

N° 18054

Vision-languages models: história e aplicações no mundo real

**Gustavo Torres Custódio
Adriana Camargo de Brito**

*Palestra apresentada no
SEBRAE, 2026. 31 slides.*

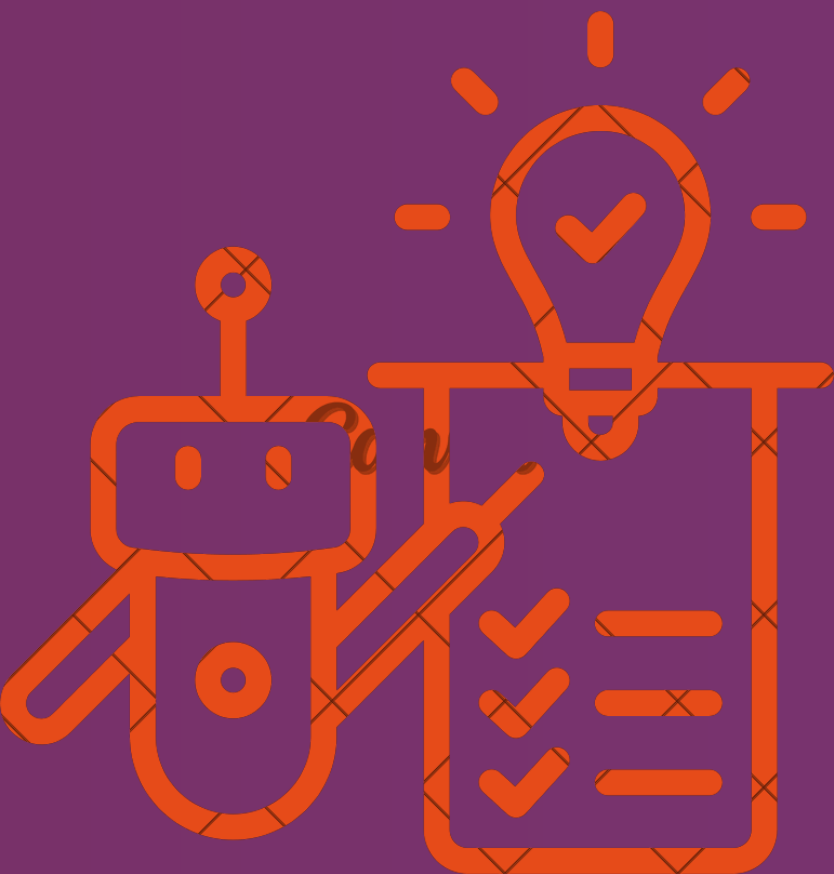
A série “Comunicação Técnica” compreende trabalhos elaborados por técnicos do IPT, apresentados em eventos, publicados em revistas especializadas ou quando seu conteúdo apresentar relevância pública.
PROIBIDO A REPRODUÇÃO, APENAS PARA CONSULTA.

VISION-LANGUAGE MODELS: HISTÓRIA E APLICAÇÕES NO MUNDO REAL



Gustavo Torres Custodio
Pesquisador IPT-SIAA

Apoio: [Adriana Camargo de Brito](#)



Realização:

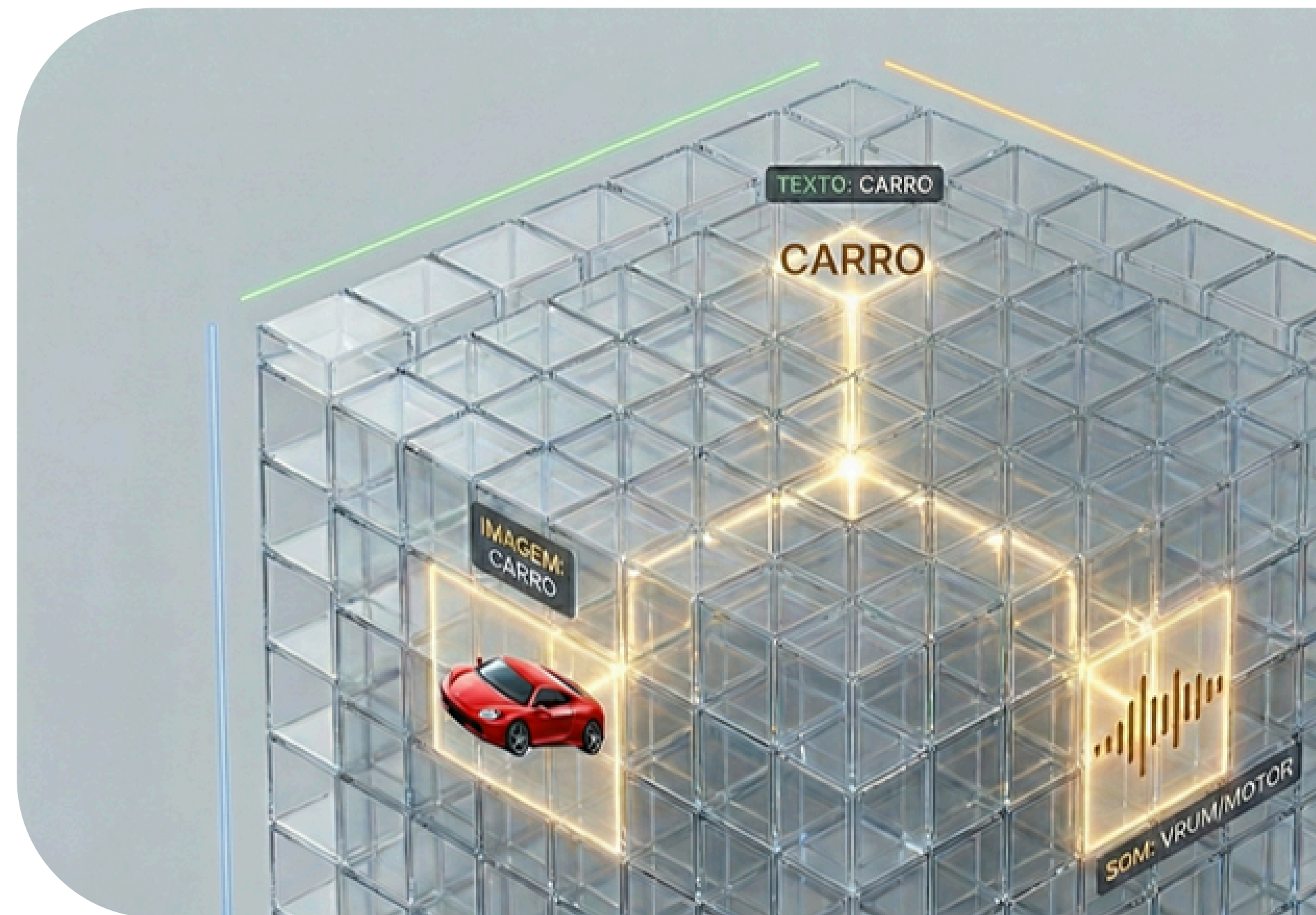


Execução:



Introdução

- Sistemas inteligentes tradicionais
 - linguagem ou visão
 - sem combinar os dois
- Avanços atuais (Transformer)
 - possível combinar esses dois domínios.
- VLMs (Vision-language models).



