SUMMIT

SÃO PAULO | 3 DE AGOSTO DE 2023

DEV203

Olá Mundo! Conheça a IA Generativa na AWS

Vinicius Caridá (he/him)
AWS Machine Learning Hero
Itaú Unibanco

Vinicius Caridá, Ph.D.

- Head of Digital Customer Service Platforms, PCP, WFM, Data and AI
- MBA Professor – FIAP

@vinicius caridá @vfcarida @vinicius caridá @vfcarida @vinicius caridá @vfcarida

É Impossível . . . eu não consigo usar GenAI

É muito complexo
Não tenho budge
Não tenho dados . . .

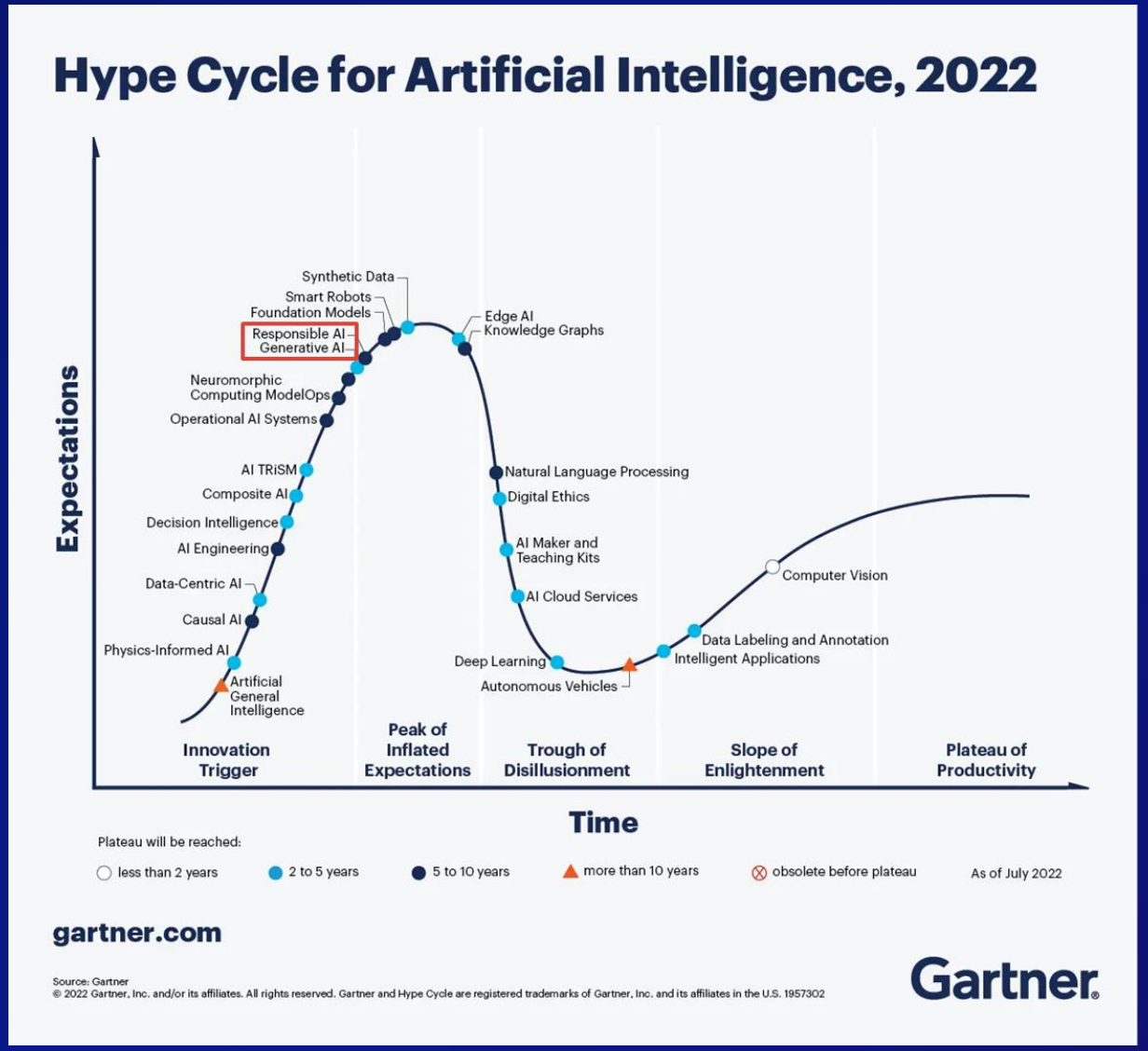
. . . vou tentar

“Do or do not. There is no try.”

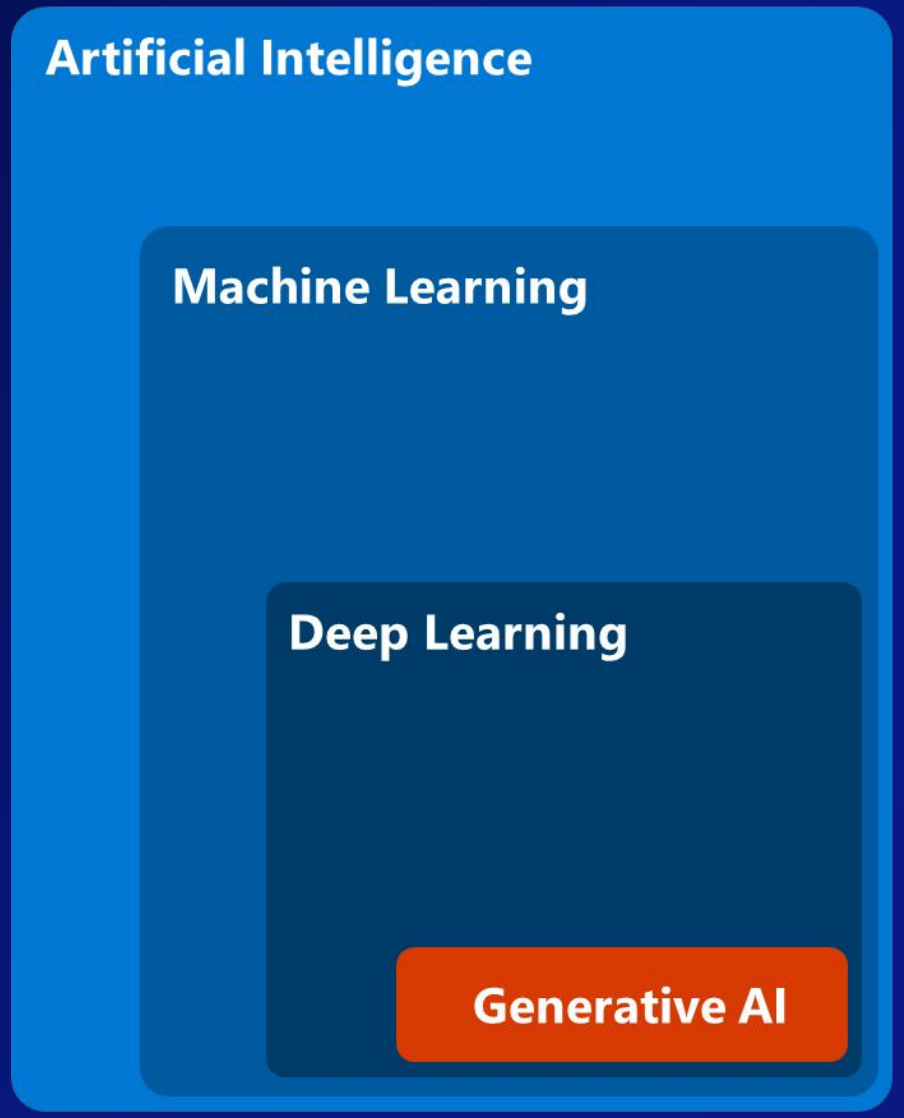
Yoda

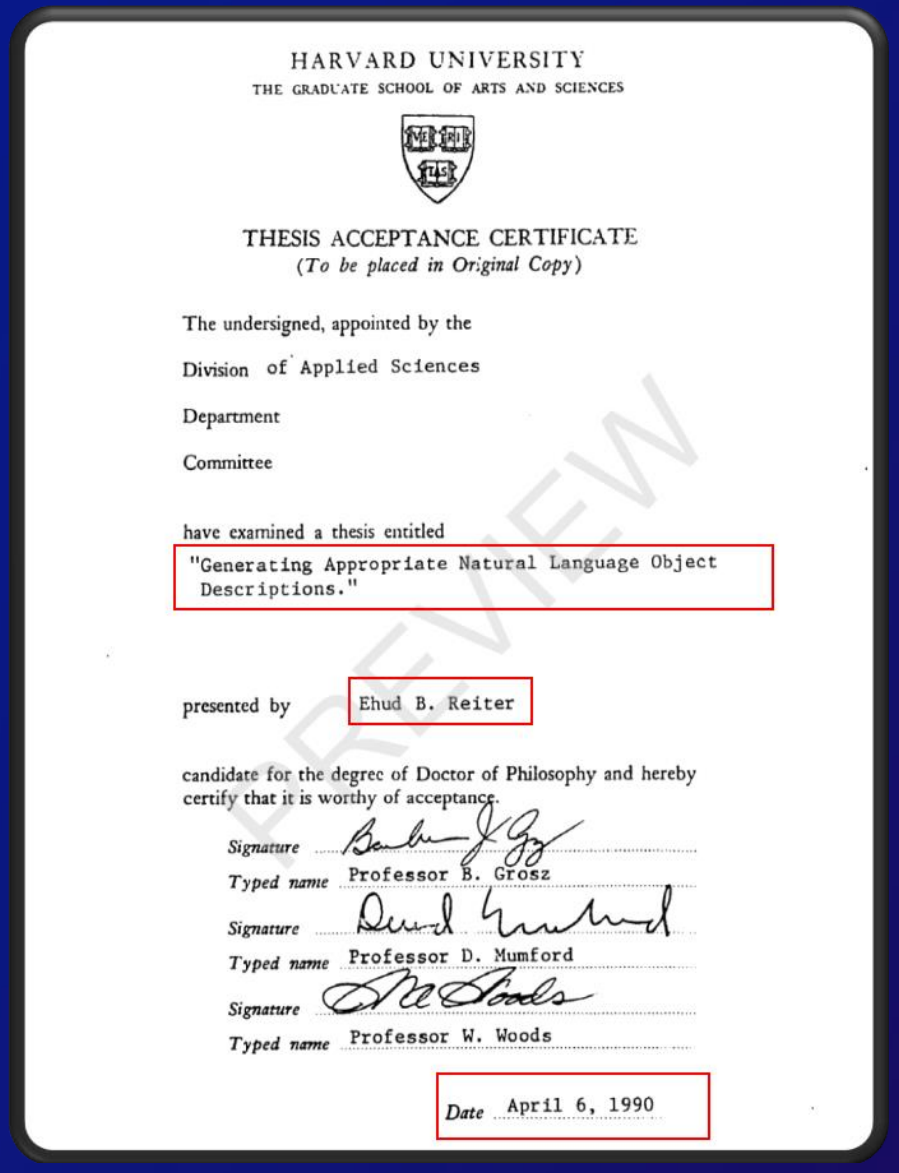
Star Wars

AI Generativa



AI Generativa





HARVARD UNIVERSITY
THE GRADUATE SCHOOL OF ARTS AND SCIENCES

THESIS ACCEPTANCE CERTIFICATE
(To be placed in Original Copy)

The undersigned, appointed by the

Division of Applied Sciences

Department

Committee

have examined a thesis entitled

"Generating Appropriate Natural Language Object Descriptions."

presented by Ehud B. Reiter

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature .....
Typed name Professor B. Grosz
Signature .....
Typed name Professor D. Mumford
Signature .....
Typed name Professor W. Woods

Date April 6, 1990

PREVIEW

https://www.proquest.com/openview/a17fb9188f537c2ab3df4091453547ba/1?pq-origsite=gscholar&cbl=18750&diss=y

The AWS AI/ML stack

Consumer

AI SERVICES

Generative AI

SCALING GENERATIVE AI
- Amazon Bedrock NEW

AMAZON BUILT FOUNDATIONAL MODEL
- Amazon Titan NEW

Specialized AI Services

BUSINESS PROCESSES
- Amazon Personalize
- Amazon Forecast
- Amazon Fraud Detector
- Amazon Lookout for Metrics

SEARCH
- Amazon Kendra

CONVERSATION
- Amazon Lex
- Contact Lens
- Voice ID
- Amazon Transcribe Call Analytics

CODE + DEVOPS
- Amazon CodeGuru
- NEW Amazon CodeWhisperer
- Amazon DevOps Guru

INDUSTRIAL
- Amazon Monitron
- Amazon Lookout for Equipment
- Amazon Lookout for Vision

HEALTH
- AWS HealthLake
- Amazon Comprehend Medical
- Amazon Transcribe Medical
- AWS HealthOmics

Core

TEXT
- Amazon Translate
- Amazon Comprehend

SPEECH
- Amazon Polly
- Amazon Transcribe

VISION
- Amazon Textract
- Amazon Rekognition
- AWS Panorama

Tuner

AMAZON SAGEMAKER

JUMPSTART NEW
Use pre-built models in SageMaker

CANVAS
No-code ML for business analysts

STUDIO LAB
Learn ML

GROUND TRUTH
Label data

STUDIO IDE
- Prepare data
- Store features
- Geospatial ML
- Build with notebooks
- Train models
- Tune parameters
- Deploy in production
- Manage and monitor

CI/CD | GOVERNANCE | RESPONSIBLE ML

Provider

ML FRAMEWORKS & INFRASTRUCTURE

PyTorch, TensorFlow | Amazon EC2 | CPUs | GPUs | NEW AWS Inferentia | NEW AWS Trainium | Habana Gaudi | FPGA

The AWS AI/ML stack

Consumer

AI SERVICES

Generative AI

SCALING GENERATIVE AI
- Amazon Bedrock

AMAZON BUILT FOUNDATIONAL MODEL
- Amazon Titan

Specialized AI Services

BUSINESS PROCESSES
- Amazon Personalize
- Amazon Forecast
- Amazon Fraud Detector
- Amazon Lookout for Metrics

SEARCH
- Amazon Kendra

CONVERSATION
- Amazon Lex
- Contact Lens
- Voice ID
- Amazon Transcribe Call Analytics

CODE + DEVOPS
- Amazon CodeGuru
- Amazon CodeWhisperer
- Amazon DevOps Guru

INDUSTRIAL
- Amazon Monitron
- Amazon Lookout for Equipment
- Amazon Lookout for Vision

HEALTH
- AWS HealthLake
- Amazon Comprehend Medical
- Amazon Transcribe Medical
- AWS HealthOmics

Core

TEXT
- Amazon Translate
- Amazon Comprehend

SPEECH
- Amazon Polly
- Amazon Transcribe

VISION
- Amazon Textract
- Amazon Rekognition
- AWS Panorama

Tuner

AMAZON SAGEMAKER

JUMPSTART
Use pre-built models in SageMaker

CANVAS
No-code ML for business analysts

STUDIO LAB
Learn ML

GROUND TRUTH
Label data

STUDIO IDE
- Prepare data
- Store features
- Geospatial ML
- Build with notebooks
- Train models
- Tune parameters
- Deploy in production
- Manage and monitor

CI/CD | GOVERNANCE | RESPONSIBLE ML

Provider

ML FRAMEWORKS & INFRASTRUCTURE

PyTorch, TensorFlow | Amazon EC2 | CPUs | GPUs | AWS Inferentia | AWS Trainium | Habana Gaudi | FPGA

Como representar contexto/significado das palavras

Você sabe qual o significado da palavra **tezgüino**?

Exemplo apresentado em: https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf – cap 14

Como representar contexto/significado das palavras

Observe a palavra **tezgüino** em diferentes contextos:

Como representar contexto/significado das palavras

Observe a palavra **tezgüino** em diferentes contextos:

- Uma garrafa de **tezgüino** está sobre a mesa.

Como representar contexto/significado das palavras

Observe a palavra **tezgüino** em diferentes contextos:

- Uma garrafa de **tezgüino** está sobre a mesa.
- Todo mundo gosta de beber **tezgüino**.

Como representar contexto/significado das palavras

Observe a palavra **tezgüino** em diferentes contextos:

- Uma garrafa de **tezgüino** está sobre a mesa.
- Todo mundo gosta de beber **tezgüino**.
- Você pode ficar bêbado com **tezgüino**.

Como representar contexto/significado das palavras

Observe a palavra **tezgüino** em diferentes contextos:

- Uma garrafa de **tezgüino** está sobre a mesa.
- Todo mundo gosta de beber **tezgüino**.
- Você pode ficar bêbado com **tezgüino**.
- **Tezgüino** é feito de milho.

