficient and accurate knowledge graph completion.

Techniques Used to Train Shared Embedding Spaces

Below are the techniques to train the shared embedding spaces:

- Co-embedding multi-modal spaces: A novel and cost-effective strategy for co-embedding multi-modal spaces can preprocess embeddings using pretrained models for all without passing gradients through these models, avoiding cost inefficiencies.
- Geometric structures: Some methods rely on the geometric structures of the underlying spaces as a proxy for parallel data, either relying on embedding similarity or detecting and exploiting invariances between pairs of low-dimensional embedding spaces.

Embedding Sharing for Multimodal AI

Embedding sharing for multimodal AI refers to the process of generating embeddings that can be used across different modalities, such as text, images, audio, and sensors.

Some key aspects of multimodal embeddings include:

Shared representation space

Applications

Architectures

Embedding Sharing for Multimodal AI

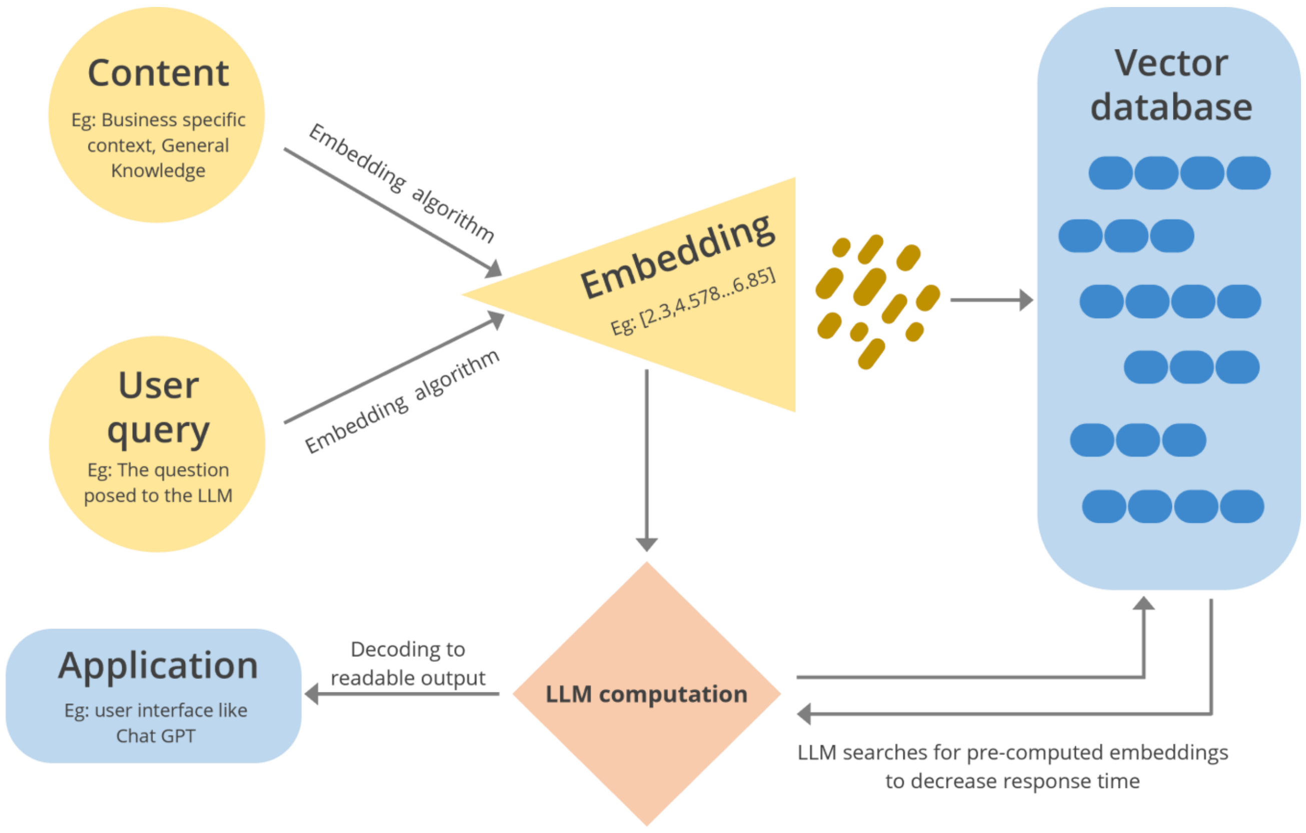
Large language models and their multimodal derivatives, such as Frozen, VL-Adapter, Flamingo, BEiT, and PaLI, show great potential for creating foundation models for multimodal AI.

Challenges:

- Multimodal learning lacks multiple sensory data, which can limit the performance of standard multimodal learning methods that rely on paired data.
- Recent research, such as ImageBind, has explored learning a single joint embedding space for multiple modalities, including text, image, audio, depth, thermal, and IMU readings.

The ongoing evolution of multimodal AI promises more innovative applications, including the integration of modalities like touch, speech, smell, and brain fMRI signals, to create richer human-centric AI models.