Realização:



Execução:



Aplicações

- Geração de legendas de imagens.
- Responder perguntas sobre imagens
- Visual Question Answering (VQA).
- Reconhecimento de letras e números em imagens (OCR).



Realização:



Execução:



Introdução

- **VLMs combinam dois problemas:**
 - Visão computacional
 - extrair significados de imagens.
 - Processamento de Linguagem Natural (PLN)
 - Compreender e interpretar textos gerados por humanos.
- Sozinhos já são complexos!



Realização:



Execução:



Soluções pré deep learning (- 2015)

• Antes do Deep Learning

- Trabalhos tentavam realizar tarefas
 - como criar legendas automáticas.
- Utilizavam combinação de
 - Processamento de Linguagem Natural
 - Machine Learning clássico
- Legendas de forma "mecânica".

Realização:



Execução:



Baby Talk: Understanding and Generating Image Descriptions (2011)

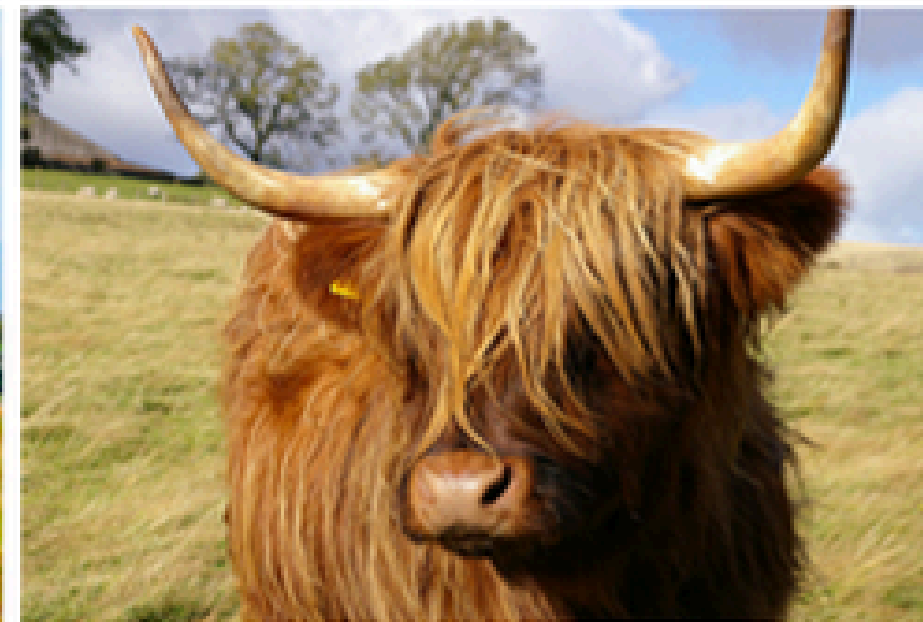
- Geração de frases utilizando um template
- "Um(a) [objeto1] [ação] de um(a) [objeto2]."
 - "Um(a) carro vermelho está ao lado de um(a) carro azul."



This is a photograph of one sky, one road and one bus. The blue sky is above the gray road. The gray road is near the shiny bus. The shiny bus is near the blue sky.



There are two aeroplanes. The first shiny aeroplane is near the second shiny aeroplane.



There are one cow and one sky. The golden cow is by the blue sky.