Como representar contexto/significado das palavras

Observe a palavra **tezgüino** em diferentes contextos:

- Uma garrafa de **tezgüino** está sobre a mesa.
- Todo mundo gosta de beber **tezgüino**.
- Você pode ficar bêbado com **tezgüino**.
- **Tezgüino** é feito de milho.

Consegue entender o que é tezgüino?

Como representar contexto/significado das palavras

Observe a palavra **tezgüino** em diferentes contextos:

- Uma garrafa de **tezgüino** está sobre a mesa.
- Todo mundo gosta de beber **tezgüino**.
- Você pode ficar bêbado com **tezgüino**.
- **Tezgüino** é feito de milho.

Com o contexto, conseguimos identificar do que se refere a palavra **tezgüino**.

Tezgüino:= é uma bebida alcoólica feita a base de milho.

Como representar contexto/significado das palavras

Como o cérebro faz isso?

Como representar contexto/significado das palavras

Quais outras palavras se “*encaixam*” nos slots das perguntas 1 até 4?

Como representar contexto/significado das palavras

Quais outras palavras se “*encaixam*” nos slots das perguntas 1 até 4?

1. Uma garrafa de _____ está sobre a mesa.

Como representar contexto/significado das palavras

Quais outras palavras se “*encaixam*” nos slots das perguntas 1 até 4?

1. Uma garrafa de _____ está sobre a mesa.
2. Todo mundo gosta de beber _____.

Como representar contexto/significado das palavras

Quais outras palavras se "*encaixam*" nos slots das perguntas 1 até 4?

1. Uma garrafa de _____ está sobre a mesa.
2. Todo mundo gosta de beber _____.
3. Você pode ficar bêbado com _____.

Como representar contexto/significado das palavras

Quais outras palavras se "*encaixam*" nos slots das perguntas 1 até 4?

1. Uma garrafa de _____ está sobre a mesa.
2. Todo mundo gosta de beber _____.
3. Você pode ficar bêbado com _____.
4. _____ é feito de milho.

Como representar contexto/significado das palavras

Inserindo contexto de forma manual . . .

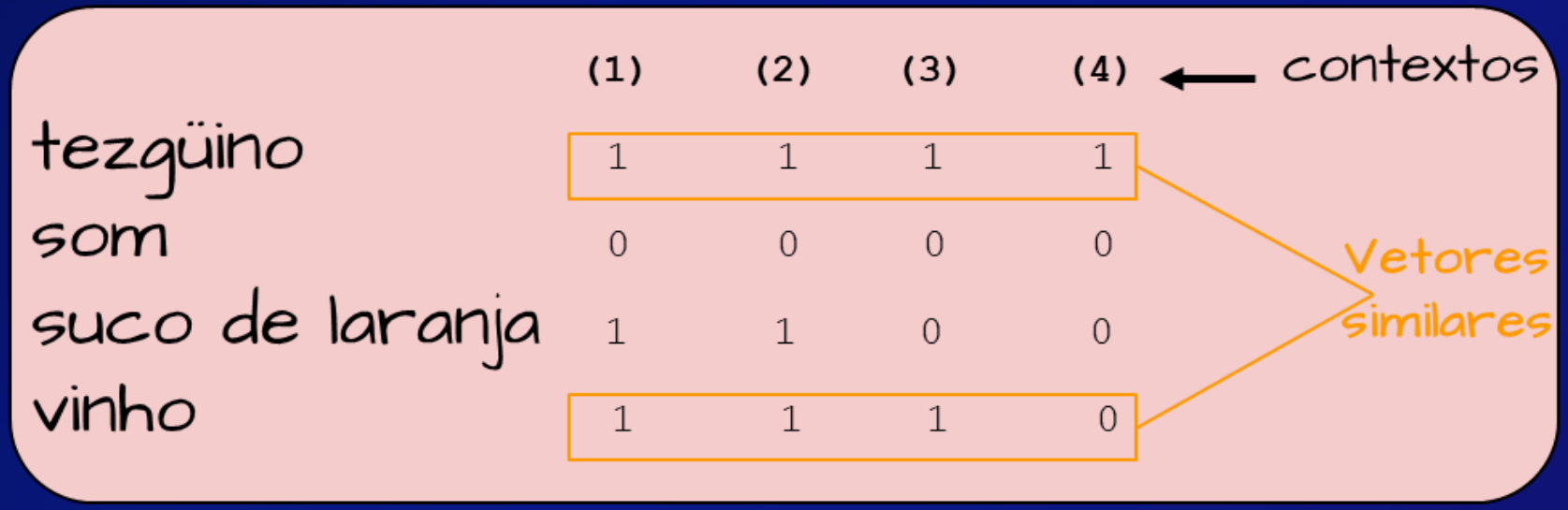
1. Uma garrafa de _____ está sobre a mesa.
2. Todo mundo gosta de beber _____.
3. Você pode ficar bêbado com _____.
4. _____ é feito de milho.

	(1)	(2)	(3)	(4) ← contextos
tezgüino	1	1	1	1
som	0	0	0	0
suco de laranja	1	1	0	0
vinho	1	1	1	0

Como representar contexto/significado das palavras

Inserindo contexto de forma manual . . .

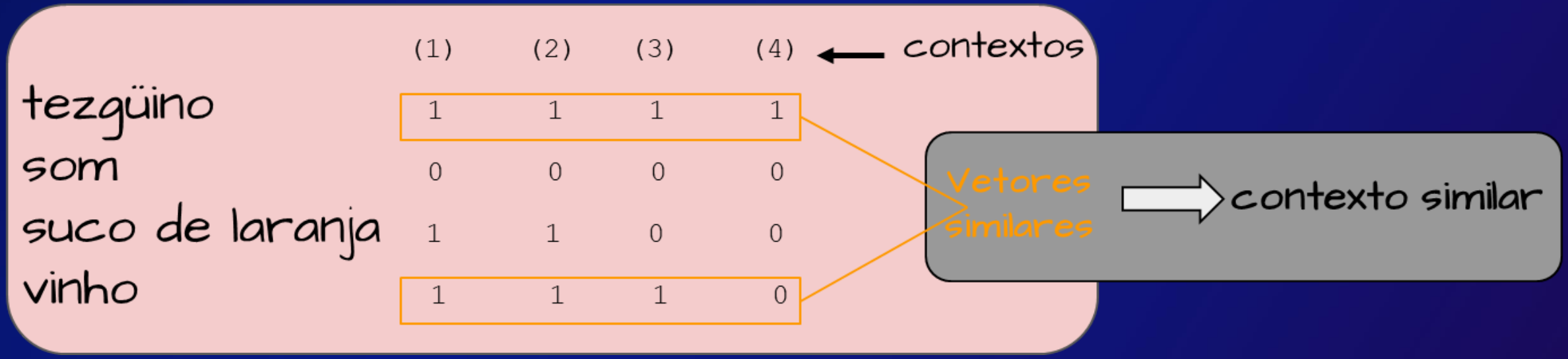
1. Uma garrafa de _____ está sobre a mesa.
2. Todo mundo gosta de beber _____.
3. Você pode ficar bêbado com _____.
4. _____ é feito de milho.



Como representar contexto/significado das palavras

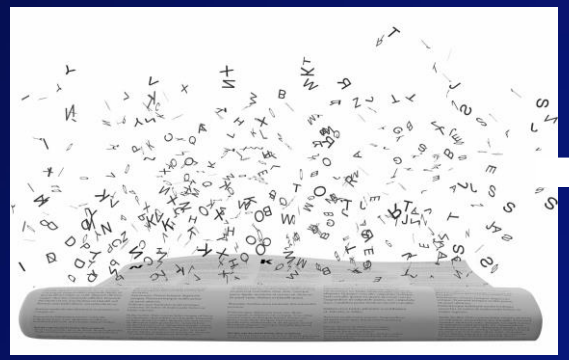
Inserindo contexto de forma manual . . .

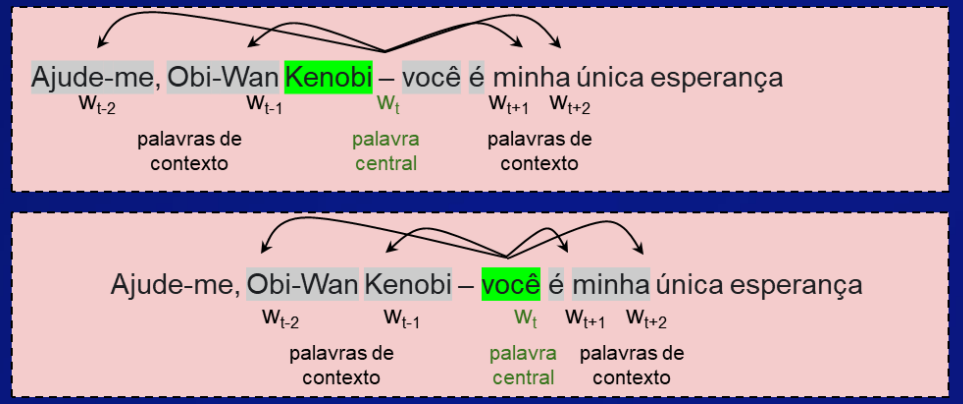
1. Uma garrafa de _____ está sobre a mesa.
2. Todo mundo gosta de beber _____.
3. Você pode ficar bêbado com _____.
4. _____ é feito de milho.

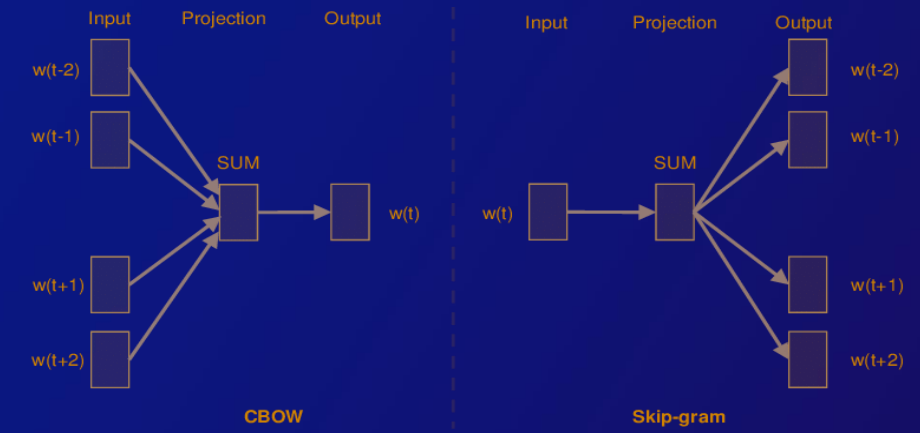


Processamento de Linguagem Natural

NATURAL LANGUAGE PROCESSING (NLP)

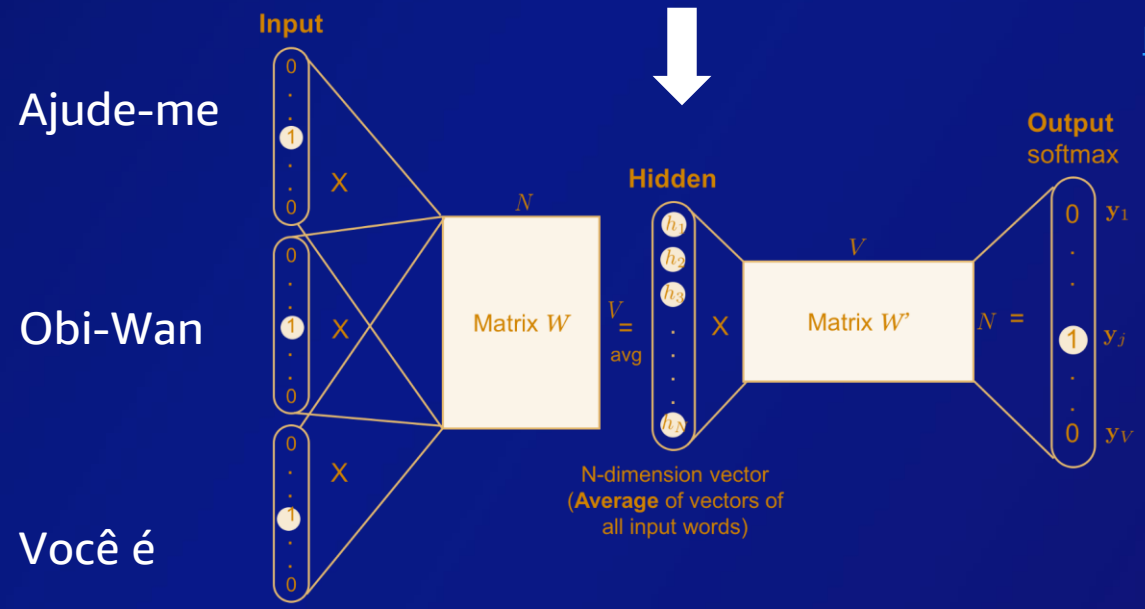






https://arxiv.org/abs/1301.3781

https://arxiv.org/pdf/1310.4546.pdf



Kenobi

Multilayer Perceptron MLP

https://dl.acm.org/doi/10.5555/1639537.1639542

Self-Supervised Learning (SSL)

https://arxiv.org/abs/2110.09327

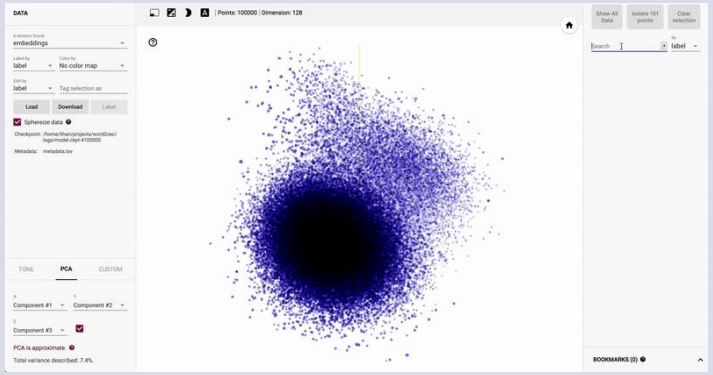
Processamento de Linguagem Natural

NATURAL LANGUAGE PROCESSING (NLP)

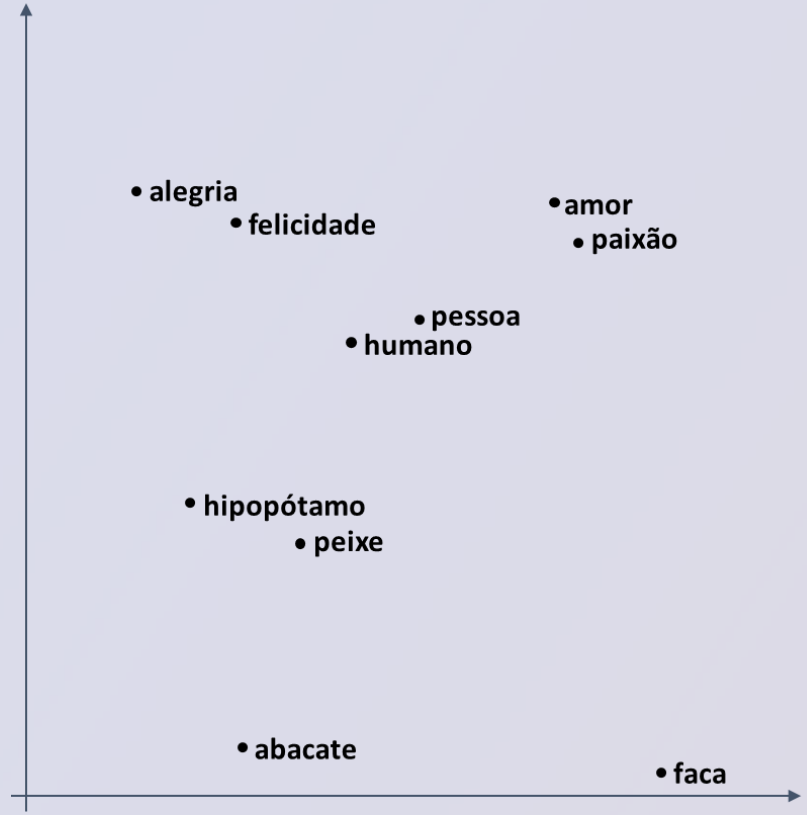
Word embedding

Representação densa

Palavras com significado similar próximas (semântica)