Techniques for Embedding Images, Text, and Code


Content
Eg: Business specific context, General Knowledge
Embedding algorithm
User query
Eg: The question posed to the LLM
Embedding algorithm
Embedding
Eg: [2,3,4,578...6.85]
Vector database
Application
Eg: user interface like Chat GPT
Decoding to readable output
LLM computation
LLM searches for pre-computed embeddings to decrease response time

LLM AI embeddings: LLM AI embeddings help generate images from texts and create code from texts, or vice versa, by converting different types of data into a common representation.

Text-generator API: Text-generator API offers an API for text and code generation and embeddings, allowing users to embed text, images, and code in the same space in up to 768 dimensions.

Techniques for Embedding Images, Text, and Code

Image embedding

To embed images, models extract features by encoding pixel data into high-dimensional vector representations. These embeddings preserve spatial and semantic relationships for tasks like search and classification.

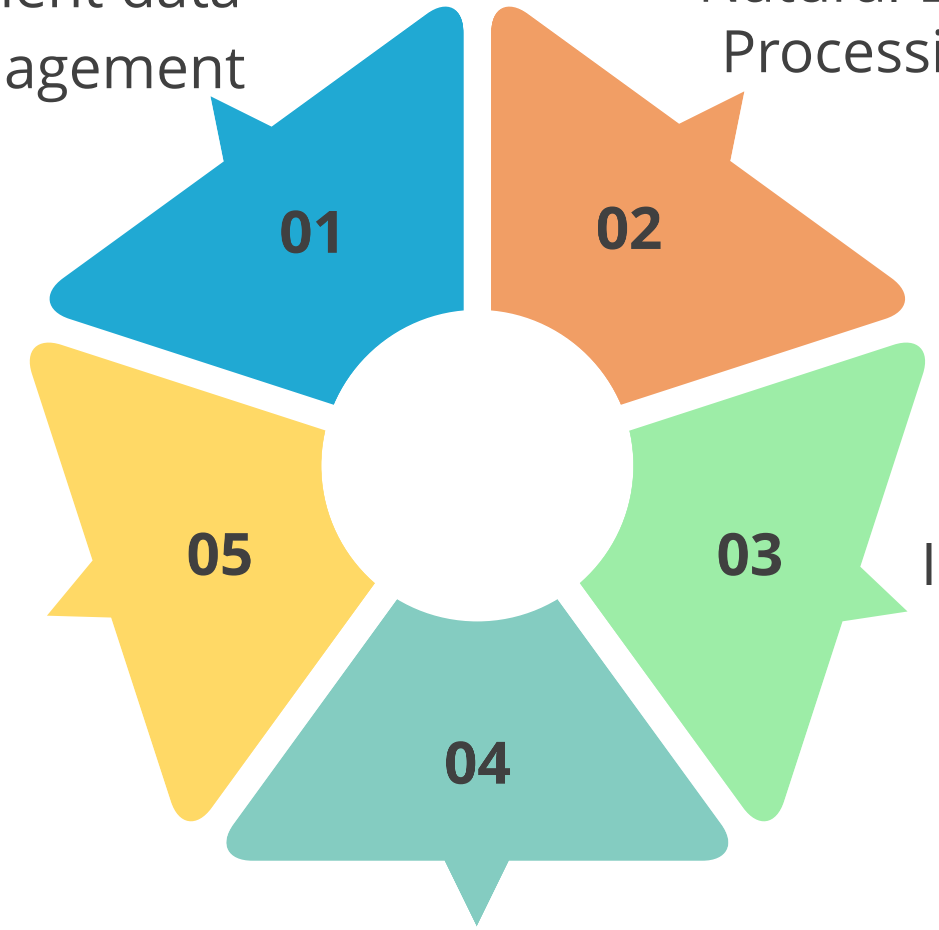
Text embedding

BERT-based embeddings capture the context and meaning of language by transforming words or sentences into dense vectors that encapsulate linguistic relationships.

GAN-based text-to-image generation

Text encoders, pretrained with image-text pairs, convert text descriptions into vector representations, which are used in GANs to produce realistic images.

Industrial Use Cases of Embeddings Space



Efficient data management

Natural Language Processing (NLP)

Image and text search

Multimodal AI

Generative AI image search

Quick Check

Which technique is commonly used in shared embedding spaces to generate meaningful representations from unlabeled data?

A. Data augmentation

B. Encoder-decoder models

C. Graphical models

D. Co-Embedding multi-modal spaces

Guided Practice

Overview

Duration: 25 minutes

In the realm of AI, the challenge of transforming images based on textual descriptions presents an intriguing opportunity to bridge the gap between linguistic semantics and visual representation. This tutorial delves into the nuanced process of altering images using text prompts, specifically through the lens of the Stable Diffusion Depth-to-Image Pipeline from Hugging Face. We will explore how to add depth or modify existing images to reflect the content of text prompts, offering a comprehensive guide to harnessing this cutting-edge technology.

Steps to Perform:

1. Install and import the necessary libraries for image transformation
2. Load a pretrained model for image transformation
3. Define helper functions
4. Load and display the image
5. Define prompts and transform the image
6. Experiment with different diffusion strengths

Key Takeaways

- Stable diffusion is a type of deep generative artificial neural network that uses a diffusion model to generate detailed images conditioned on text descriptions.
- The combination of autoencoders and contrastive learning has led to novel approaches for self-supervised learning and representation learning.
- Shared embedding spaces refer to the process of mapping different types of data into a single, unified embedding space.

Q&A