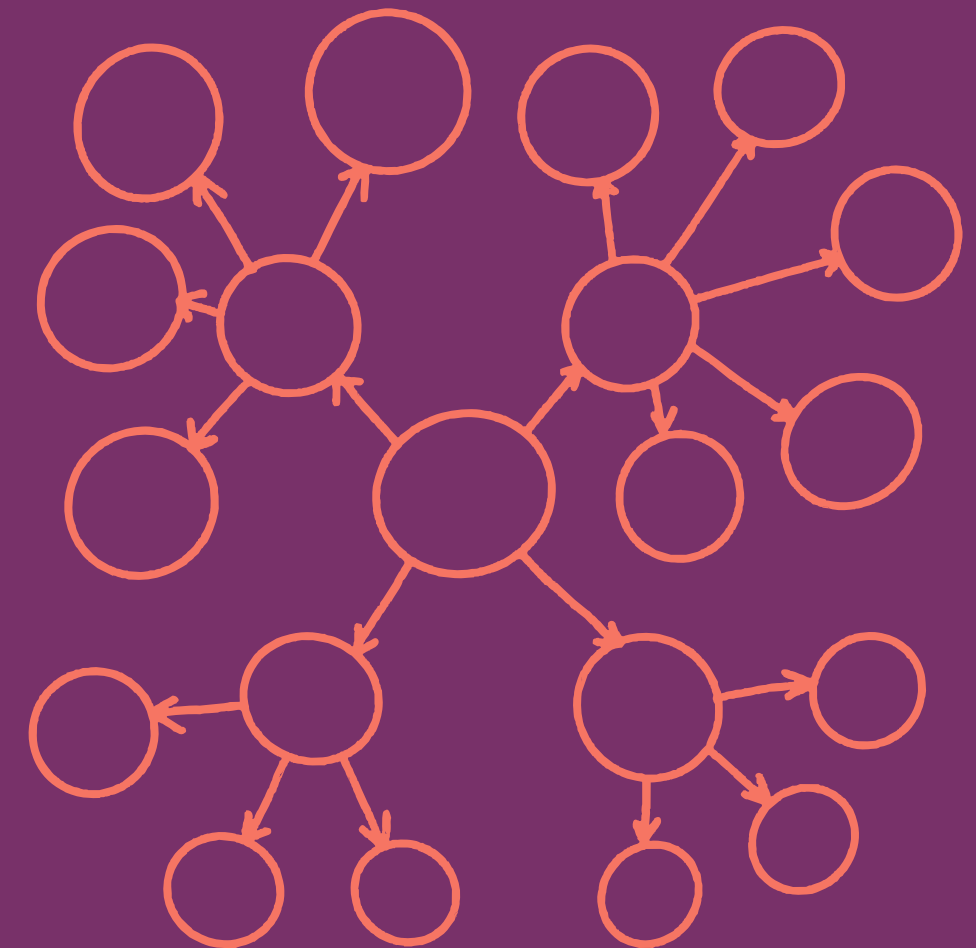
Problemas

- **Limitado**

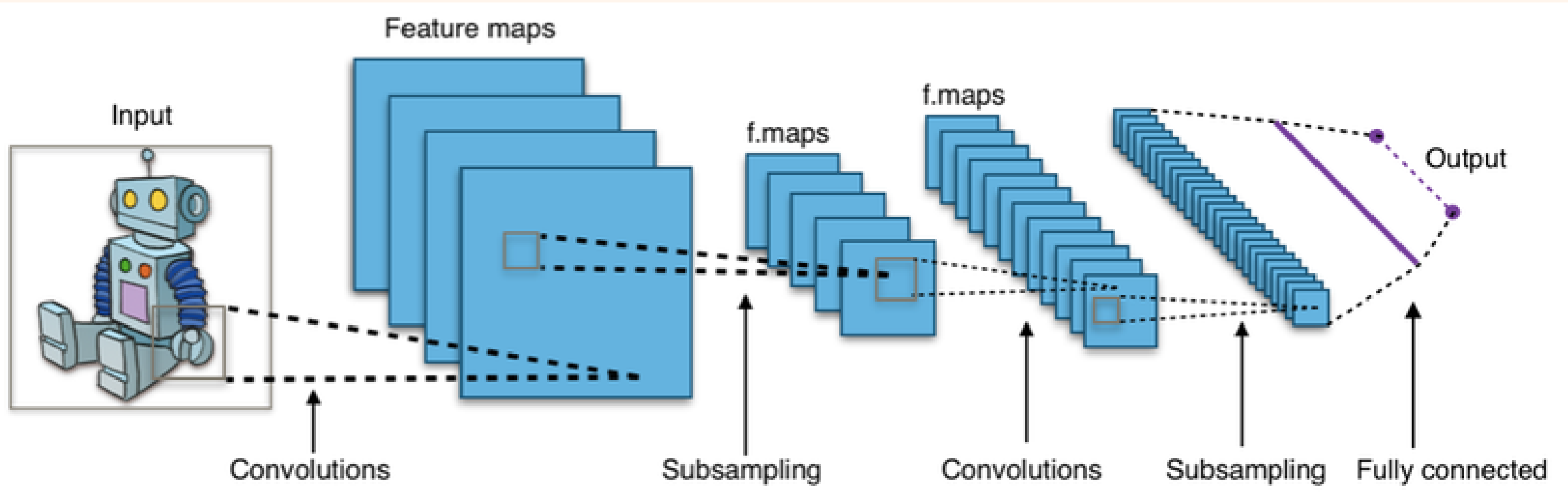
- Frases sempre com o mesmo formato
 - geradas de forma mecânica.

- **Complexo**

- Para gerar uma legenda era necessário:
 - Algoritmo de detecção de objetos em imagem.
 - Algoritmo de Machine Learning para dizer o que era o objeto
 - (lista enorme de objetos possíveis).
 - Achar adjetivos que melhor combinam com o objeto em lista.
 - Criar um grafo com substantivos, verbos adjetivos, etc.



Deep Learning (2015-2019)



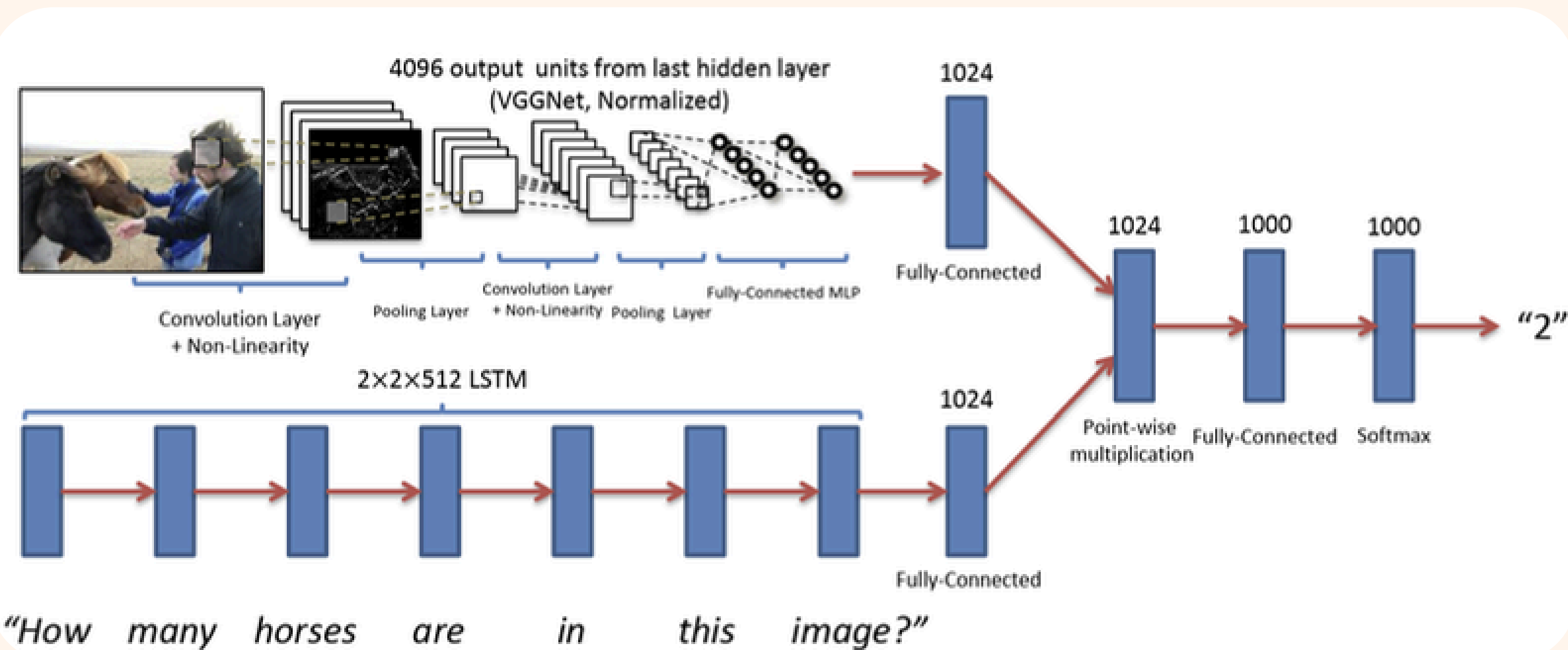
- **Redes Neurais Convolucionais (CNNs)**

- poderosas para imagens - exploram a estrutura espacial (2D) dos dados, capturam padrões locais como bordas, texturas e formas.