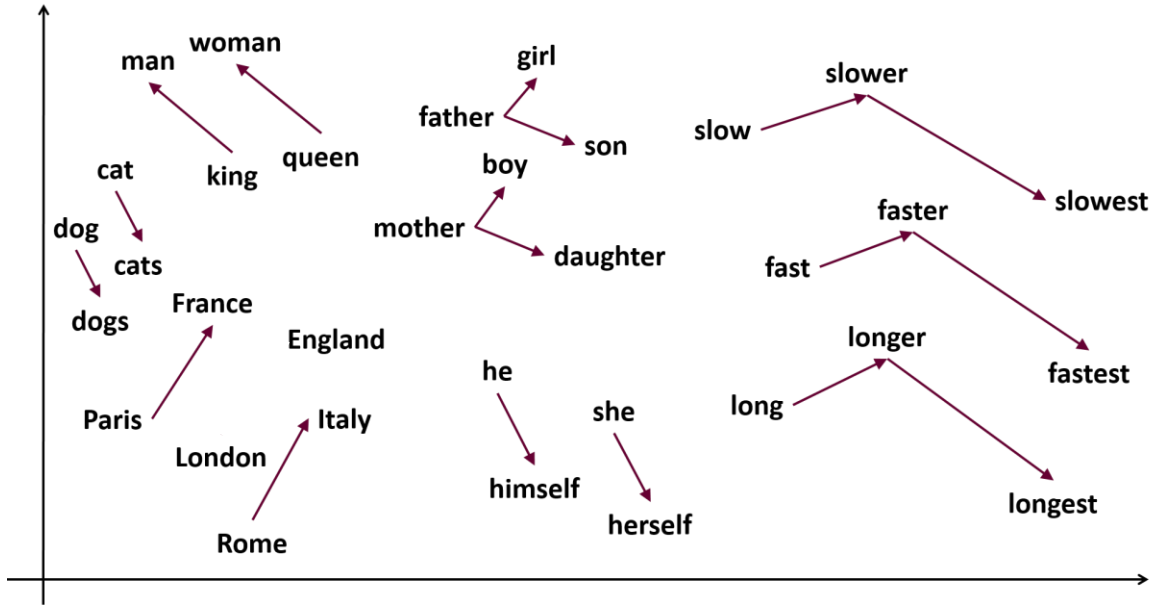


	tamanho do embedding				
a →	.1234	.4763	.0948	.9384	.7363
abafar →	.6463	.2347	.9434	.3232	.5832
acampar →	.5546	.7293	.1039	.3220	.1342
acordo →	.9482	.3293	.4201	.4403	.8494
adeus →	.9584	.7564	.9924	.2134	.1123
aéreo →	.1111	.4949	.2847	.2856	.1345
afiado →	.6754	.4342	.9683	.3219	.1103
água →	.9831	.3456	.1234	.9873	.3435
águia →	.4426	.8464	.9868	.4458	.3244
.	.				
.	.				
.	.				
zimbábue →	.9857	.7746	.2234	.3342	.9948



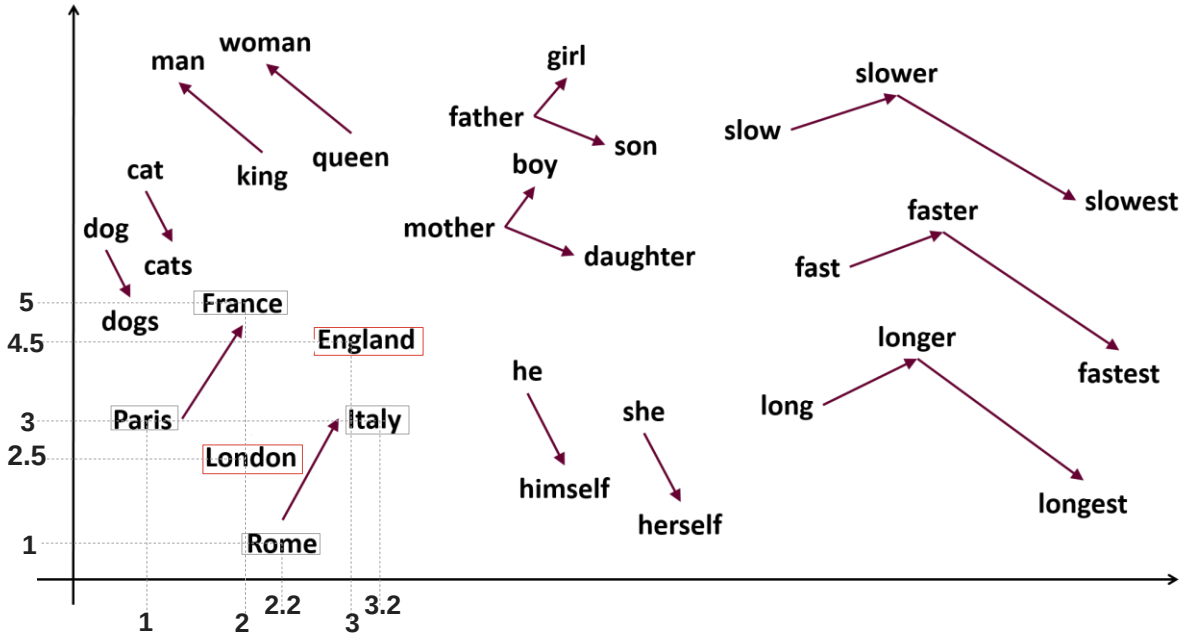
Processamento de Linguagem Natural

NATURAL LANGUAGE PROCESSING (NLP)



Processamento de Linguagem Natural

NATURAL LANGUAGE PROCESSING (NLP)



Paris [1, 3]

France [2, 5]

London [2, 2.5]

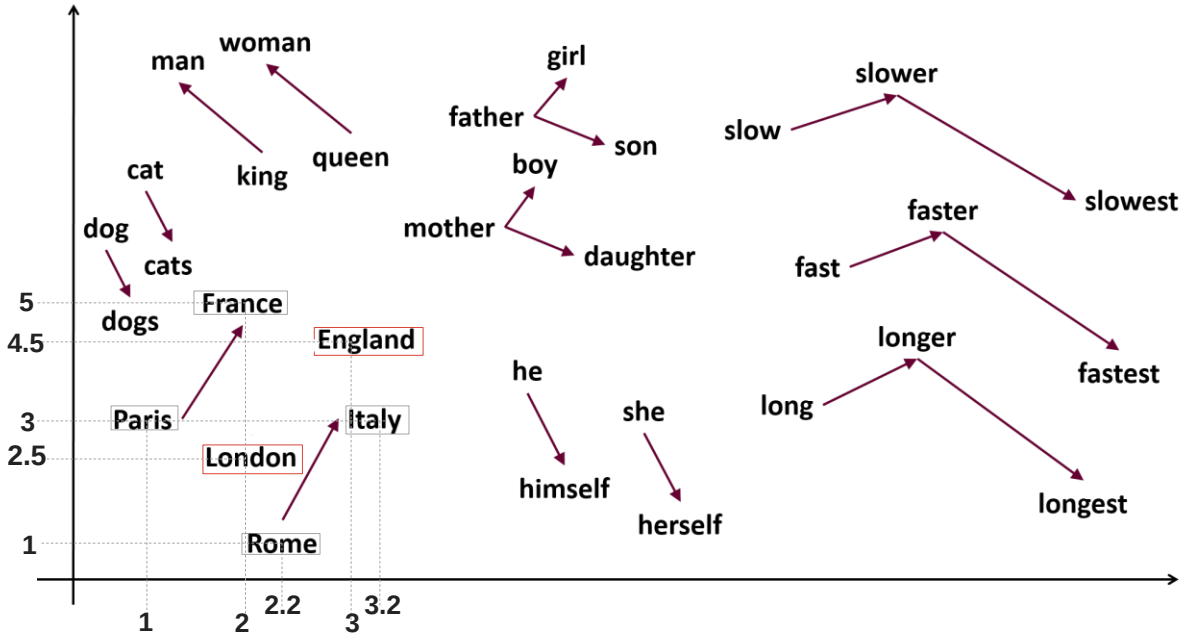
England [3, 4.5]

Rome [2.2, 1]

Italy [3.2, 3]

Processamento de Linguagem Natural

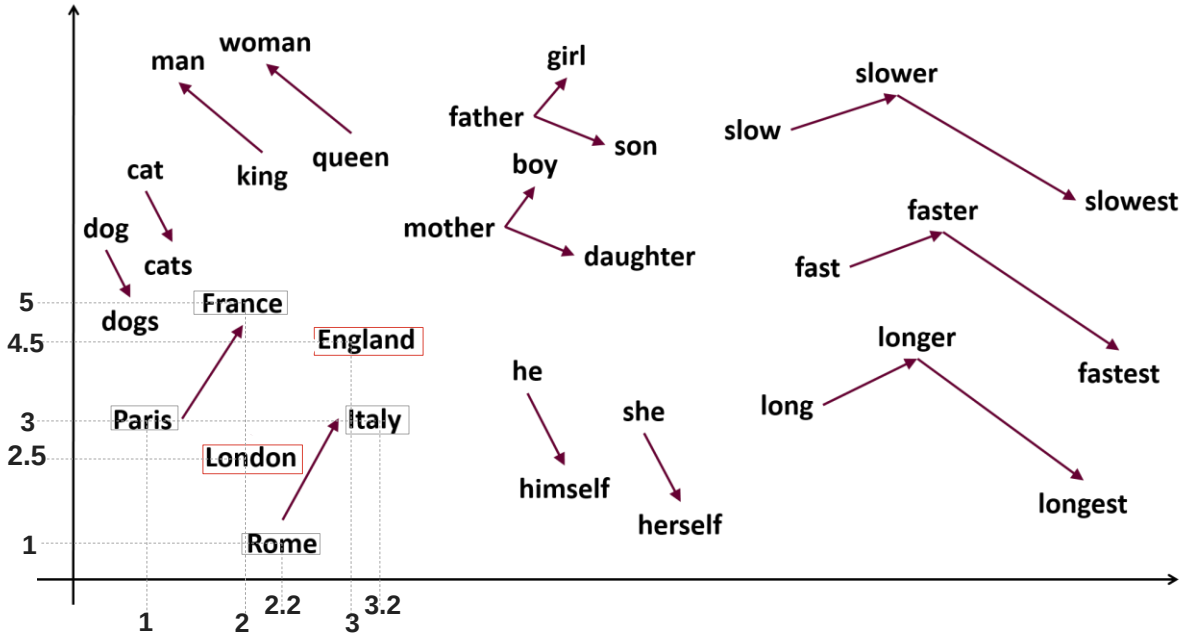
NATURAL LANGUAGE PROCESSING (NLP)



Qual a capital da Inglaterra?

Processamento de Linguagem Natural

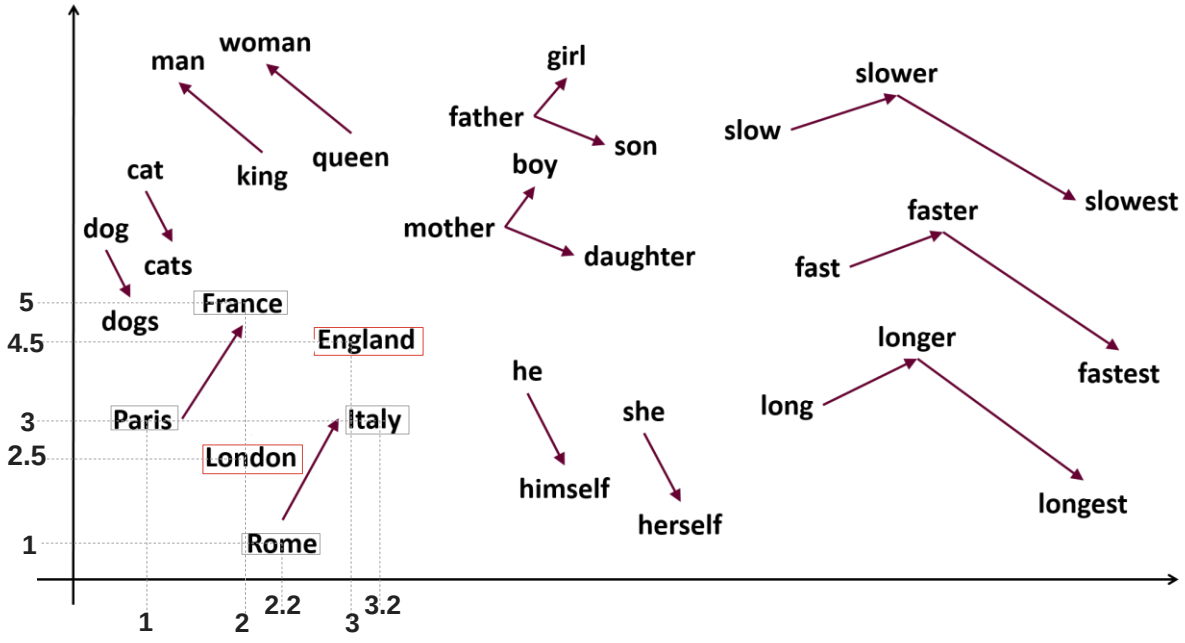
NATURAL LANGUAGE PROCESSING (NLP)



Paris – France + England = ?

Processamento de Linguagem Natural

NATURAL LANGUAGE PROCESSING (NLP)

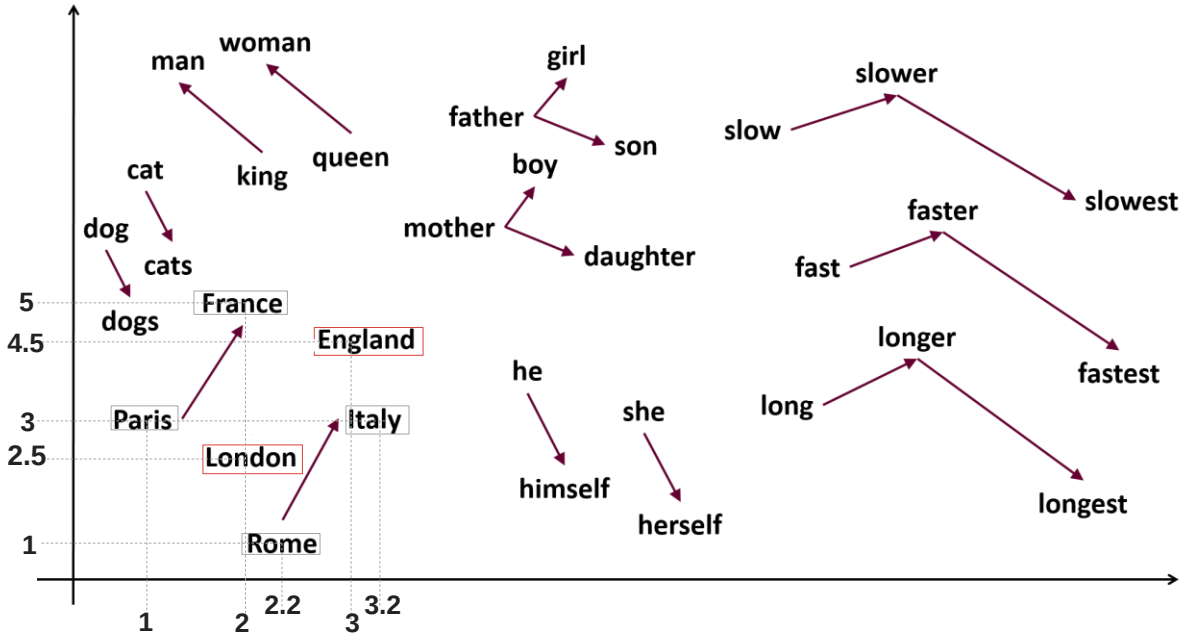


Paris – France + England = ?

Paris [1, 3]
France [2, 5]
=
Result. [–1, –2]
+
England [3, 4.5]
=
Result. [2, 2.5]
==
London [2, 2.5]

Processamento de Linguagem Natural

NATURAL LANGUAGE PROCESSING (NLP)



Qual a capital da Inglaterra?

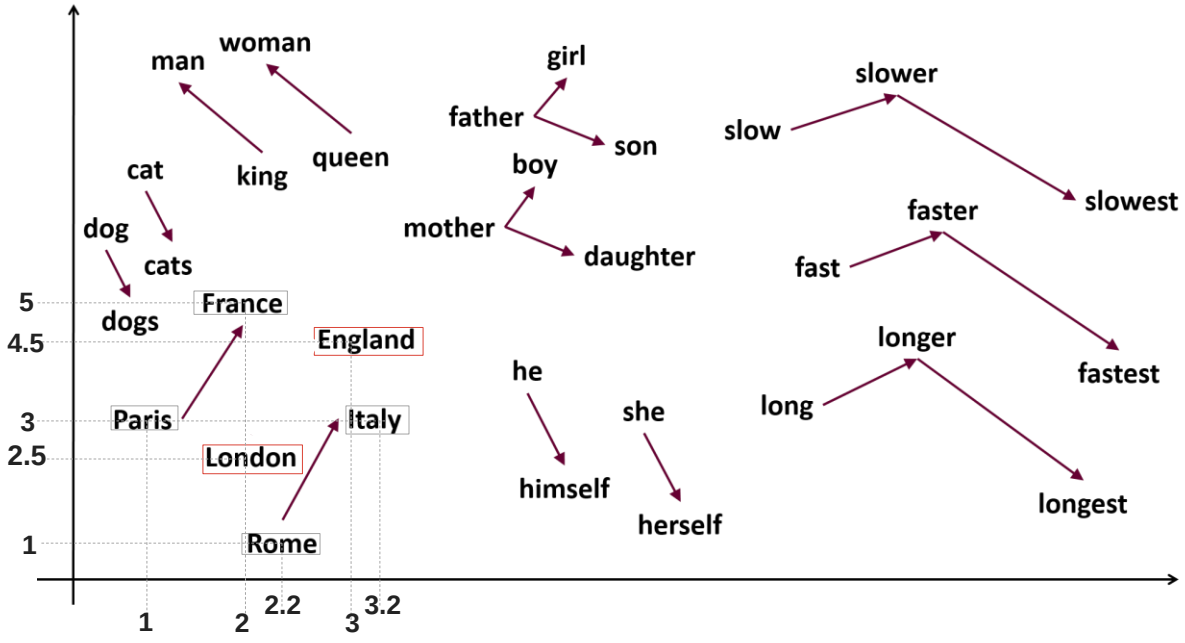
Paris – France + England = ?

Paris [1, 3]
France [2, 5]
=
Result. [–1, –2]
+
England [3, 4.5]
=
Result. [2, 2.5]
==
London [2, 2.5]

Paris – France + England = London

Processamento de Linguagem Natural

NATURAL LANGUAGE PROCESSING (NLP)



Qual a capital da Inglaterra?

Paris – France + England = ?

Paris [1, 3]
France [2, 5]
=
Result. [–1, –2]
+
England [3, 4.5]
=
Result. [2, 2.5]
==
London [2, 2.5]

Rome – Italy + England = ?

Rome [2.2, 1]
Italy [3.2, 3]
=
Result. [–1, –2]
+
England [3, 4.5]
=
Result. [2, 2.5]
==
London [2, 2.5]

Paris – France + England = London

Rome – Italy + England = London

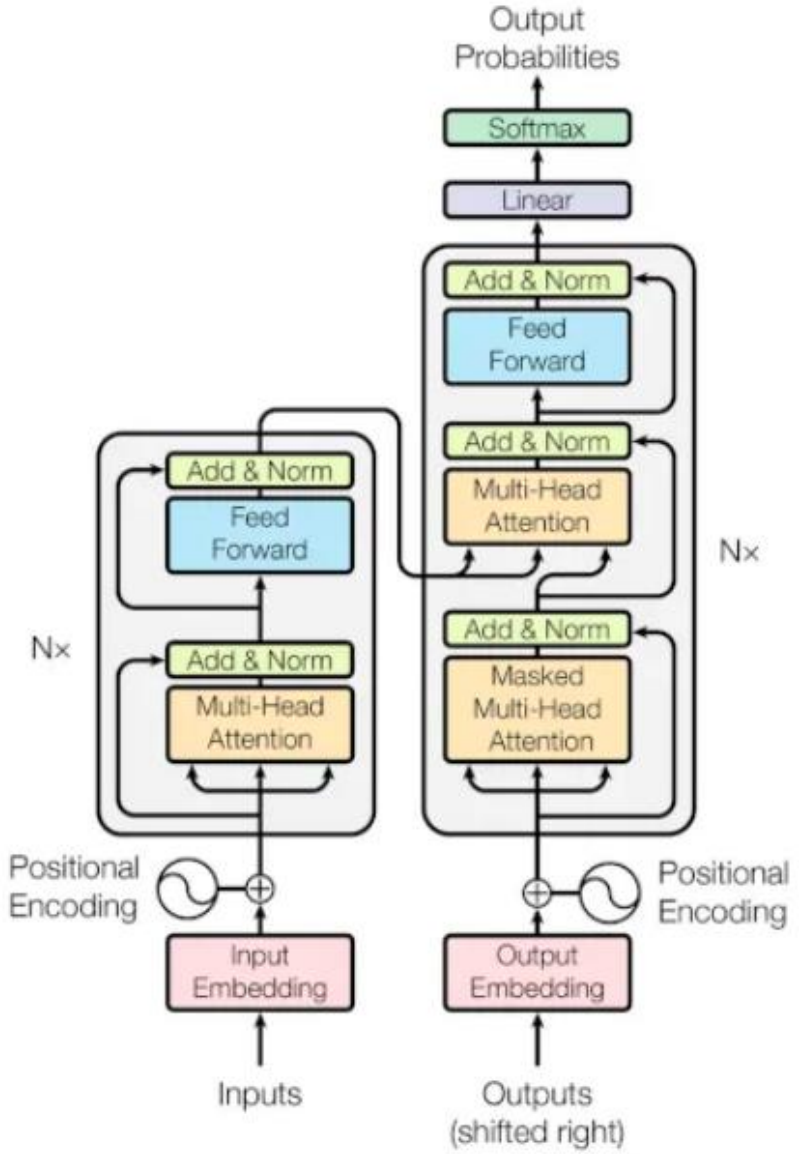
Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

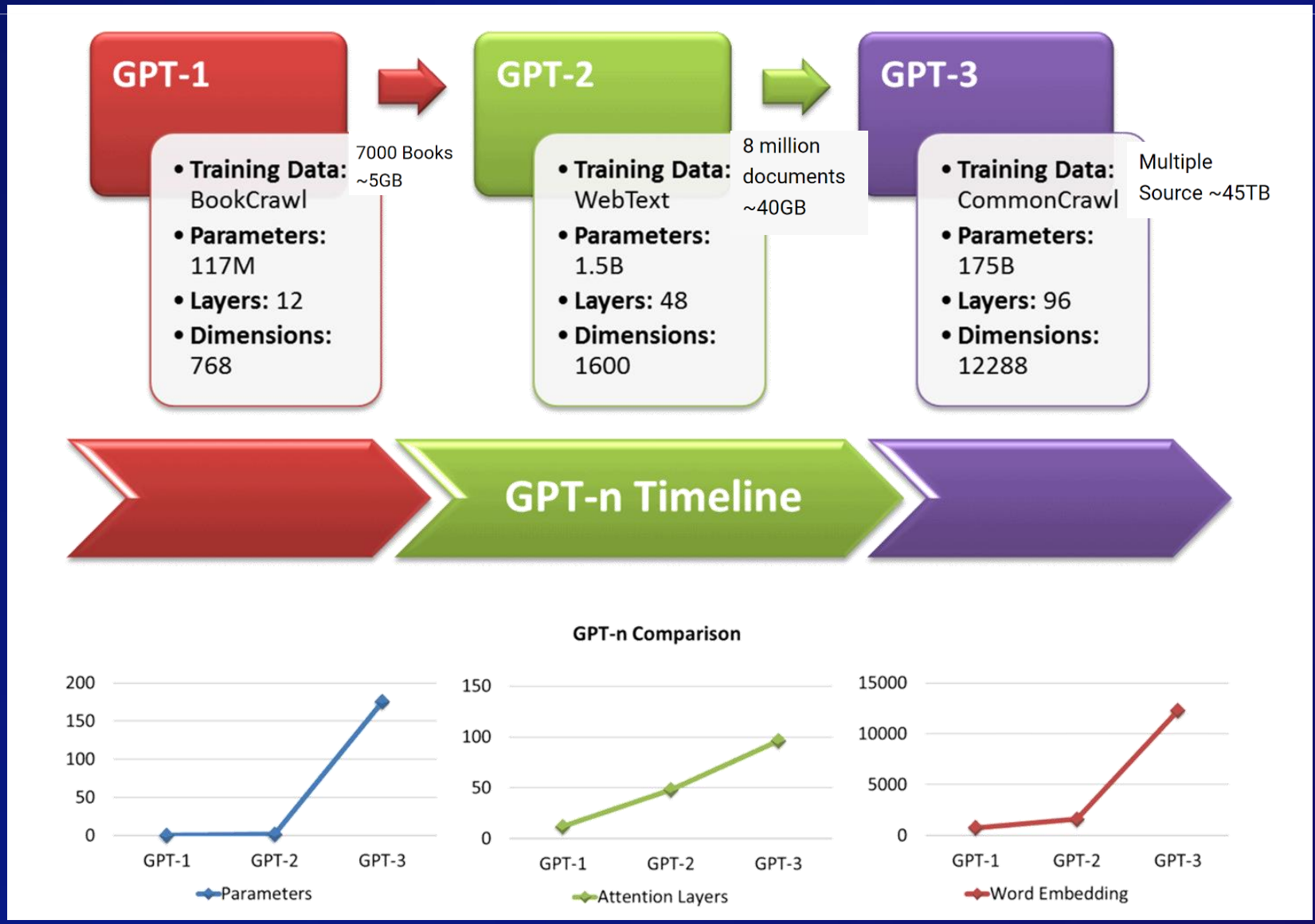
Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task,

https://arxiv.org/abs/1706.03762

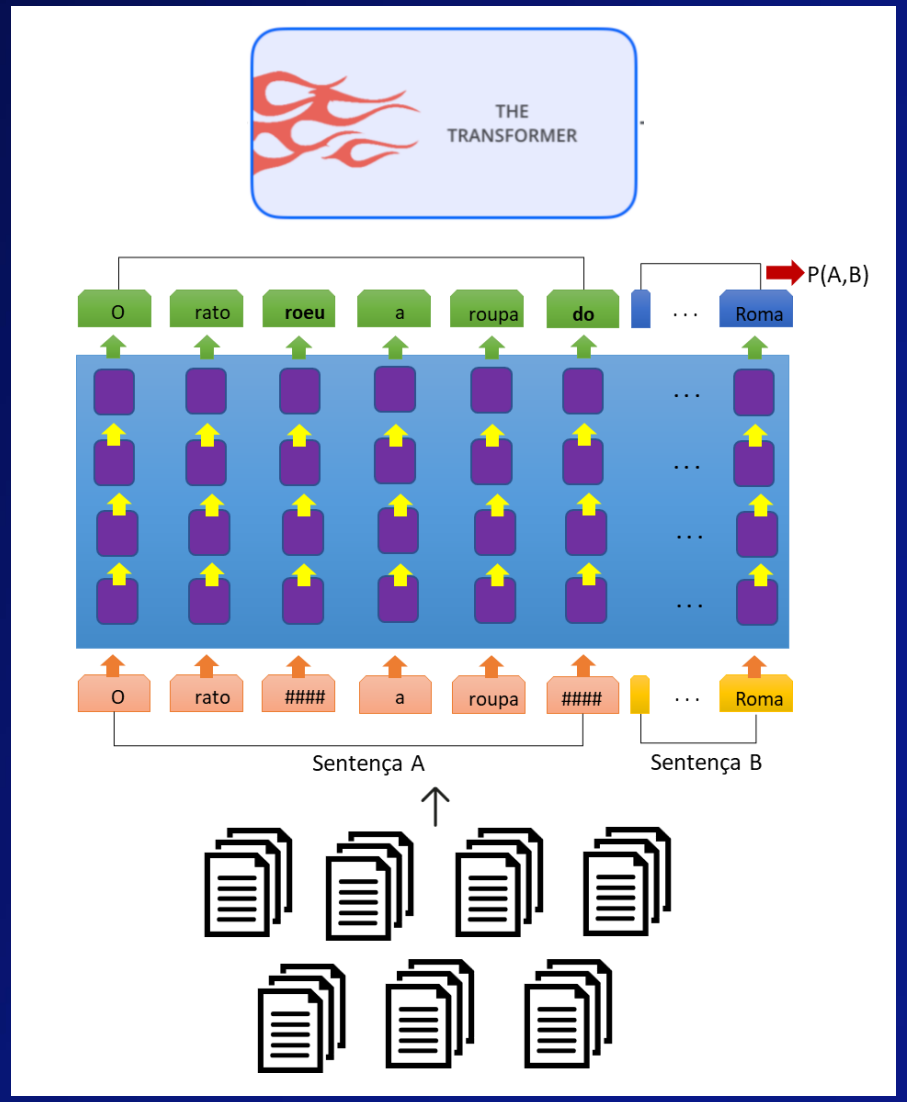


Família GPT

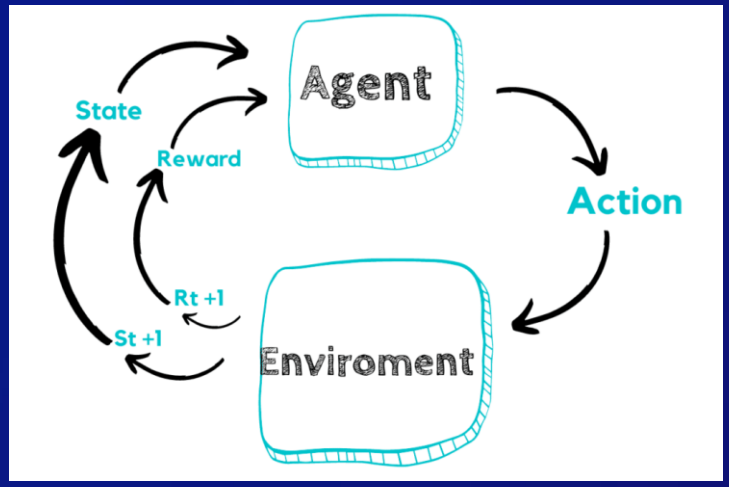


https://businessolution.org/gpt-3-statistics/

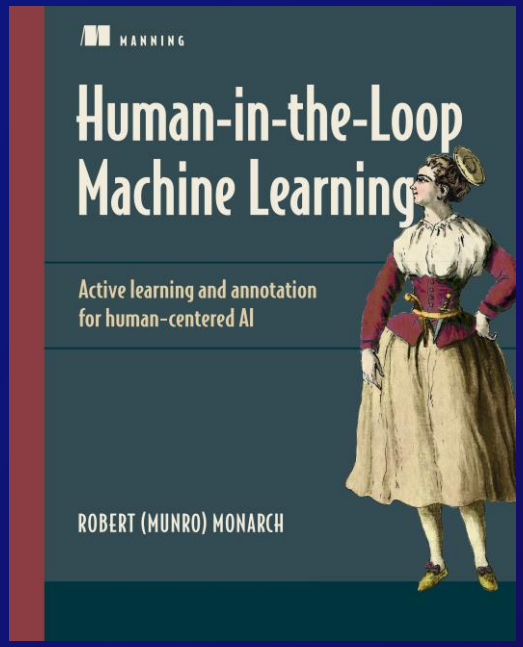
ChatGPT



+



Reinforcement Learning



Reinforcement Learning from Human Feedback (RLHF)

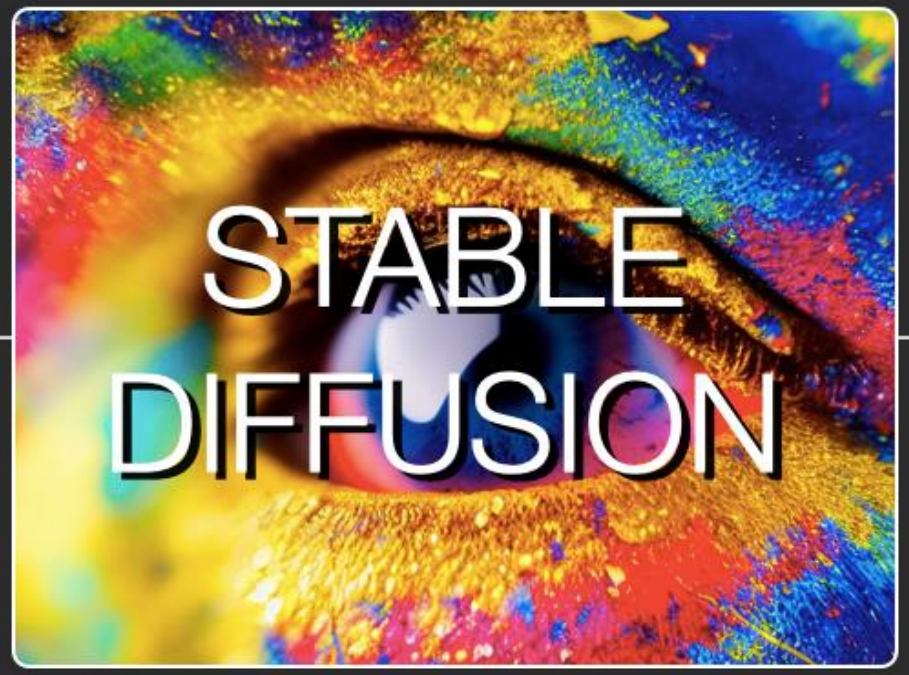
Quanto Custa os Modelos GPTs

LLM Training Costs on MosaicML Cloud			
Model	Billions of Tokens (Compute-optimal)	Days to Train on MosaicML Cloud	Approx. Cost on MosaicML Cloud
GPT-1.3B	26B	0.14	$2,000
GPT-2.7B	54B	0.48	$6,000
GPT-6.7B	134B	2.32	$30,000
GPT-13B	260B	7.43	$100,000
GPT-30B *	**610B**	**35.98**	**$450,000**
GPT-70B **	1400B	176.55	$2,500,000