- **Redes Neurais Recorrentes (RNNs)**

- eficazes para texto - modelam sequências e processam dados palavra por palavra.

Deep Learning (2015-2019)



- **VQA: Visual Question Answering (2015)**
- Formalizou o desafio de responder perguntas sobre as imagens (VQA).
- Combina uma rede LSTM (RNN) e uma CNN para responder perguntas.

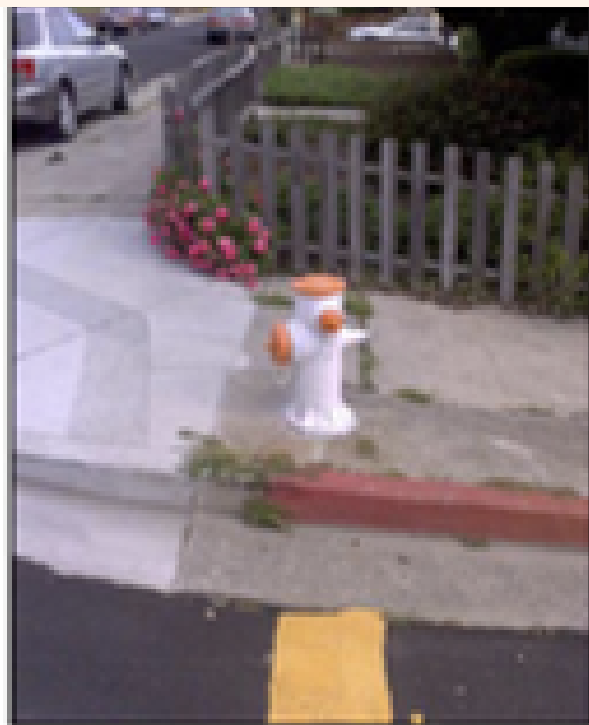
Realização:

Execução:

Deep Learning (2015-2019)

- **VQA: Visual Question Answering (2015)**

- Criou um dataset extenso com imagens, perguntas e respostas.
- Respostas curtas para perguntas simples.
- Início da visão integrada de imagem e texto.



Can you park here?

no
no
no

no
no
yes

What color is the hydrant?

white and orange
white and orange
white and orange

red
red
yellow



What kind of store is this?

bakery
bakery
pastry

art supplies
grocery
grocery

Is the display case as full as it could be?

no
no
no

no
yes
yes



How many bikes are there?

2
2
2

3
4
12

What number is the bus?

48
48
48

4
46
number 6

Deep Learning (2015-2019)

• Problemas

- Dificuldades com frases longas.
- Limitado a fornecer um conjunto de respostas padrão:
 - "2."
 - "sim."
 - "não."
- Apenas respostas curtas - em essência, é similar a um classificador.

Realização:



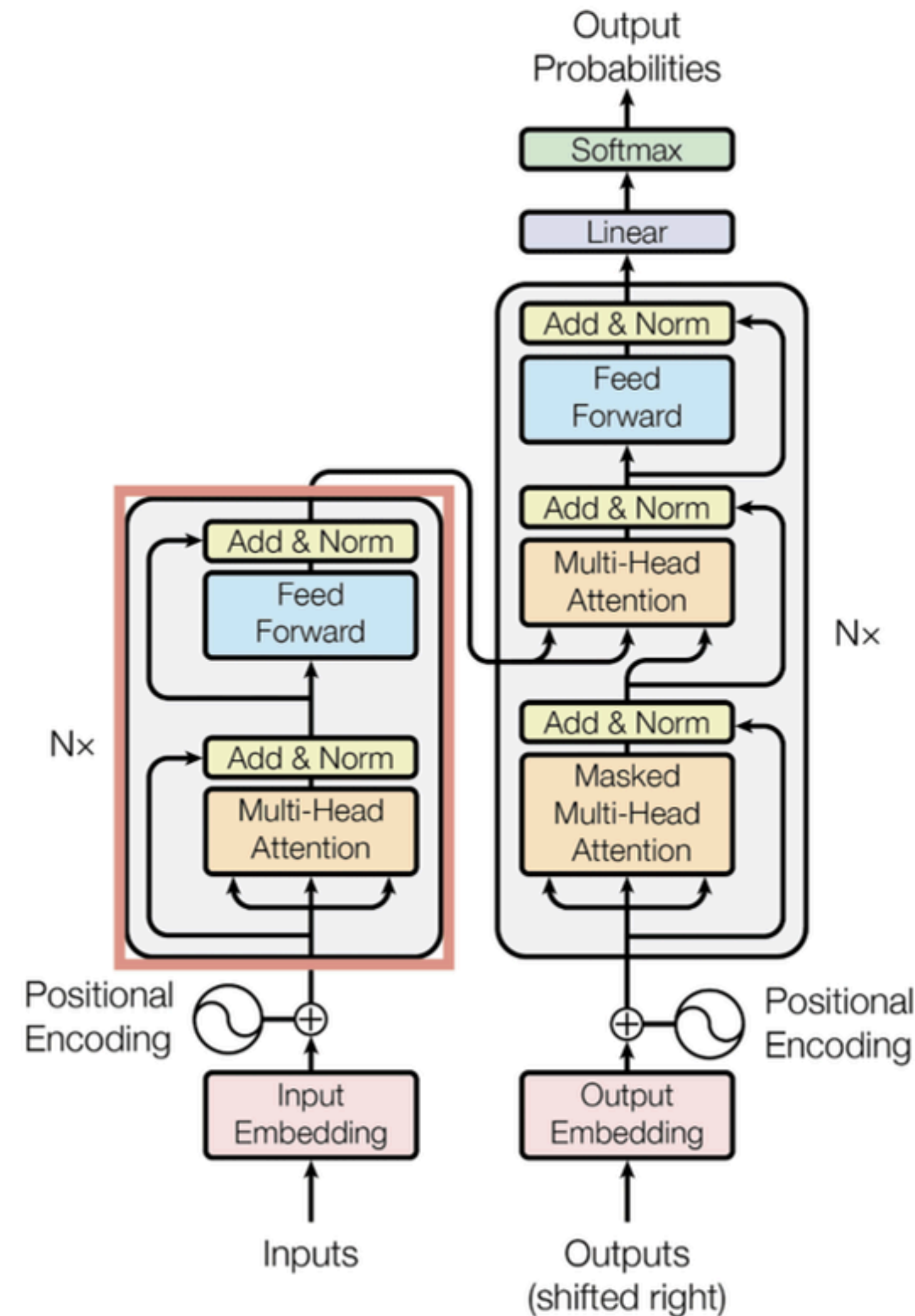
Execução:



Deep Learning (2019-hoje)

• Surgimento da arquitetura Transformer

- Artigo Attention is all you need (2017)
- Aplicação inicial em tradução automática.
- Revolução que possibilitou a criação das LLMs e VLMs.
- Arquiteturas multimodais (imagens, texto e vídeo).



Realização:



Execução:



Deep Learning (2019-hoje)

- **Dados são quebrados em tokens.**

- Textos viram palavras.
- Imagens viram pedaços de imagens (patches).
- Cria uma matriz dizendo quanto um token deve "prestar atenção" a outro.
- Os modelos aprendem a relação entre:
 - palavras e palavras (LLM).
 - patches e patches (Visual Transformer).
- (em VLMs) palavras e imagens (VLM).

Realização:



Execução:



Deep Learning (2019-hoje)

- Regiões da imagem semanticamente relacionadas (ex.: cabeça e corpo do cachorro)
 - recebem maiores pesos de atenção entre si.
- Regiões não relacionadas (ex.: cachorro e grama)
 - apresentam menores pesos de atenção.
- Usamos a mesma lógica para texto!



Realização:

Execução:

Deep Learning (2019-hoje)

• Vantagens

- Compreensão global do contexto da frase / imagem.
- Não sofre com sequências longas como RNNs.
- O modelo aprende onde prestar atenção dependendo da entrada.
- Não usa regras fixas (flexibilidade).



Realização:

Execução:

Deep Learning (2019-hoje)



- Uso intensivo de memória e GPU.
- Interpretabilidade.
- Não possui uma capacidade de raciocínio real.



eu preciso lavar meu carro e o lava-rápido está a 200 m de distância. Devo ir de carro ou andando?

Se está a só 200 metros, ir andando quase sempre é a melhor escolha.

Em termos práticos:

- **Tempo:** você leva uns 2–3 minutos a pé. De carro, entre entrar, manobrar e parar, pode dar praticamente o mesmo ou até mais.
- **Comodidade:** evita ligar o carro à toa e procurar espaço no lava-rápido.
- **Economia:** menos gasto de combustível (mesmo que pequeno).
- **Simplicidade:** menos stress — só vai, deixa o carro e pronto.

Realização:



Execução:



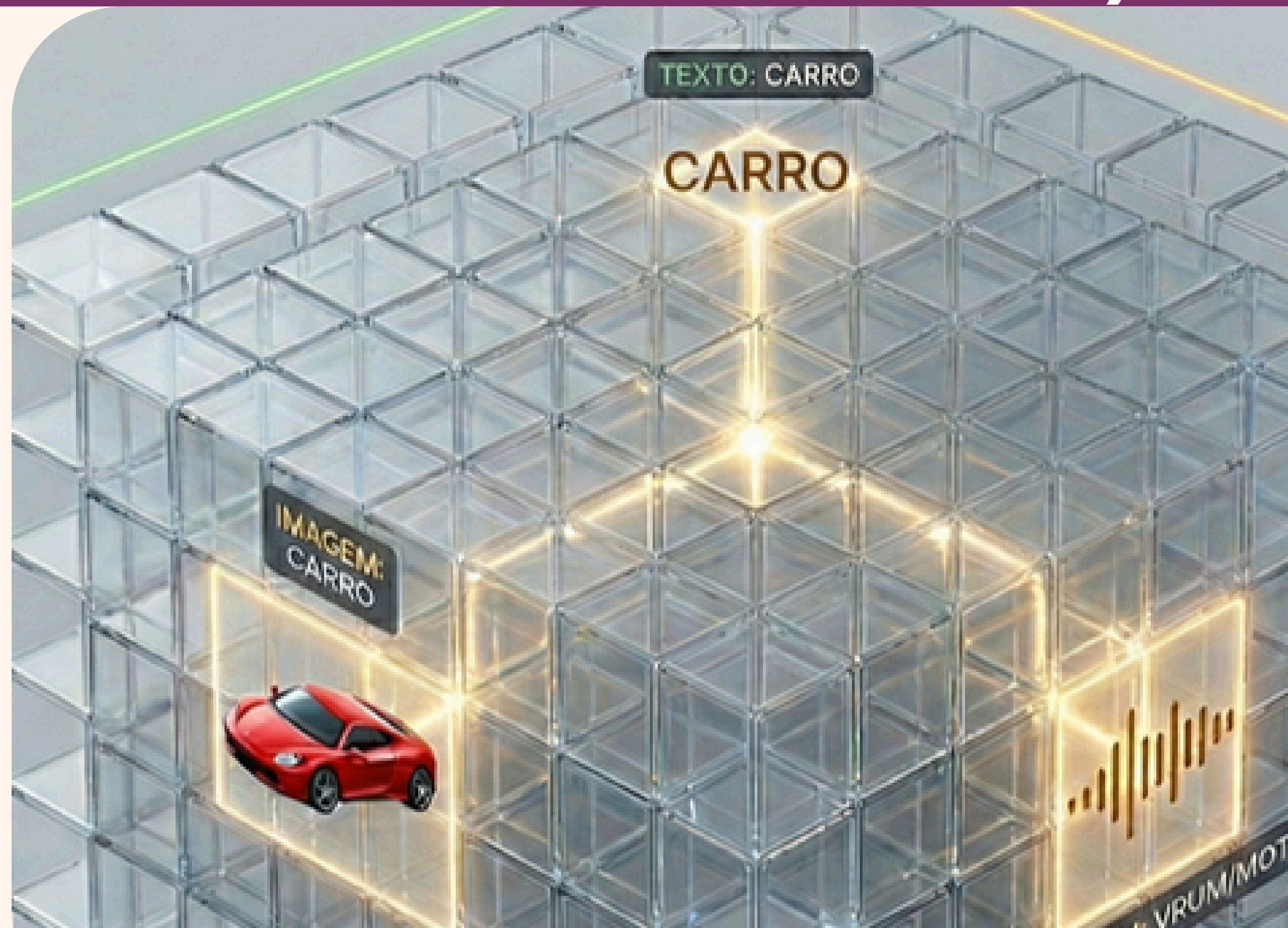
CLIP (CONTRASTIVE LANGUAGE-IMAGE PRE-RAINING)