Original GPT-3, which had more parameters (175 billion parameters) but was trained on less data (300 billion tokens)

https://www.mosaicml.com/blog/gpt-3-quality-for-500k#:~:text=The%20bottom%20line%3A%20it%20costs,10x%20less%20than%20people%20think

Attention Is All You Need



https://arxiv.org/abs/1706.03762

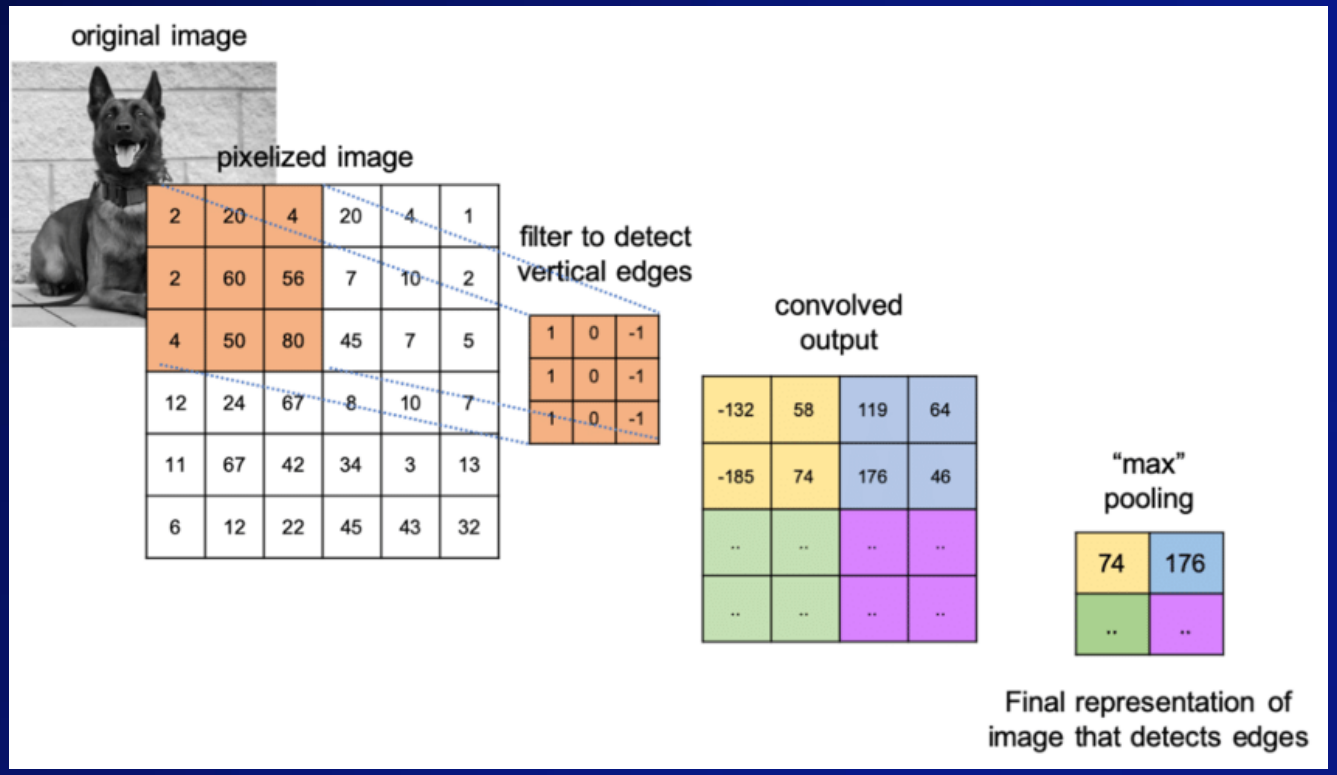
Output Probabilities

Softmax

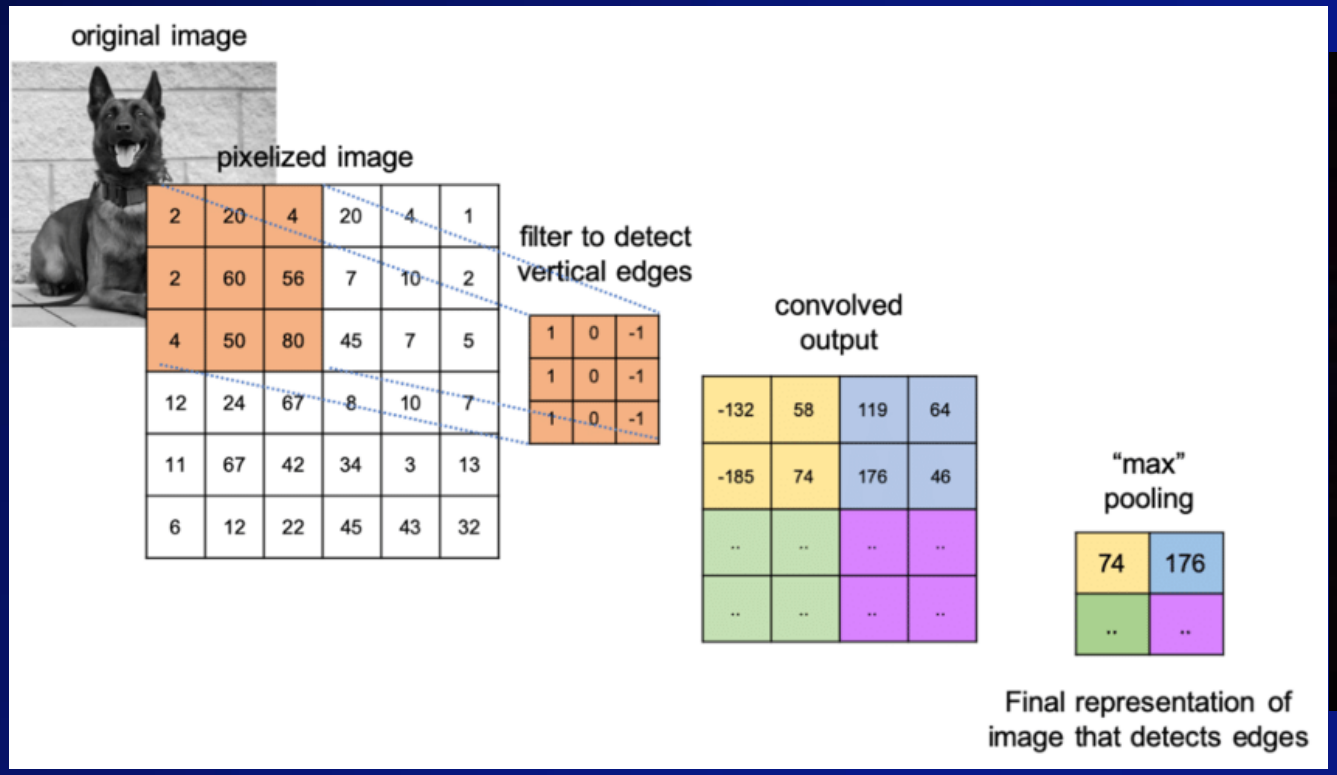
Inputs

Outputs (shifted right)

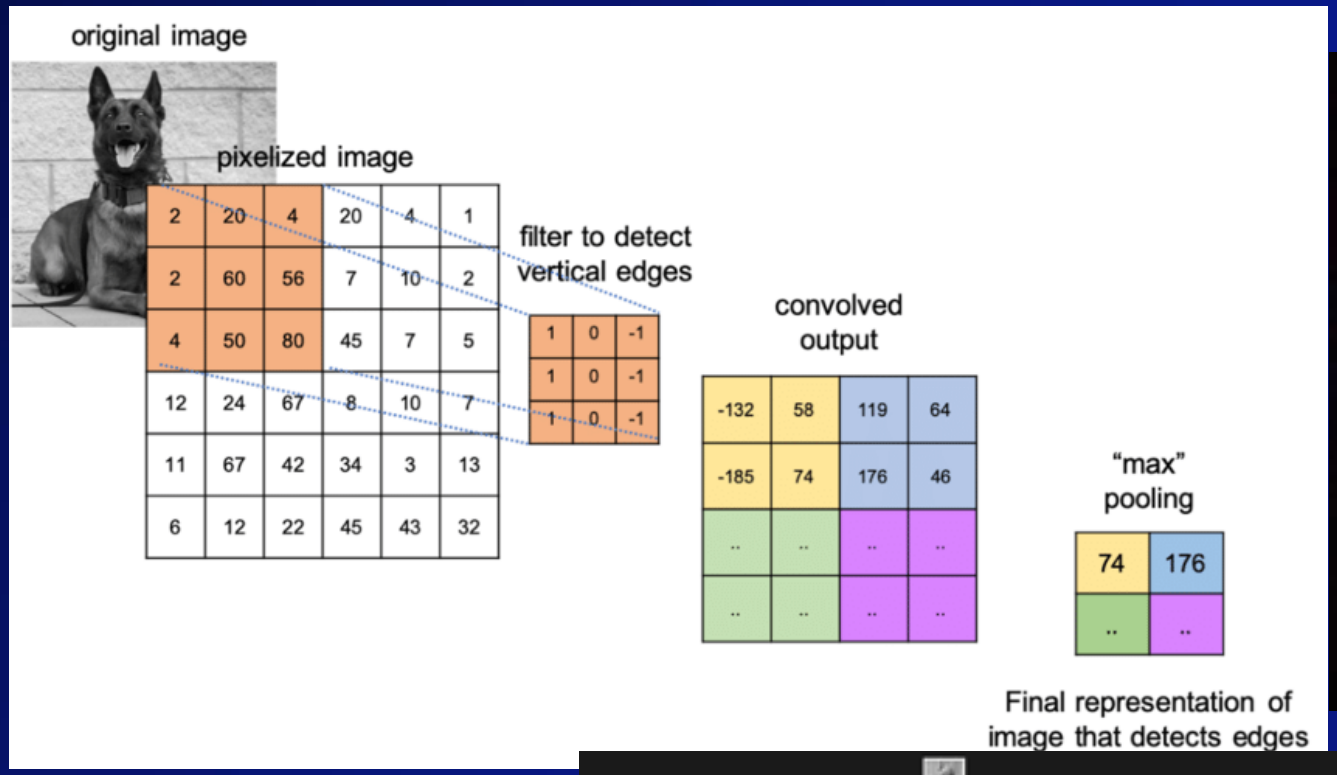
Stable Diffusion

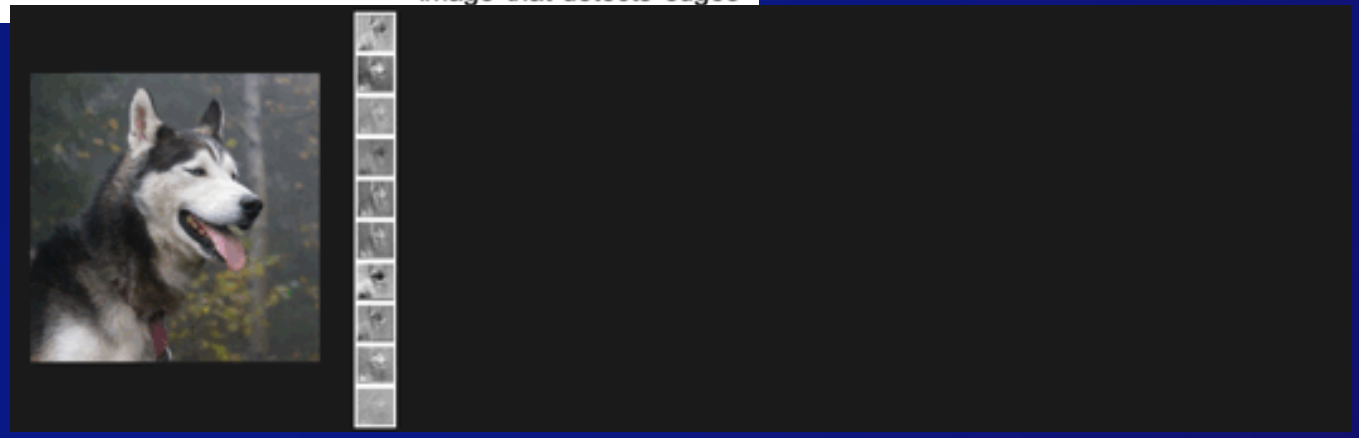


Stable Diffusion

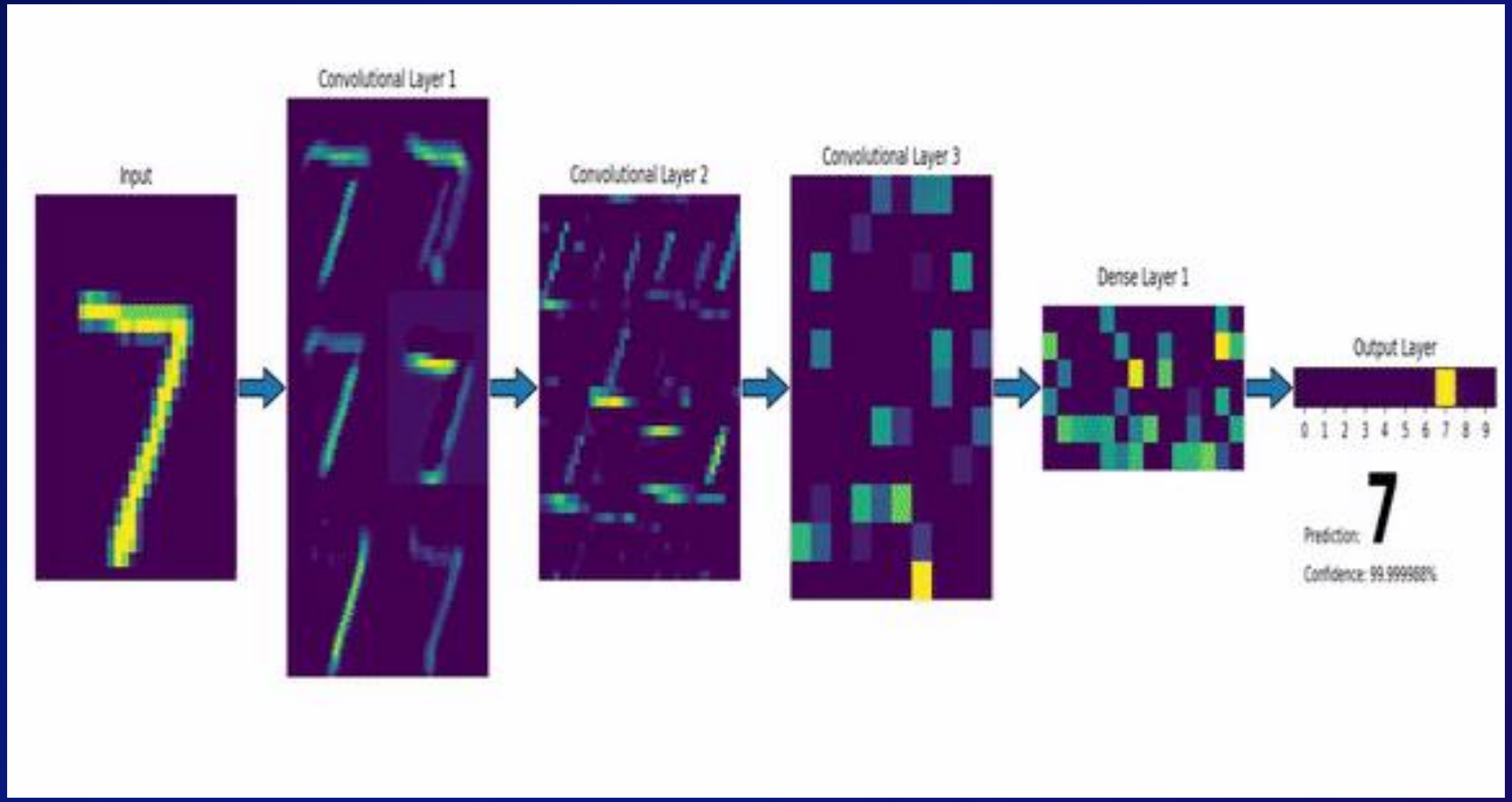


Stable Diffusion





Stable Diffusion



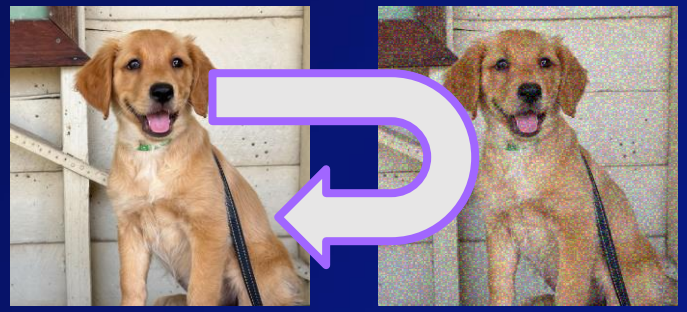
Modelos de difusão

Diffusion models

Adds noise and learns how to work backwards to the original image

Modelos de difusão

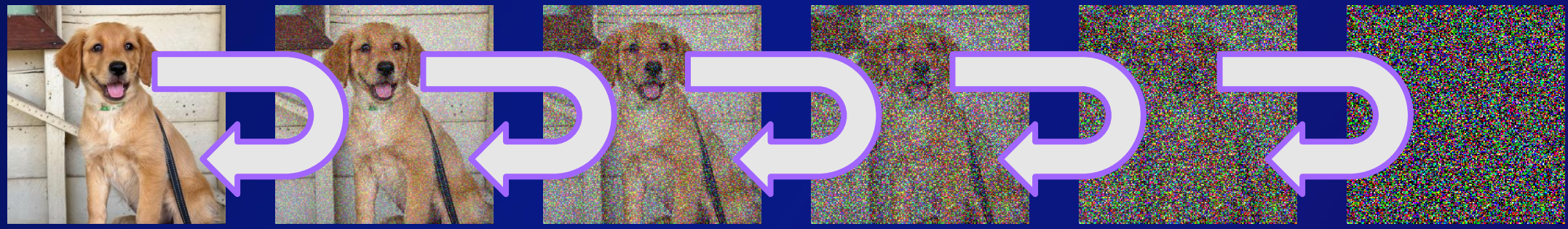
Diffusion models



Adds noise and learns how to work backwards to the original image

Modelos de difusão

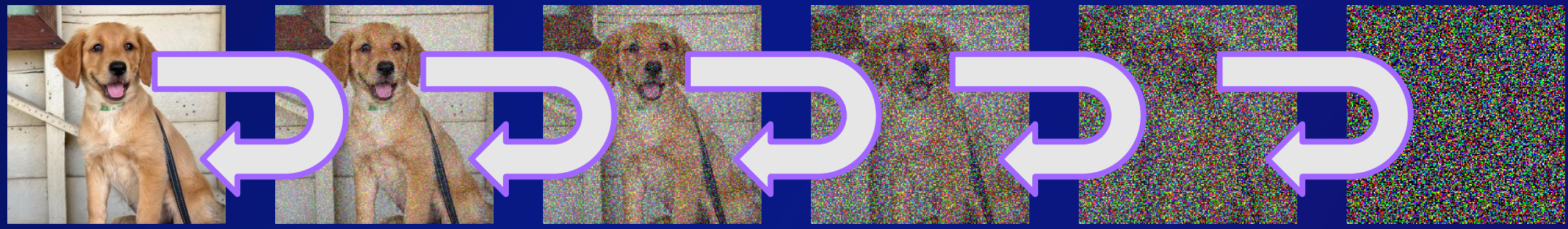
Diffusion models



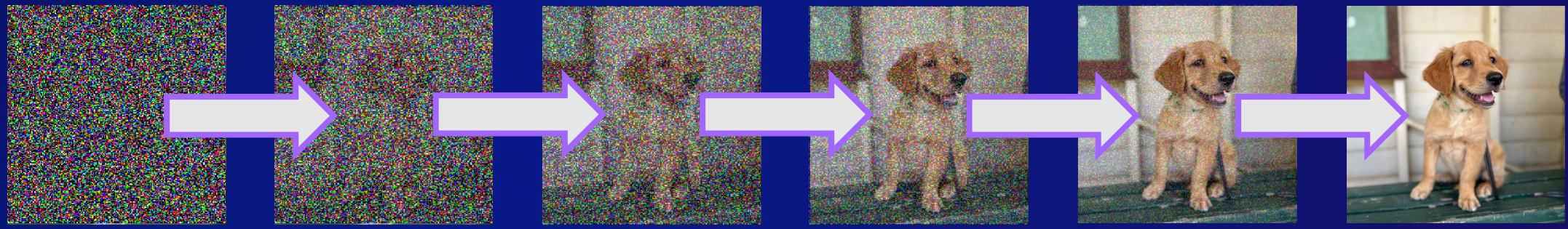
Adds noise and learns how to work backwards to the original image

Modelos de difusão

Diffusion models



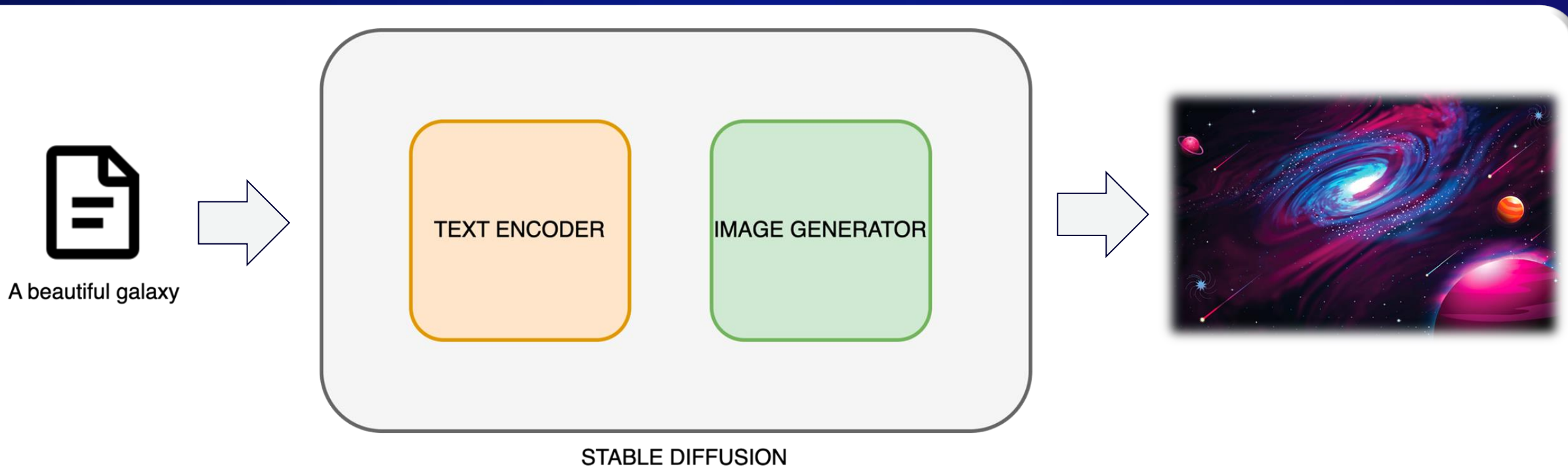
Adds noise and learns how to work backwards to the original image



Trained model works from random noise to generated an image

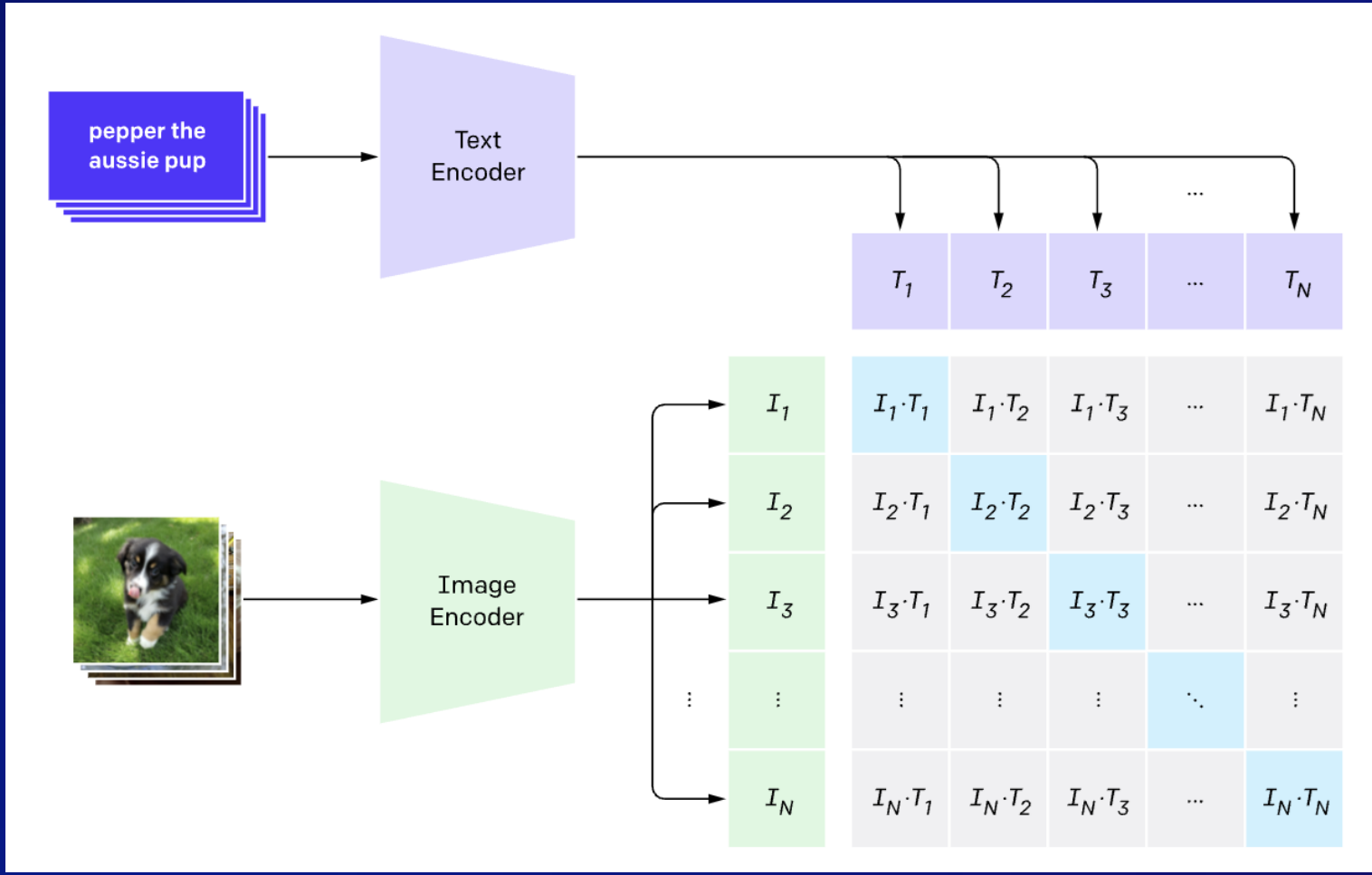
Textual Inversion

Stable Diffusion



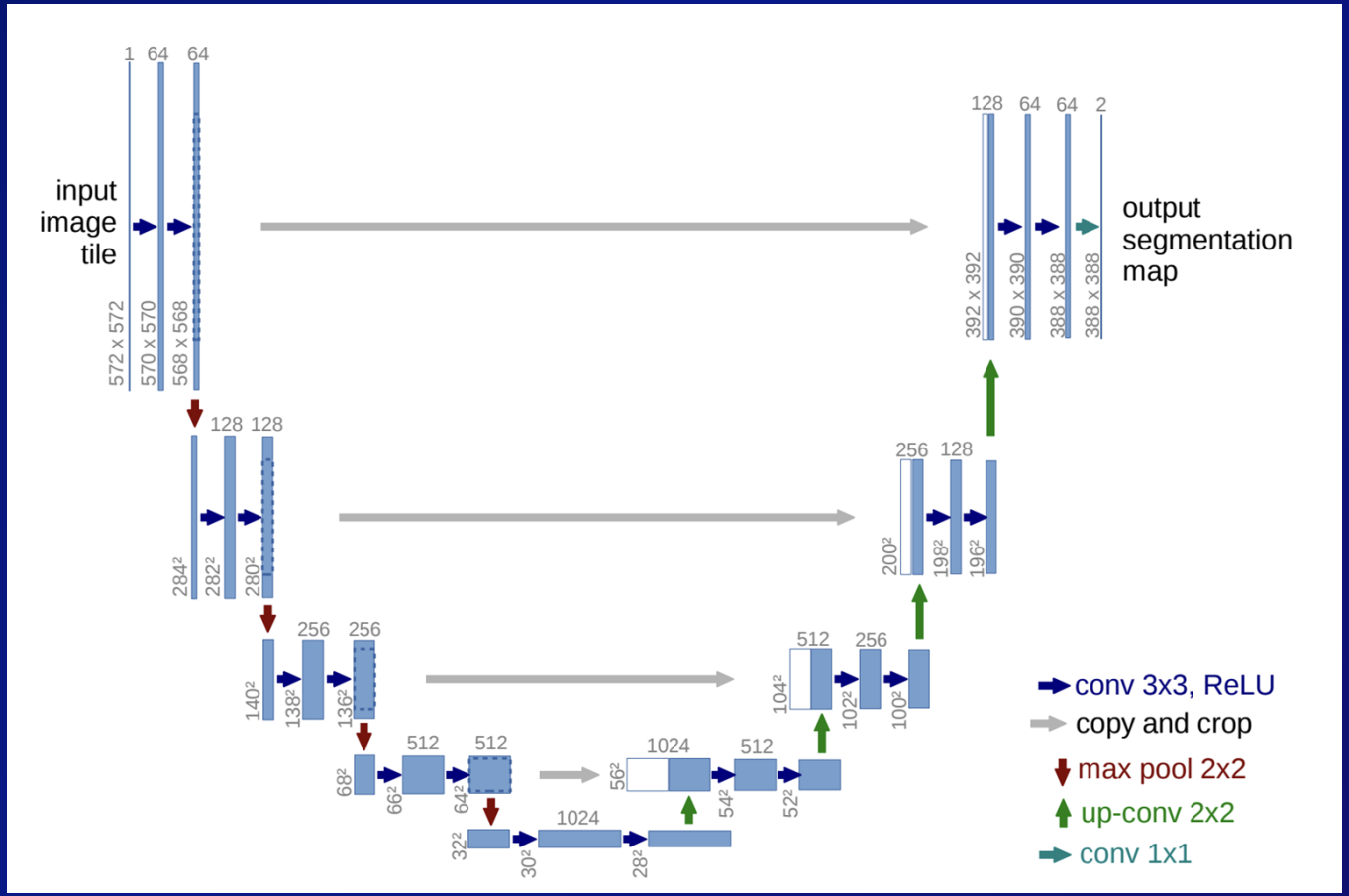
Stable Diffusion

Text Encoder – CLIP



Stable Diffusion

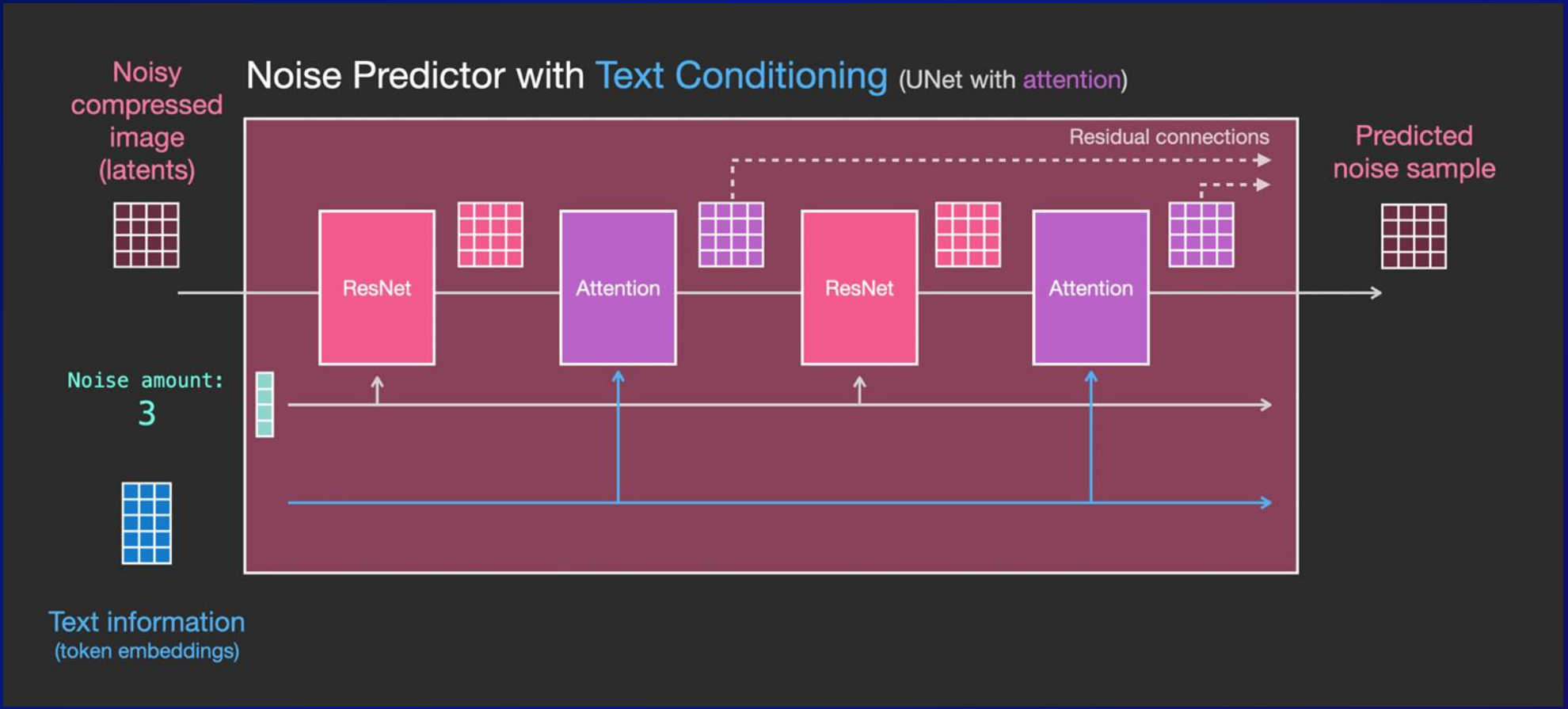
Preditor de ruído – U-Net Condicionada



Olaf Ronneberger, Philipp Fischer, and Thomas Brox: U-Net: Convolutional Networks for Biomedical Image Segmentation

Stable Diffusion

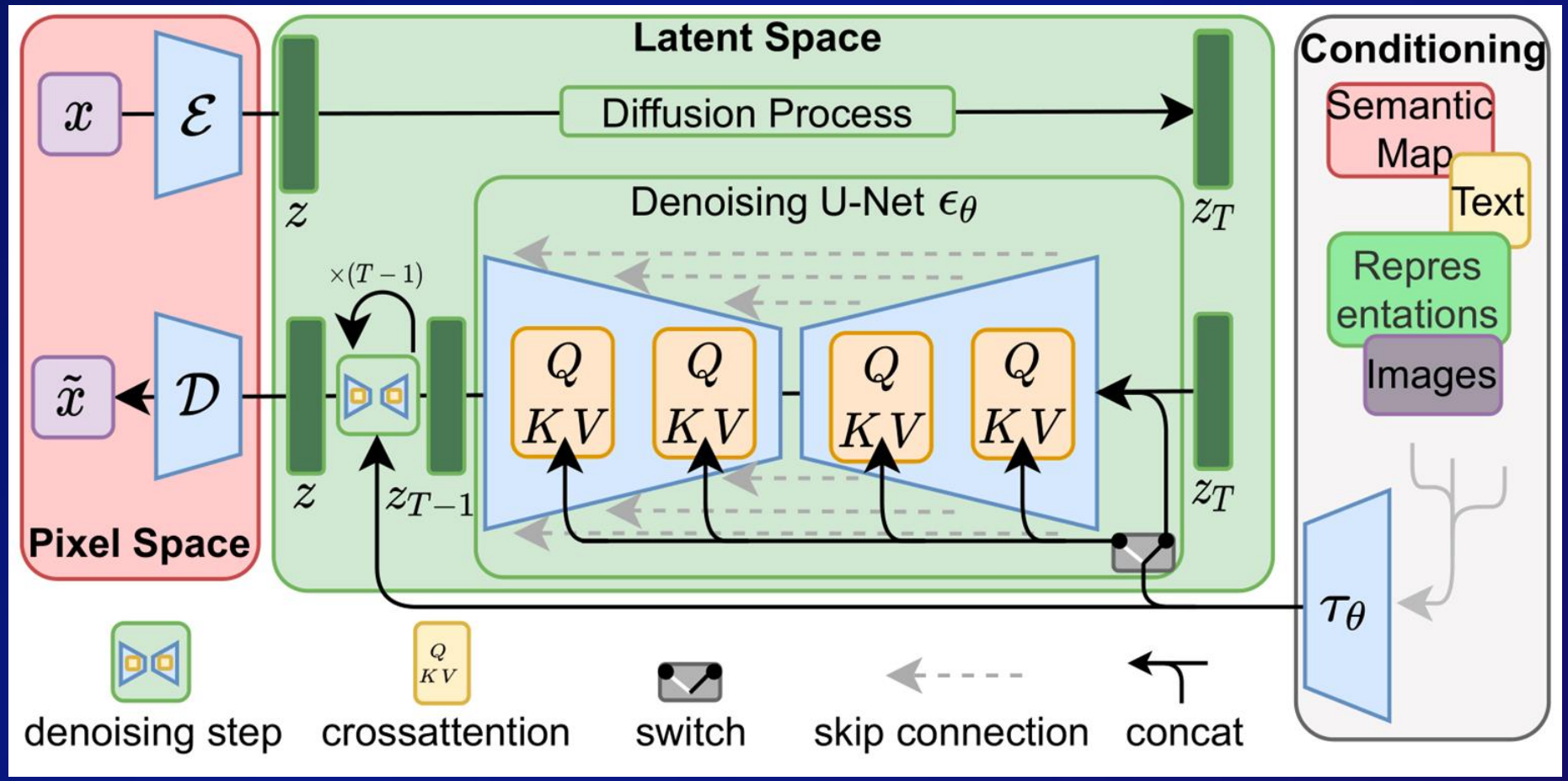
Preditor de ruído – U-Net Condicionada



The Illustrated Stable Diffusion

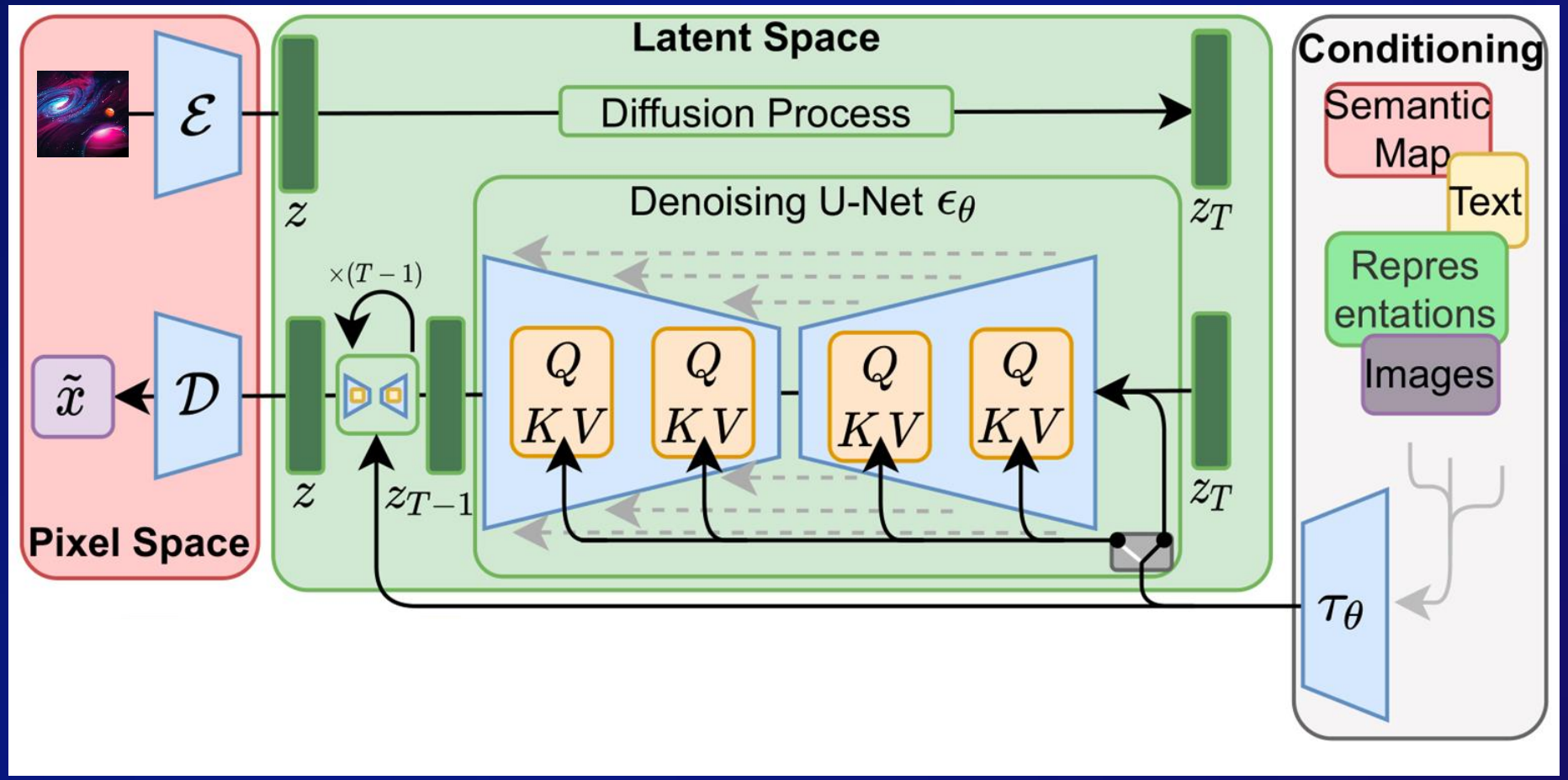
Stable Diffusion

Treinamento



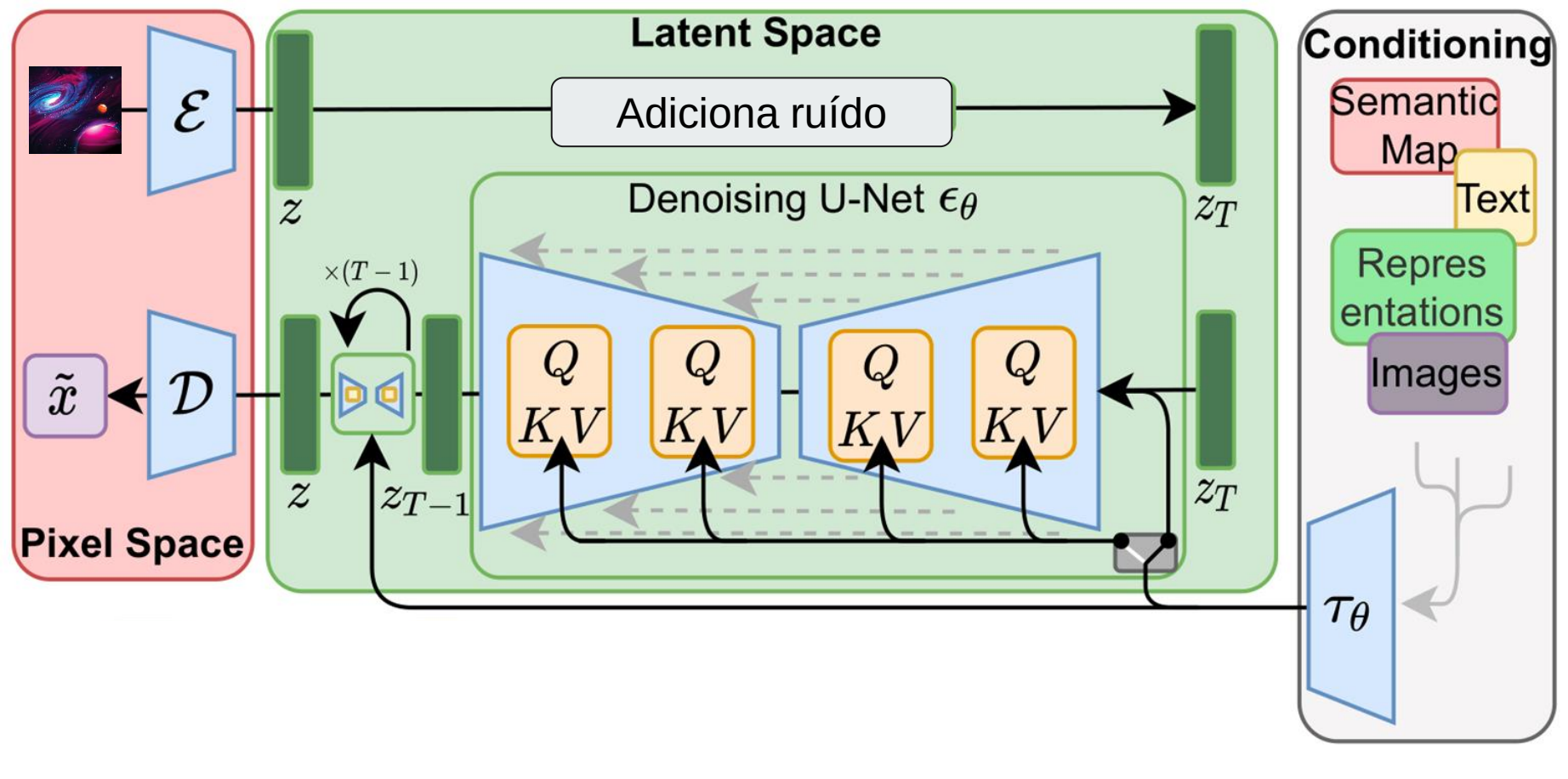
Stable Diffusion

Treinamento



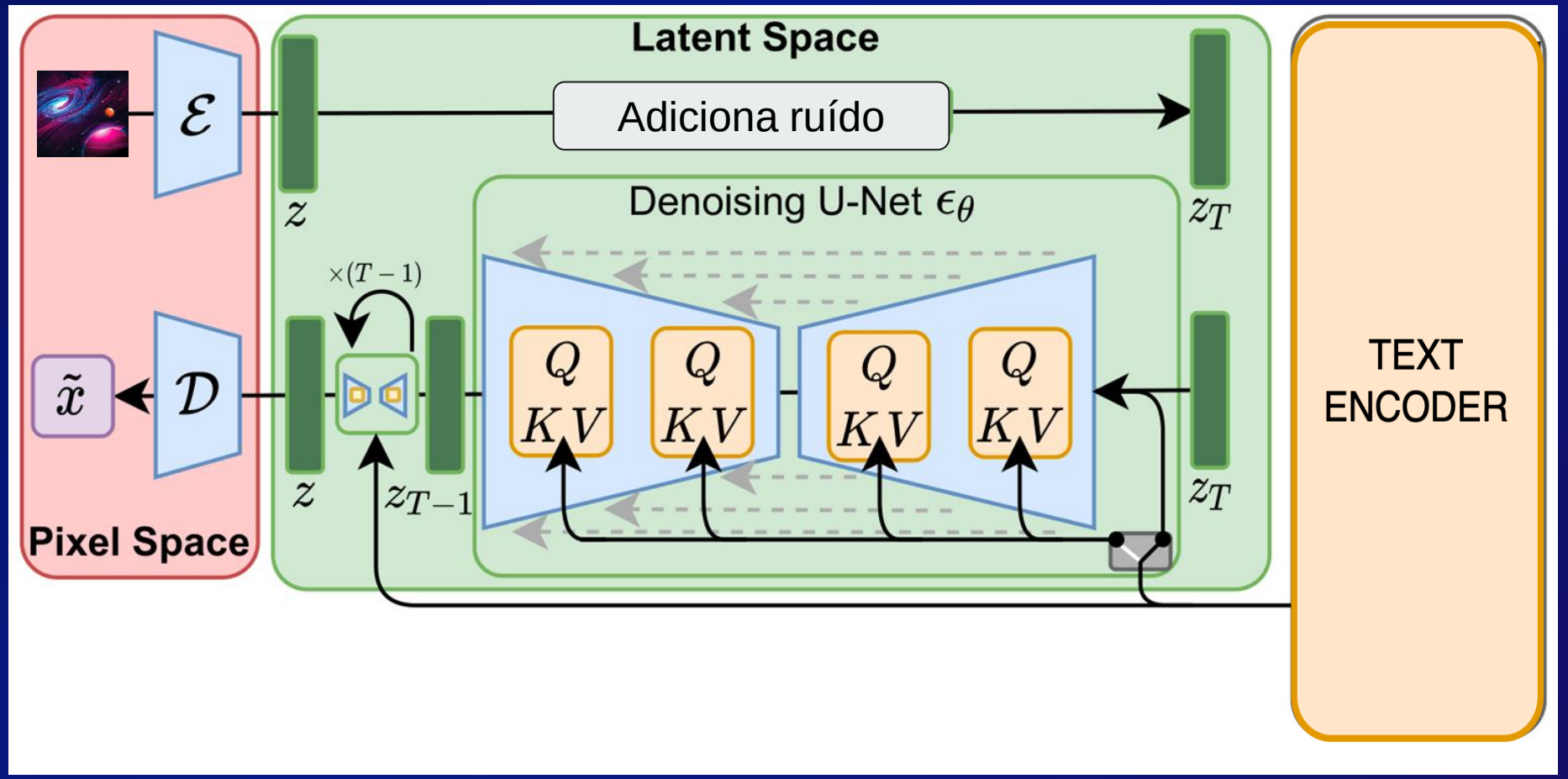
Stable Diffusion

Treinamento



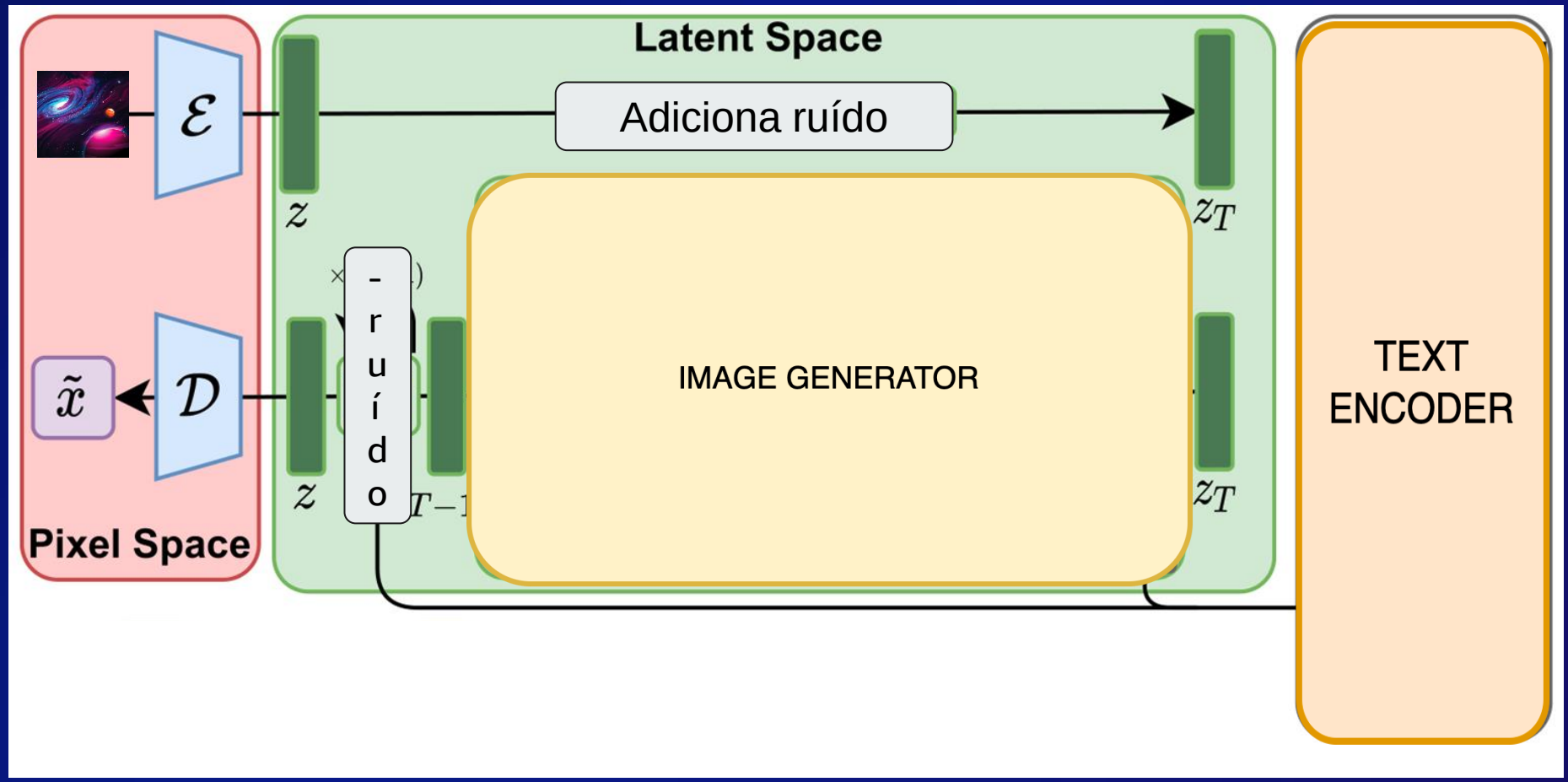
Stable Diffusion

Treinamento



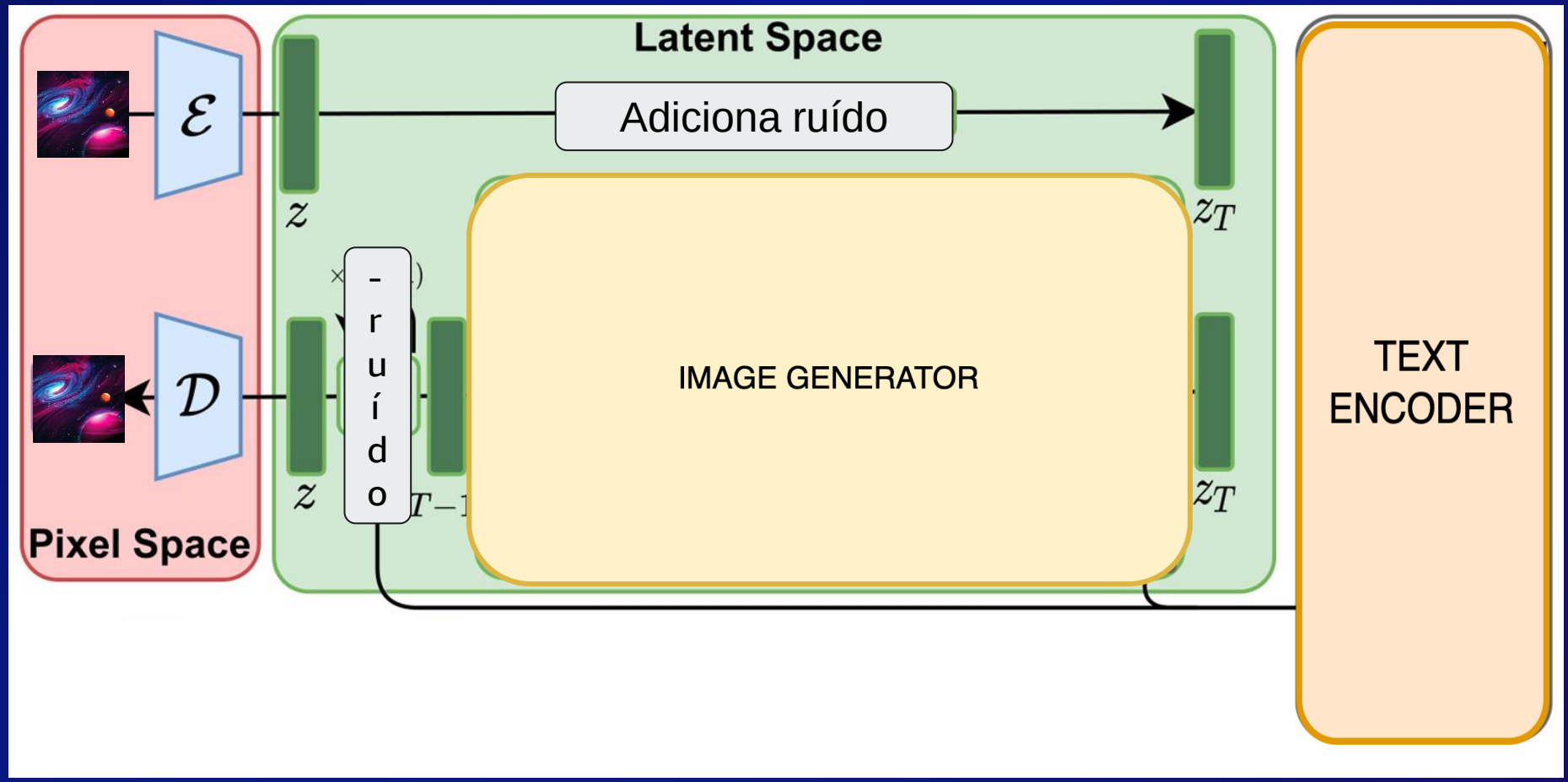
Stable Diffusion

Treinamento



Stable Diffusion

Treinamento



Textual Inversion

Dados fora do treinamento

Textual Inversion

Dados fora do treinamento

- Um gato em uma praça ✅

Textual Inversion

Dados fora do treinamento

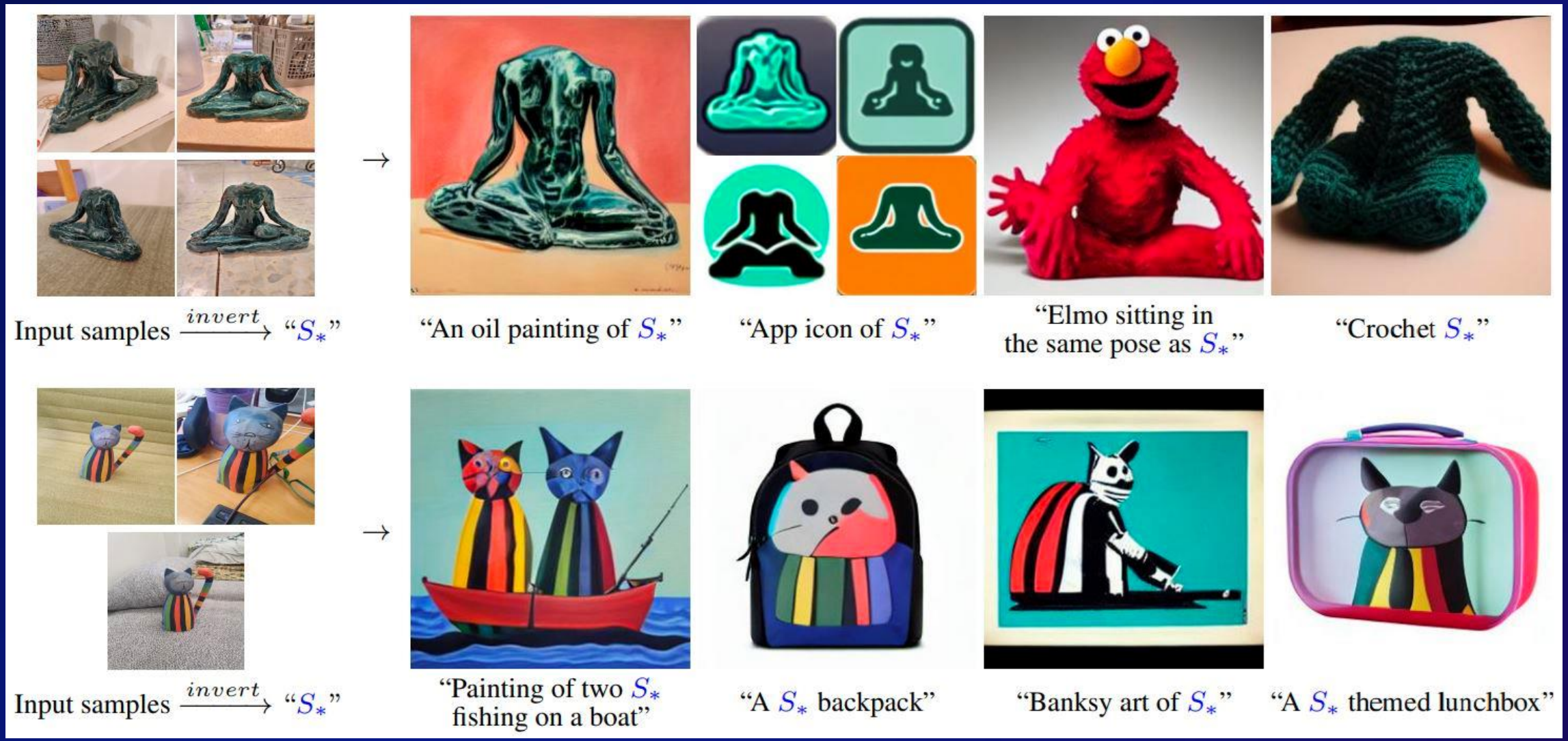
- Um gato em uma praça ✅
- Um cachorro na grama ✅

Textual Inversion

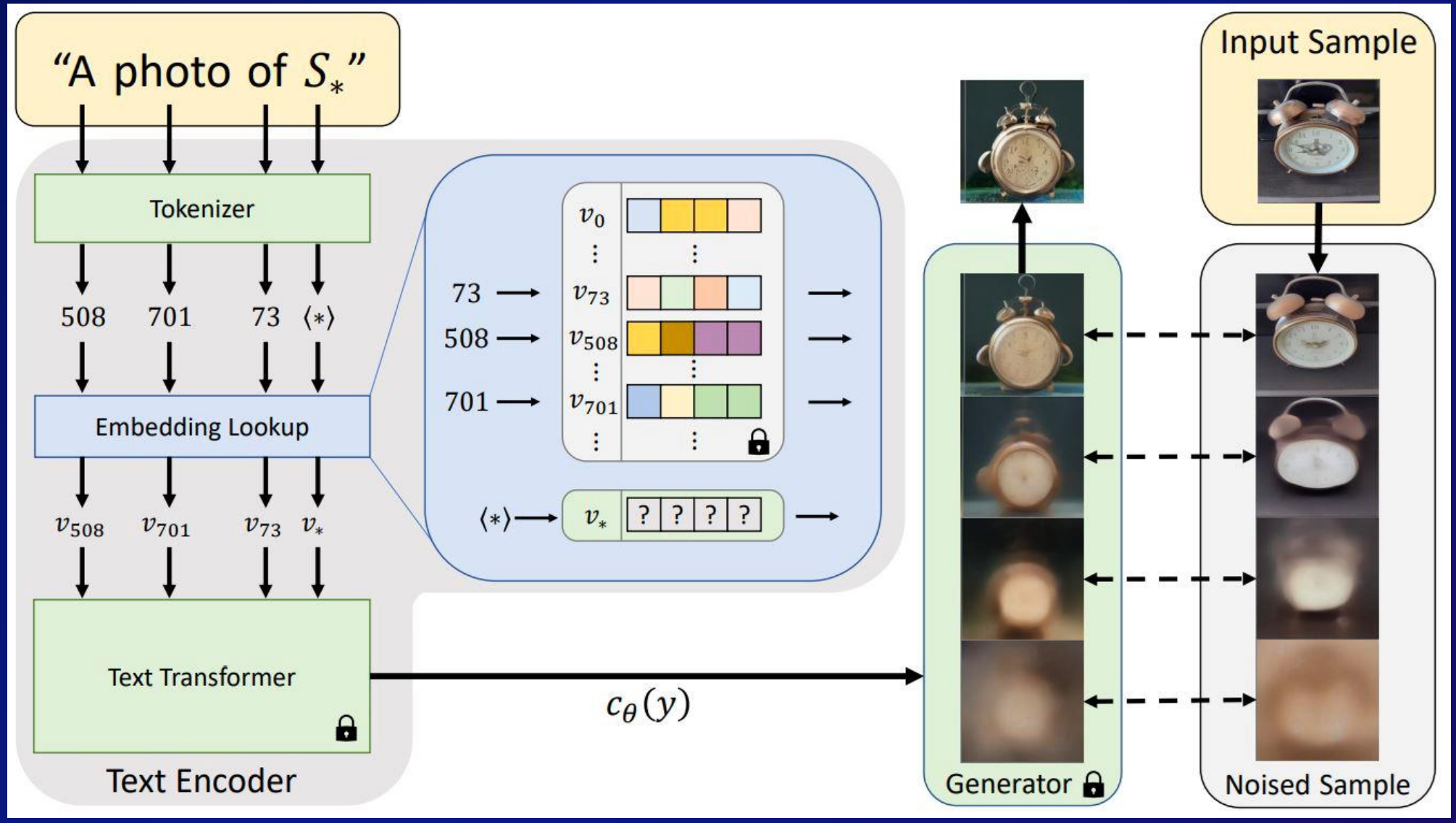
Dados fora do treinamento

- Um gato em uma praça ✅
- Um cachorro na grama ✅
- Um coelho no mato ❌

Textual Inversion



Textual Inversion



Stable Diffusion

Let’s do it!

Obrigado!

Por favor complete a pesquisa da sessão no aplicativo móvel

Vinicius Caridá
@vfcarida
https://linktr.ee/vfcarida