- Apesar dessas limitações, a capacidade de aprender representações gerais permitiu o surgimento de modelos multimodais.
- Modelos que trabalham com mais de um tipo de dado.
- Um dos mais importantes é o CLIP.

Realização:



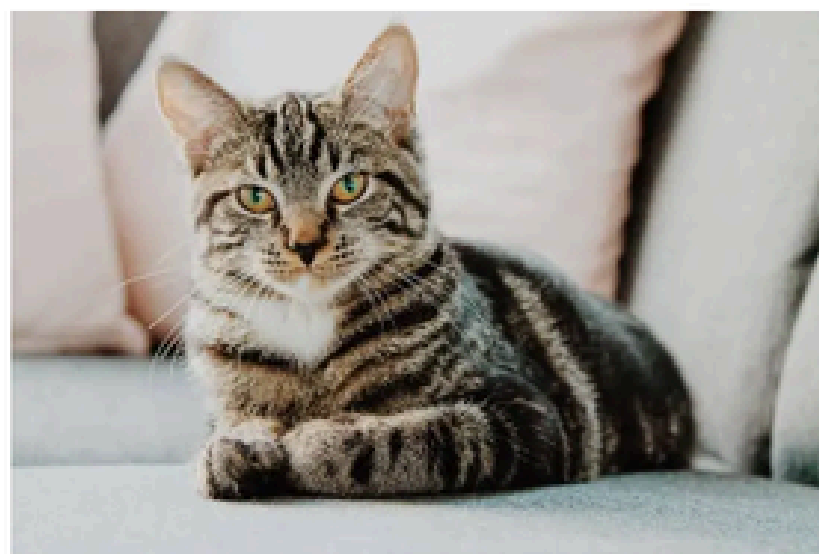
Execução:



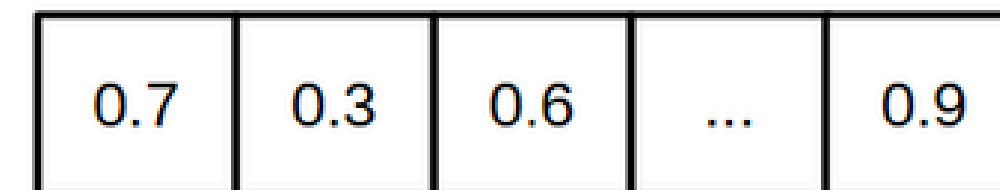
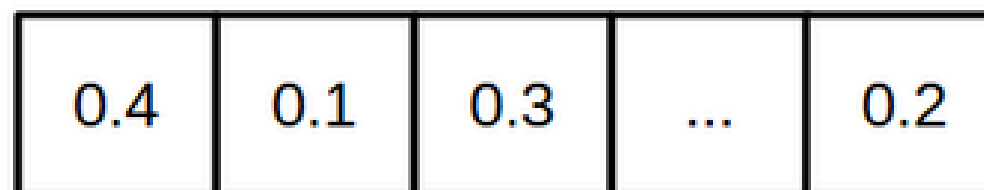
CLIP (CONTRASTIVE LANGUAGE-IMAGE PRE-TRAINING)

- **Em VLMs**

- imagens e textos são convertidos em embeddings.
- Um embedding é
 - uma representação vetorial numérica de tamanho fixo.
- O vetor tem tamanho fixo
 - então é possível "medir a distância" entre imagens e texto.



Um gato deitado no sofá olhando para frente.

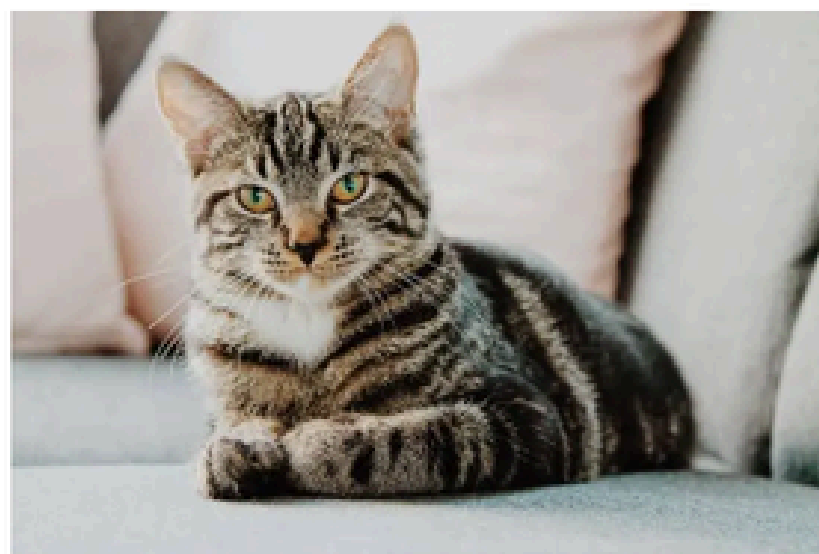


512 números em cada um

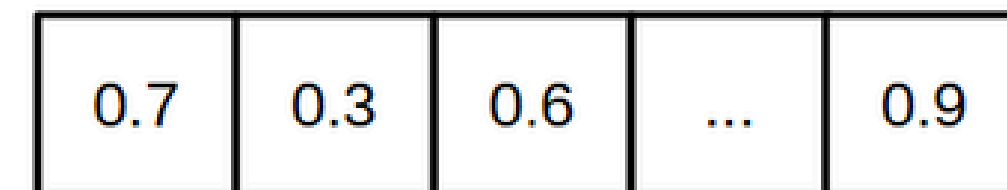
CLIP (CONTRASTIVE LANGUAGE-IMAGE PRE-TRAINING)

- **Em VLMs**

- imagens e textos são convertidos em embeddings.
- Um embedding é
 - uma representação vetorial numérica de tamanho fixo.
- O vetor tem tamanho fixo
 - então é possível "medir a distância" entre imagens e texto.



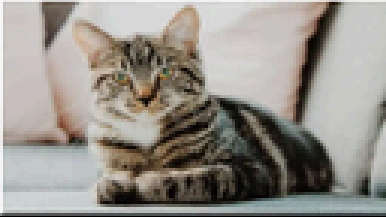



Um gato deitado no sofá olhando para frente.



512 números em cada um

CLIP (CONTRASTIVE LANGUAGE-IMAGE PRE-RAINING)

- O CLIP é um modelo treinado para
 - associar imagens a suas legendas.
- Ele busca
 - diminuir a distância entre pares corretos de imagens e legendas.

				
Isso é um gato	■			
Um cavalo		■		
Imagem de um cachorro			■	
Foto de uma girafa				■

Realização:

Execução:

CLIP (CONTRASTIVE LANGUAGE-IMAGE PRE-TRAINING)

• Aplicações

- Busca de imagens

Query by

text

image

Enter text/image URL here:

two cats

Submit

Image Search Engine



Realização:

Execução:

CLIP (CONTRASTIVE LANGUAGE-IMAGE PRE-TRAINING)

- O CLIP é parte de outros modelos aplicados em Visual Question Answering (VQA)
 - Exemplo:
 - LLaVA



Realização:

Execução:

Llava



- Modelo baseado no CLIP
- utilizado para Visualquestion answering (VQA)
- Combina o CLIP com o modelo de linguagem LLaMA.
- CLIP compreende as representações de imagens
 - Enquanto o LLaMA gera o texto descrevendo a imagem.

- Outros modelos similares:
 - Qwen-VL;
 - DeepMind Flamingo.



Realização:



Execução:





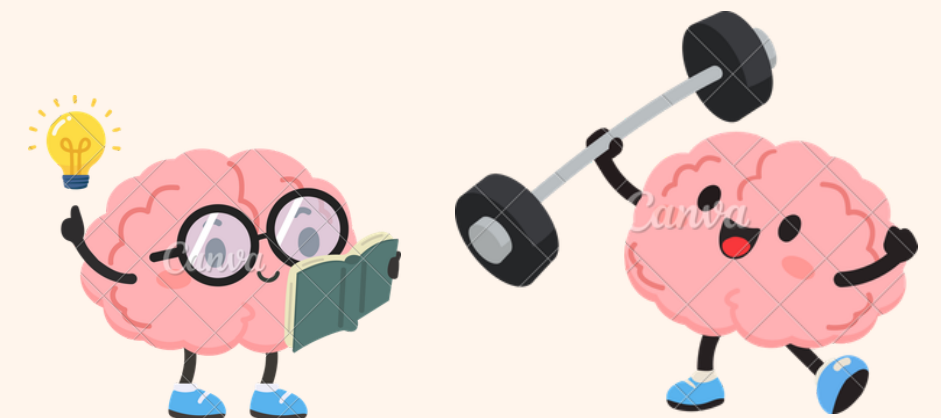
Llava

- **2015**
 - CNN + RNN.
 - Respostas fixas
 - (sim, não, 2, 3, etc.)
- **2023**
 - Melhorias significativas no processo de responder perguntas sobre imagens (VQA)
 - LLM + modelo de visão.
 - Geração de textos longos explicando a imagem.

Realização:



Execução:



Visual input example, Chicken Nugget Map:

Llava

Sometimes I just look at pictures of the earth from space and I marvel at how beautiful it all is.



User

Can you explain this meme in detail?

LLaVA

The meme in the image is a creative and humorous take on food, with a focus on chicken nuggets, as the center of the universe. The meme begins with the phrase "Sometimes I just look at pictures of the Earth from space and I marvel at how beautiful it all is..." The main image features a tray of cooked chicken nuggets, with the chicken pieces positioned to resemble the continents and islands on a world map. The meme playfully suggests that the chicken nuggets represent the Earth, and the various locations depicted in the photo are actually chicken nugget versions of different places. The punchline of the meme appears in smaller text at the bottom, reading "I mean, it's not the real Earth, but how beautiful it

Aplicações VLMs

- **LLaVA-Med:** Training a Large Language-and-Vision Assistant for Biomedicine in One Day (2023)
- O LLaVA foi refinado utilizando 660.000 pares de imagens / texto.
- Treinado para responder perguntas em imagens médicas:
- Em 8 GPUs A100 de 40 GB cada.
 - Parte do treinamento realizado utilizando legendas sintéticas geradas pelo GPT-4.



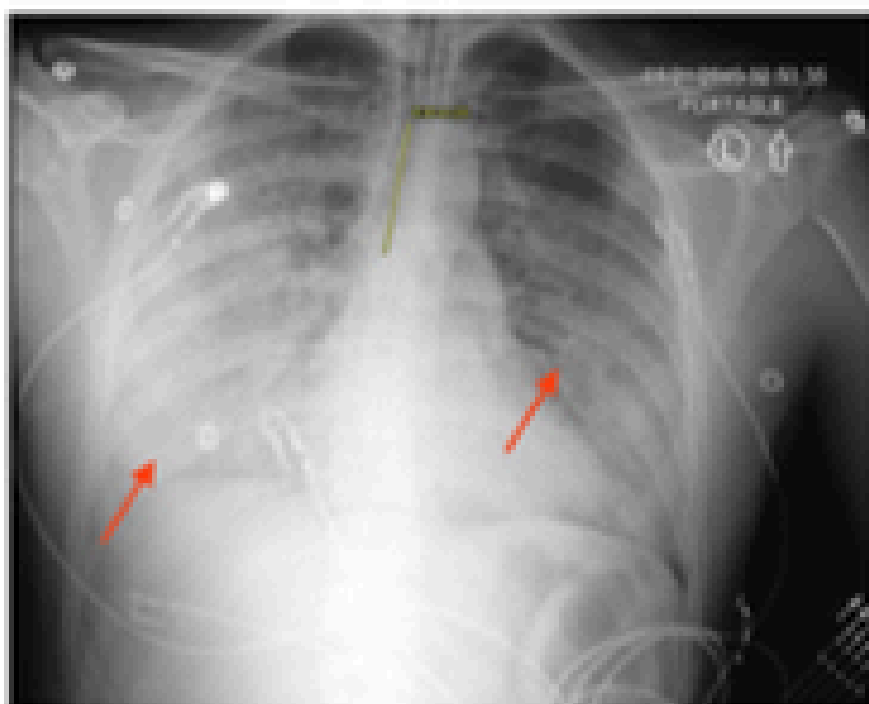
Realização:



Execução:



Visual input example, Biomedical image from [6]



Caption: Chest x-ray (CXR) on Day 2 of admission post-intubation (yellow line showing the level of the endotracheal tube). Rapidly worsening ground-glass opacities are seen throughout the lungs with relative subpleural sparing (red arrows)

Source: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7398722/>

User

What is shown in this image?

LLaVA-Med

The image is a chest X-ray (CXR) that shows bilateral patchy infiltrates, which are areas of increased opacity in the lungs. These infiltrates can be indicative of various lung conditions, such as infections, inflammation, or other lung diseases.

User

Are there any devices implanted in this patient?

LLaVA-Med

Yes, it appears that the patient has several devices implanted. These devices may include endotracheal tubes, central venous catheters, and/or pulmonary artery catheters. These devices are typically used for monitoring and managing various medical conditions, as well as for providing support during surgical procedures.

O modelo foi capaz responder perguntas em mandarim, apesar de treinado em inglês.



Biomedical image from the SLAKE Bilingual dataset. [23]



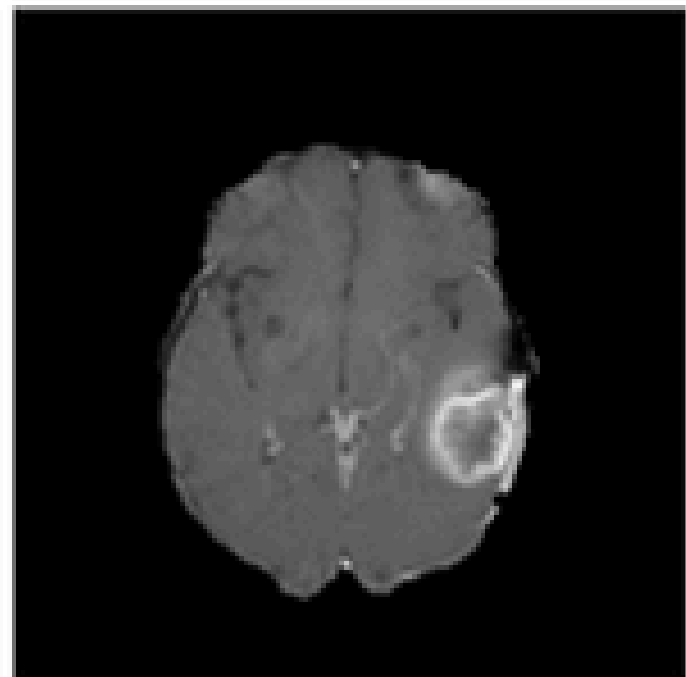
User
LLaVA-Med

这张图片的成像方式是什么？

The image is an abdominal computed tomography (CT) scan, which is a cross-sectional view of the patient's abdomen. The image is taken in the portal phase, which means it is captured after the administration of contrast material to enhance the visibility of certain structures within the abdomen.

Answer

CT



User
LLaVA

这张图片展示的是核磁共振的哪种类型？

The image shows a typical appearance of a nodular lesion on T1-weighted MR images. The lesion appears hyperintense, which means it is brighter than the surrounding tissue. This can be indicative of various types of lesions, and further evaluation would be needed to determine the specific nature and cause of the lesion.

Answer

T1

Outras Aplicações

- **Outras Aplicações**

- Descrição de imagens em E-Commerce.
- Acessibilidade (descrição imagem para texto).
- Descrição de gráficos.
 - Compreensão de documentos (OCR).

— LLaVA Applications —

Visual Question Answering



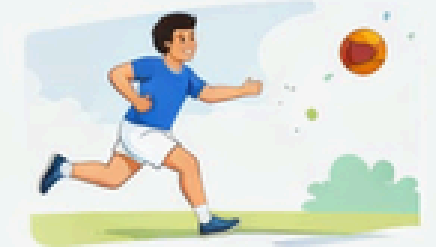
Q Where is the cat sleeping?
A. In a pet bed.

Image Captioning



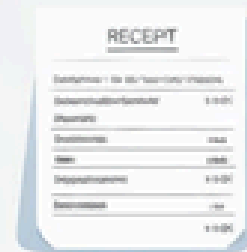
A Golden Retriever resting on a gray couch.

Multimodal Reasoning



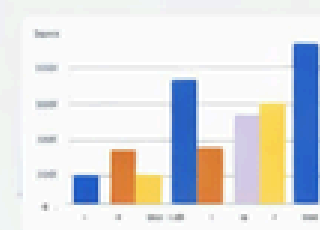
Q Why is the boy running?
A. He is chasing after the ball.

Document Understanding



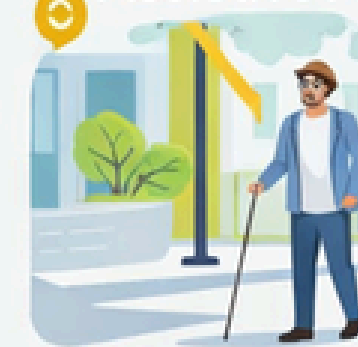
Summarize the information in the receipt.

Chart Interpretation



Q What trend is shown in this graph?

Assistive AI



Q Describe what is in front of me.

E-Commerce & Visual Search



Q Describe this item.

Realização:

Execução:

Conclusão

- Métodos iniciais eram limitados e baseados em regras.
- Deep Learning trouxe melhorias, mas ainda com respostas restritas.
- Transformers e VLMs permitiram integração multimodal real de geração de linguagem



Realização:



Execução:



Desafios

- Necessidade de grande quantidade de dados.
- Baixa interpretabilidade.
- Dificuldade de aplicação em sequências temporais e vídeos longos.
- Alto custo de memória e GPU.

Realização:



Execução:

