Învățare automată în vedere artificială

Curs 7

De la **Transformers** la **Multimodal DL**

Overview

1. Transformers
2. CLIP
3. Multimodal Deep Learning
    - CogVLM

I. Transformers

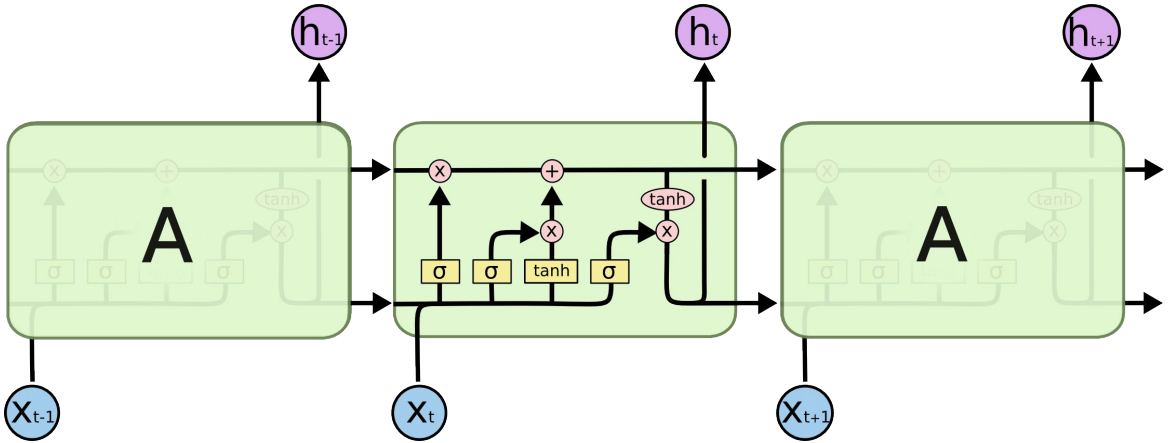
Transformers - Context

- Au aparut initial in NLP
  - Secvențe de cuvinte
    - Ordinea conteaza
    - Număr variabil de cuvinte
    - Sensul unui cuvânt este dependent de restul

Transformers - Motivatie

Pana in 2017, progresul in NLP a fost mai incet.

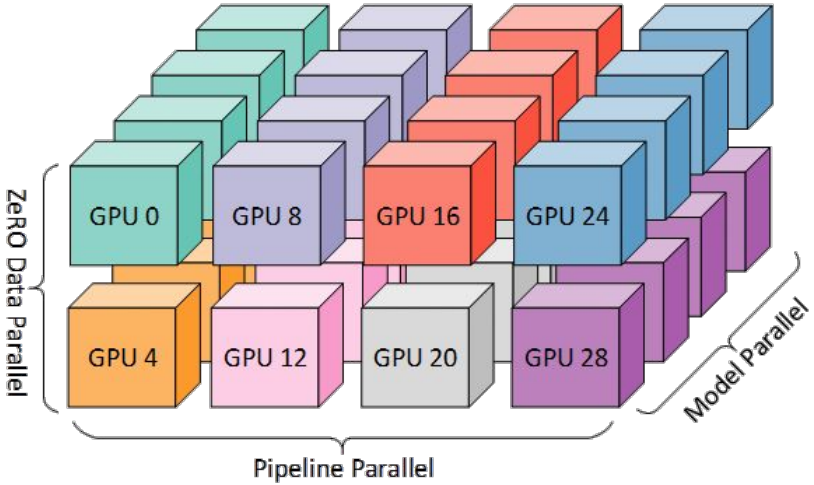
Procesare secvențială.



Sursa: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Transformers - Motivatie

- Depasirea problemelor RNN
  - Optimizare dificila
  - Procesare lenta



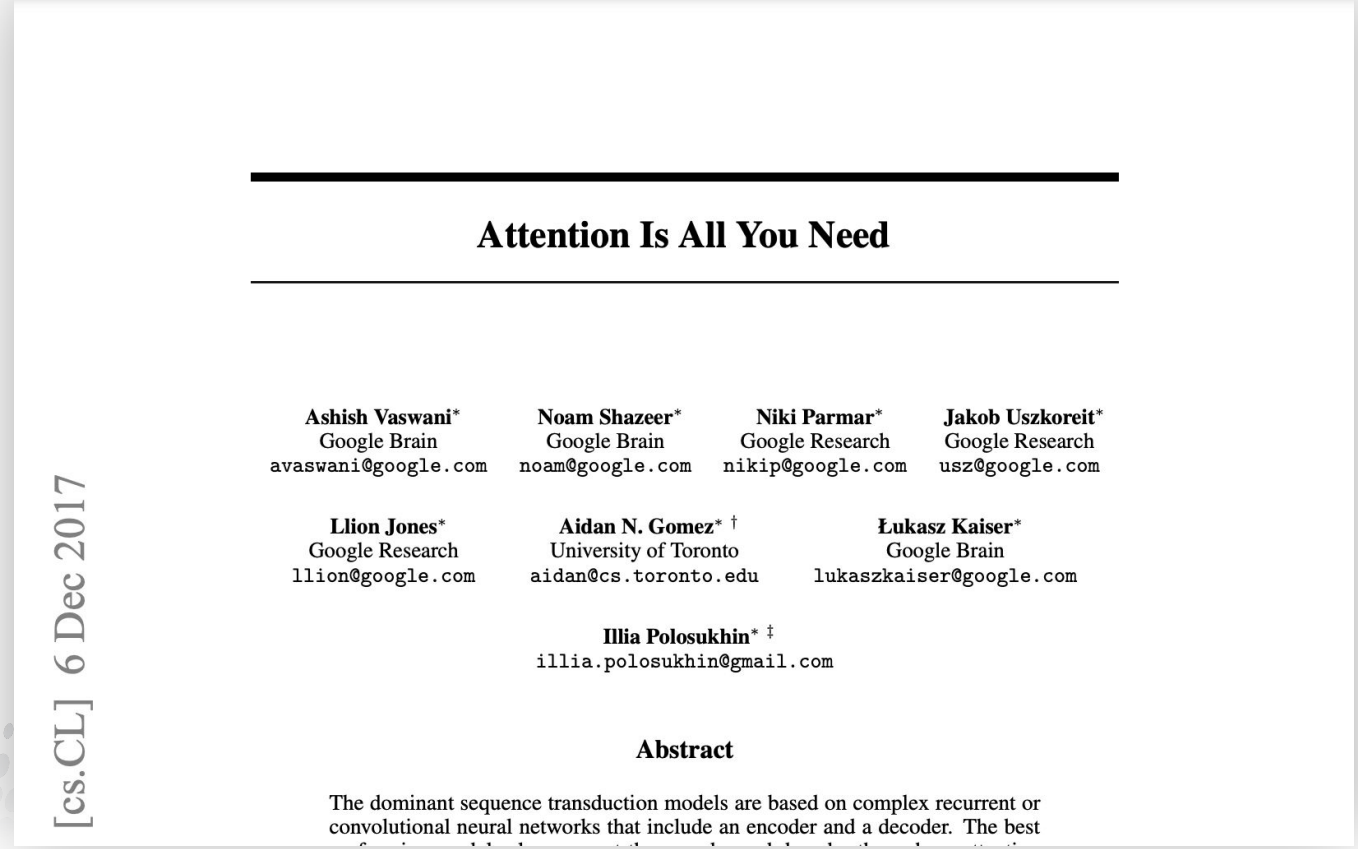
Sursa: https://huggingface.co/docs/transformers/perf_train_gpu_many

Attention

He went to the bank and learned of his empty account, after which he went to a river bank and cried.

Sursa: Google Images

Transformers



[cs.CL] 6 Dec 2017

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

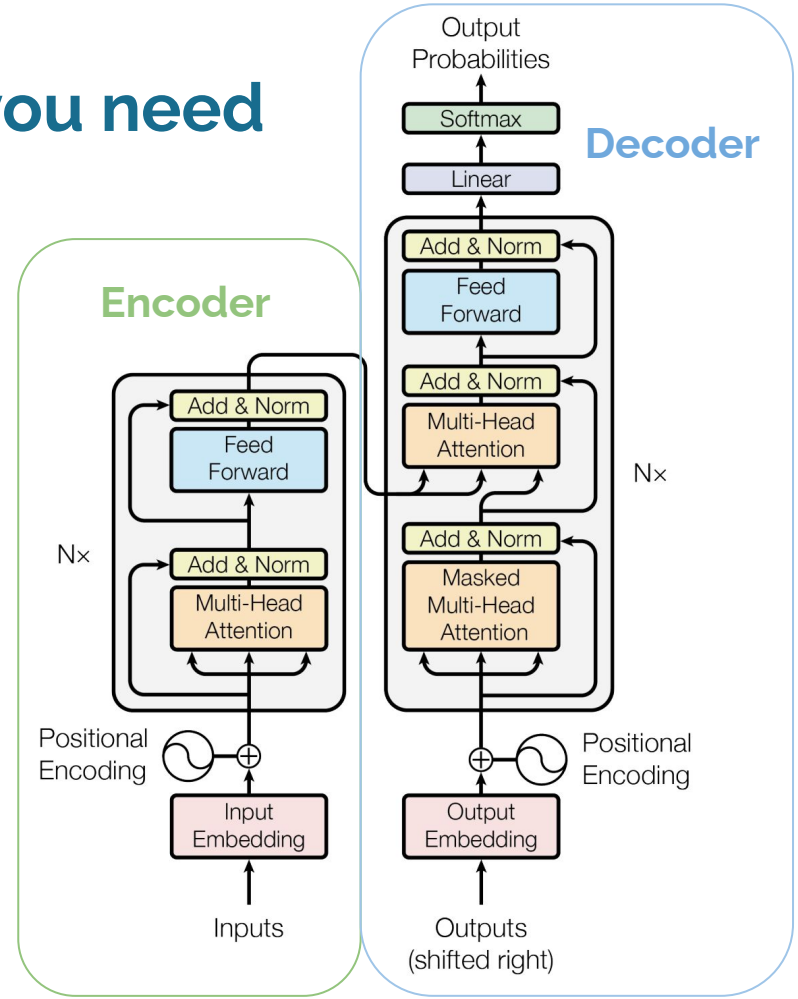
Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best

Attention is all you need

Proceseaza input-ul

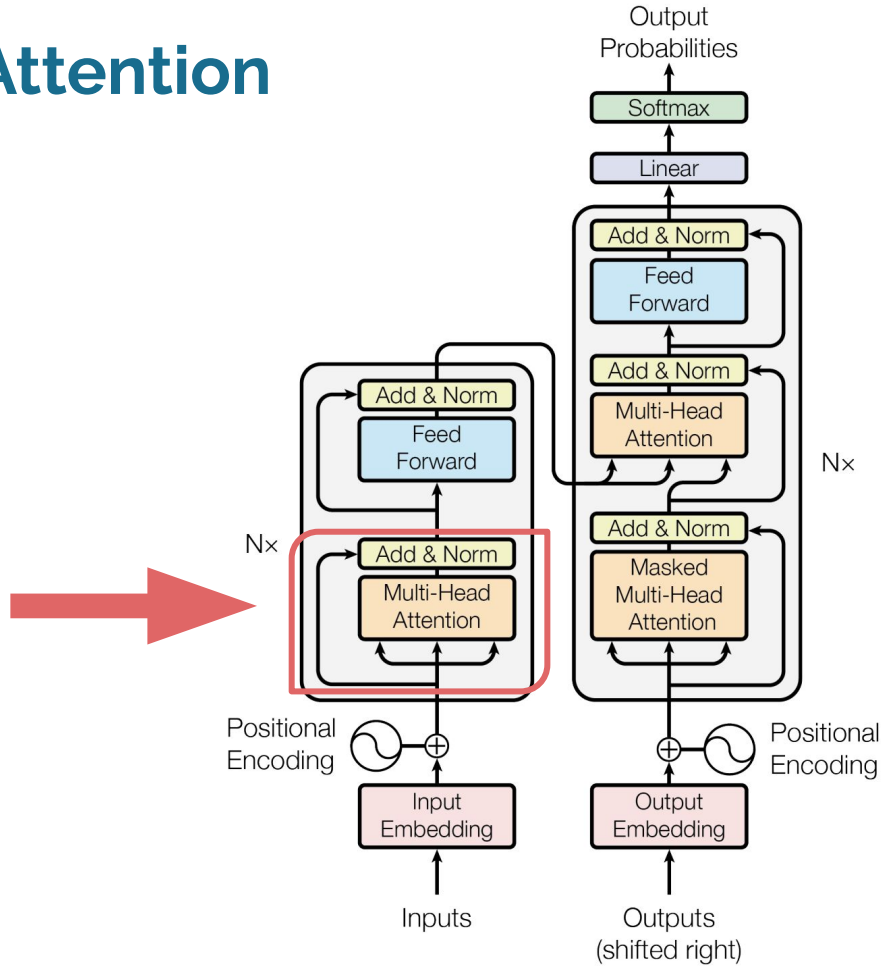
- Embedding In
- Embedding Out



Genereaza secvențial

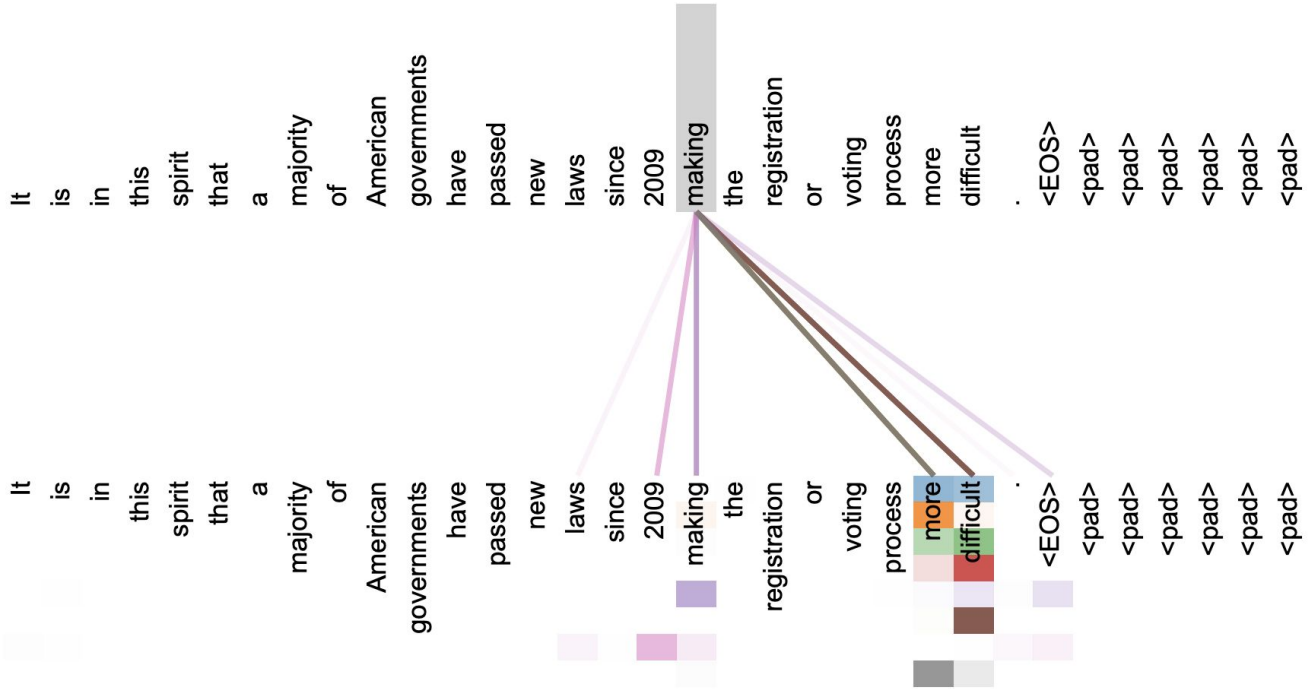
Sursa: Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

Transformers - Self-Attention



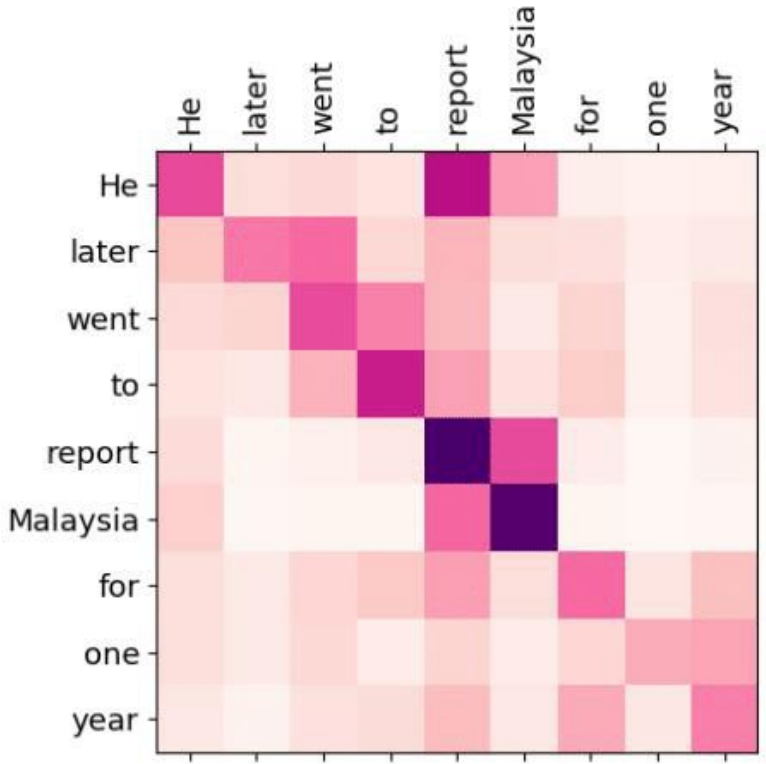
Sursa: Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

Transformers - Self-Attention



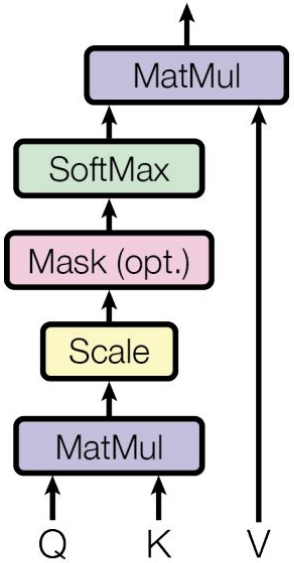
Sursa: Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

Transformers - Self-Attention



Transformers - Self-Attention

Scaled Dot-Product Attention



$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Sursa: Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

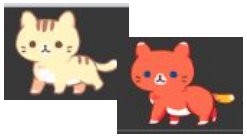
Transformers - Self-Attention

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Query(Q)

Intuitiv: informația de care suntem interesați, pentru care facem un search

- "What are some interesting facts about cats?"



Sursa: Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

Transformers - Self-Attention

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Key(K)

Intuitiv: ajuta la identificare info relevante legate de *query*

- attribute: "feline," "kitten", "purr"

Sursa: Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

Transformers - Self-Attention

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Value(V)

Intuitiv: conținutul care răspunde la întrebare

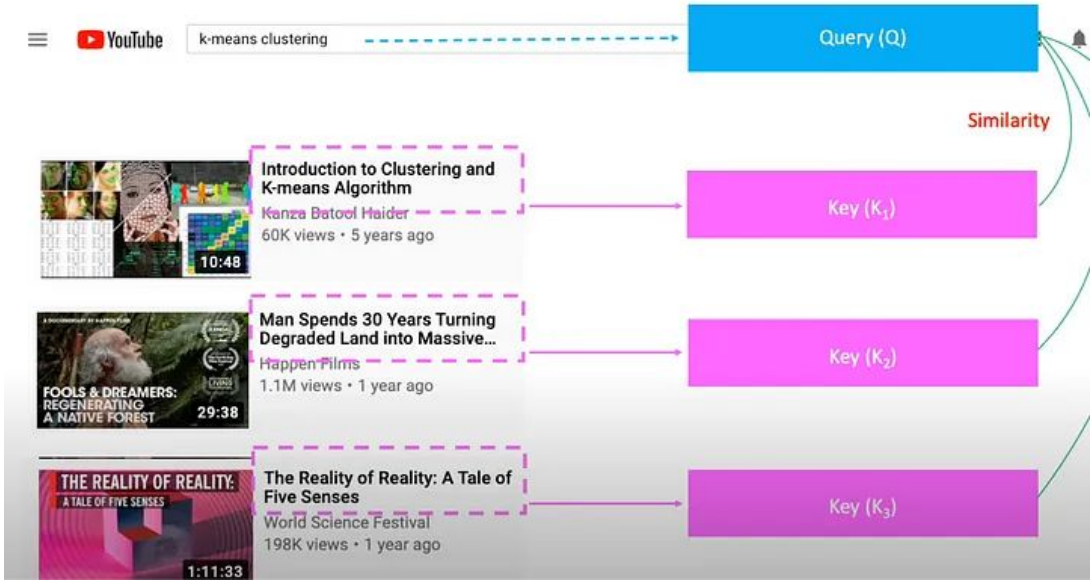
- descrieri legate de pisici, comportament, caracteristici

Sursa: Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

Transformers - Self-Attention

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

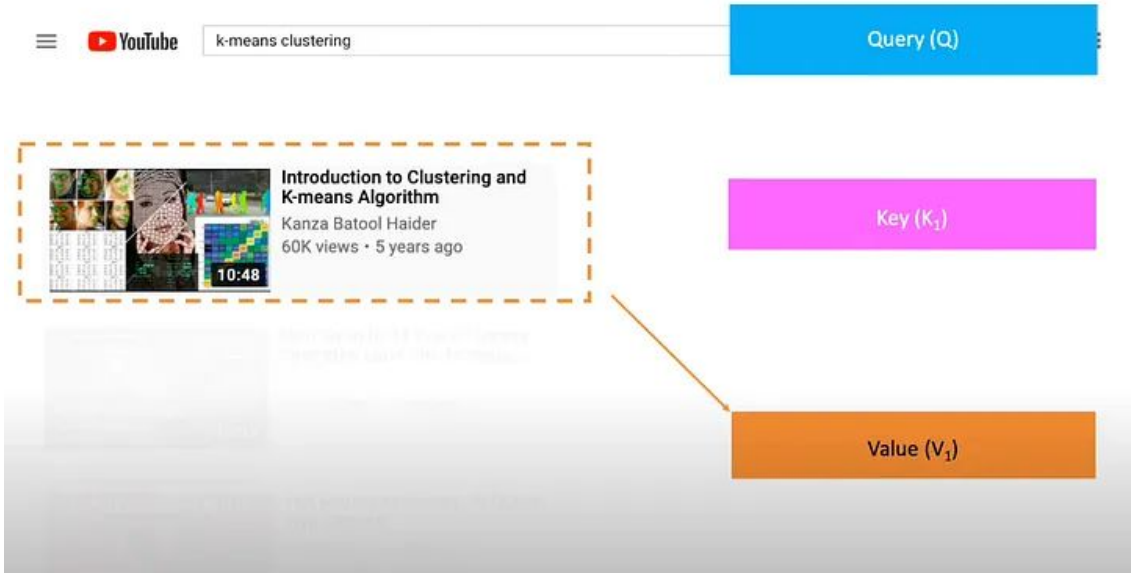
- Similarity



Sursa: medium.com

Transformers - Self-Attention

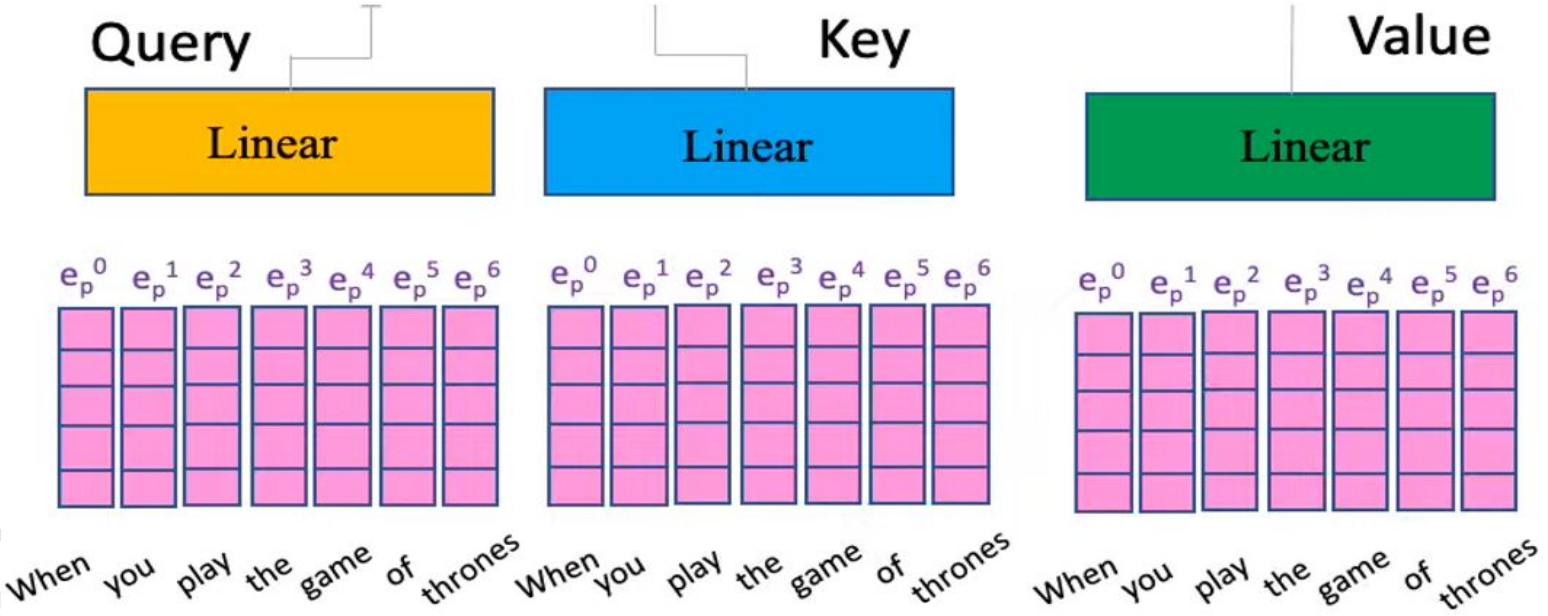
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$



Sursa: medium.com

Transformers - Self-Attention

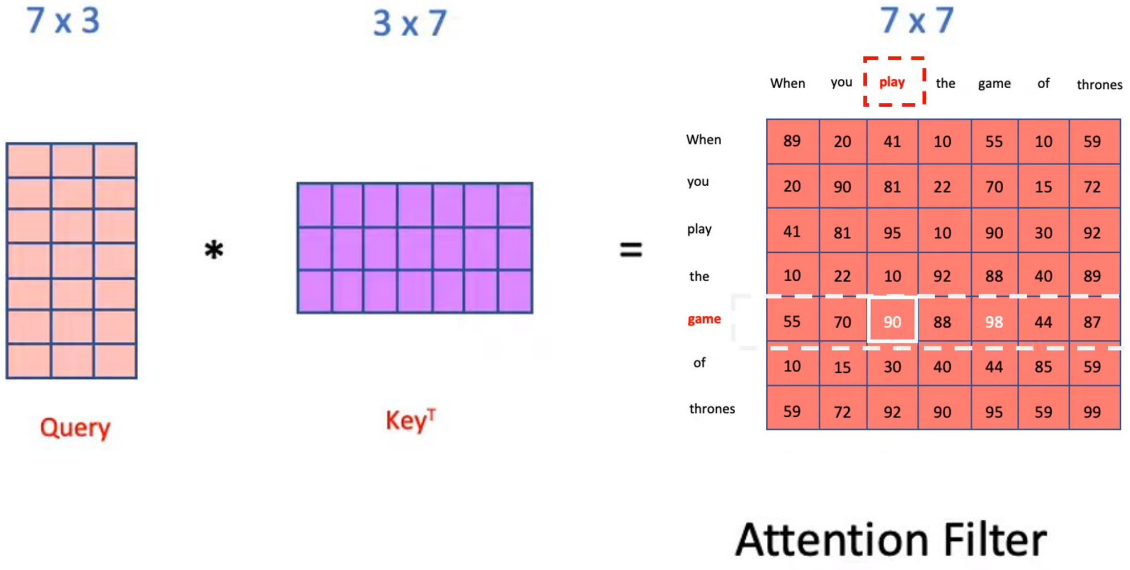
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$



Sursa: medium.com

Transformers - Self-Attention

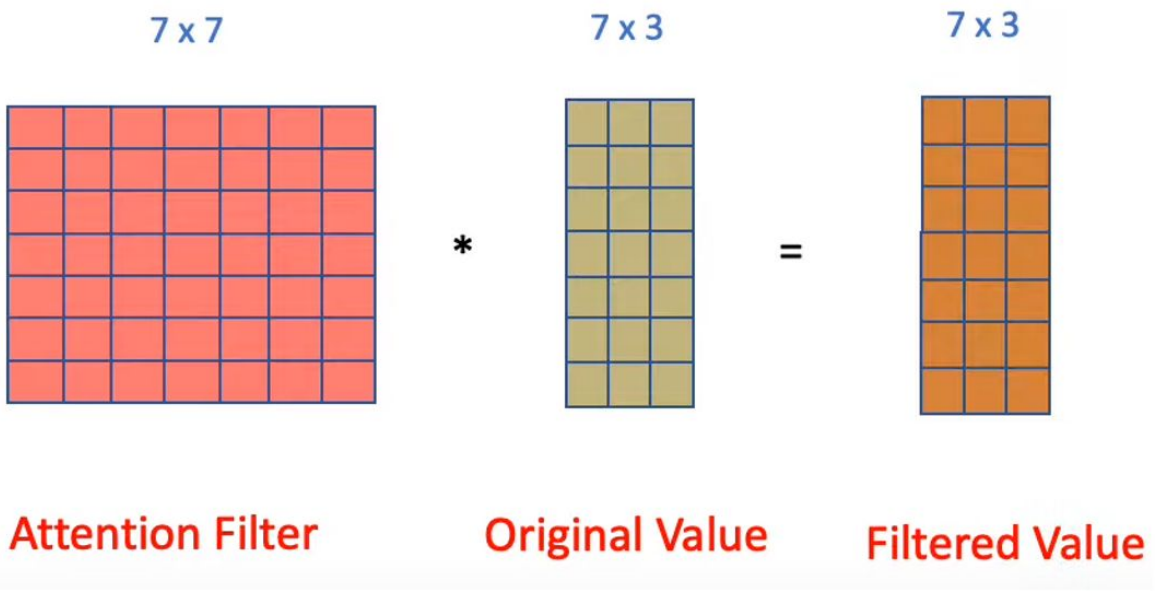
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$



Sursa: medium.com

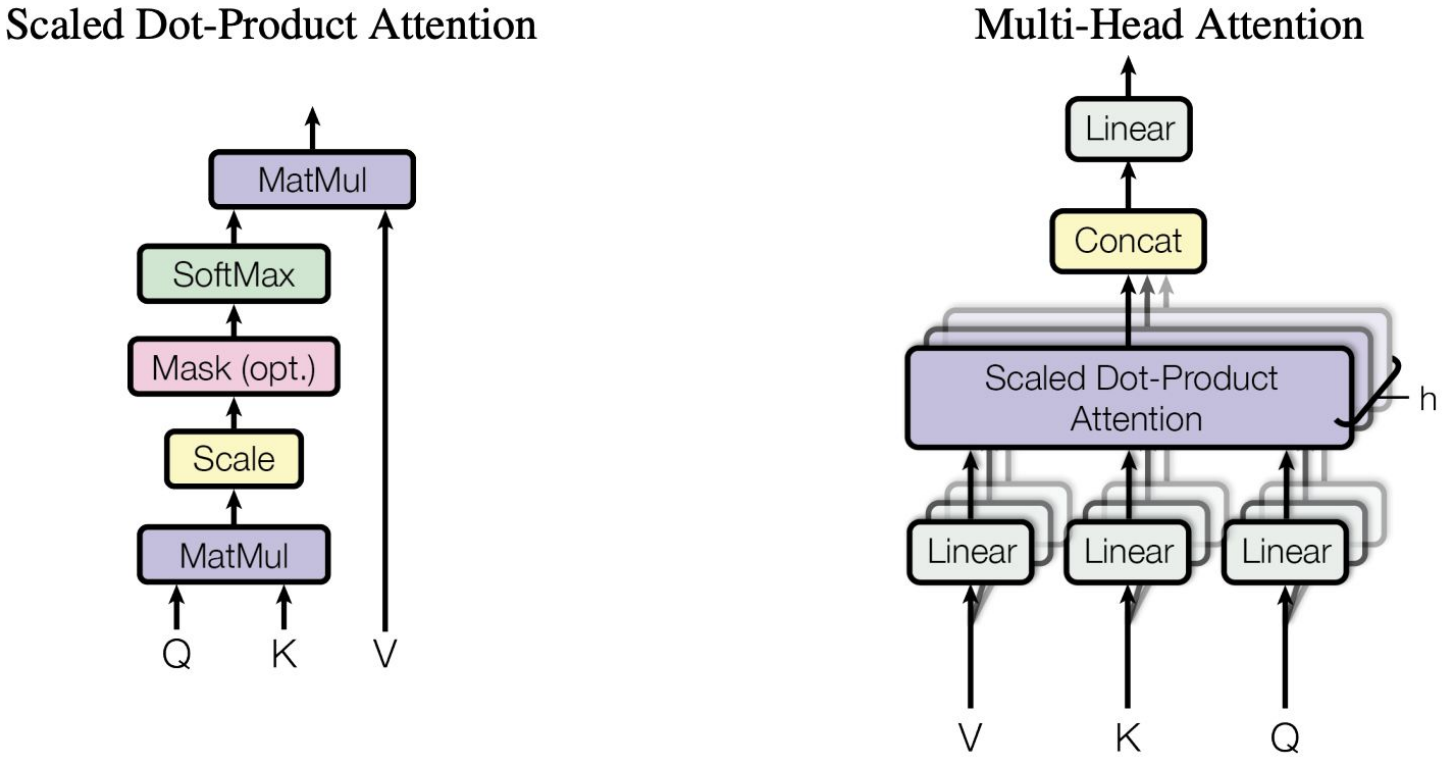
Transformers - Self-Attention

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$



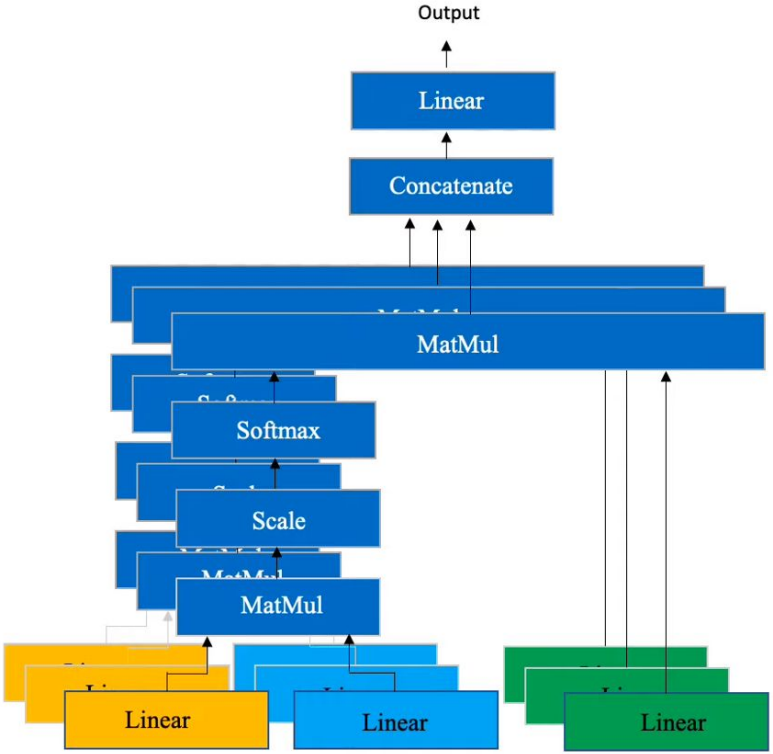
Sursa: medium.com

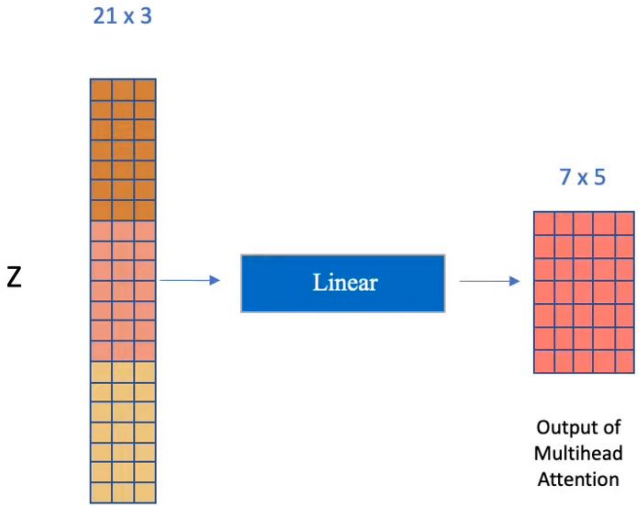
Transformers - Self-Attention



Sursa: Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

Transformers - Self-Attention

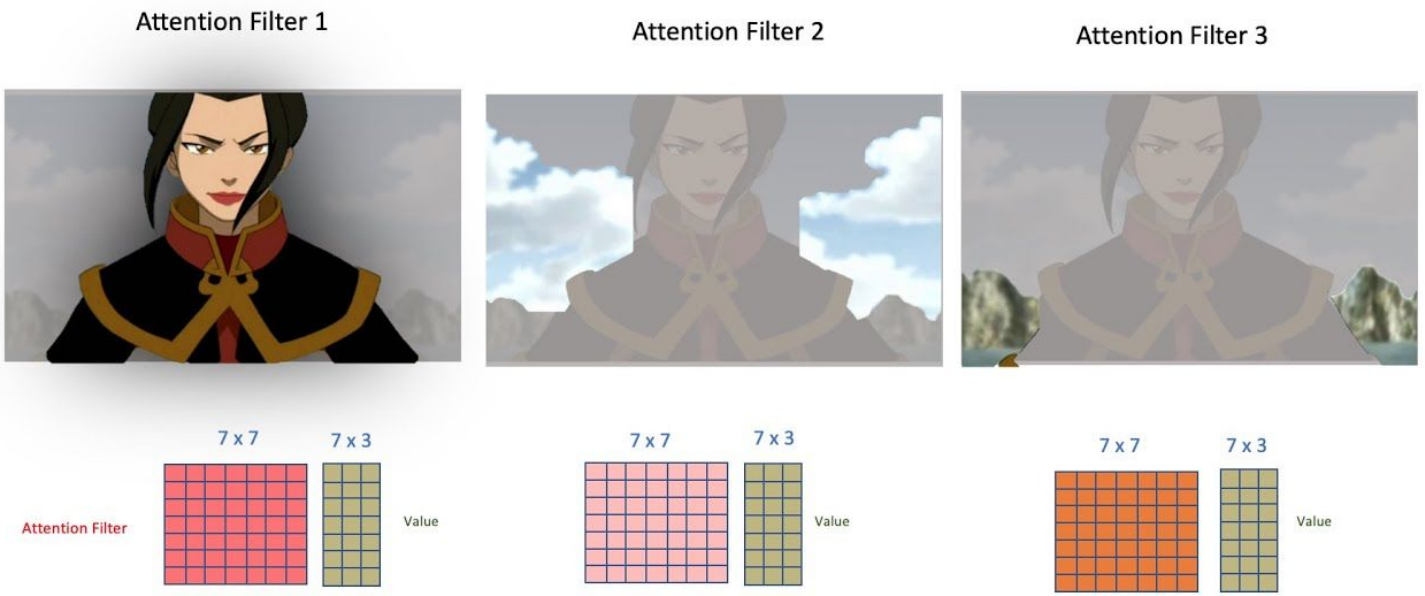




Sursa: medium.com

Transformers - Self-Attention

Intuition for Multi-Head Attention

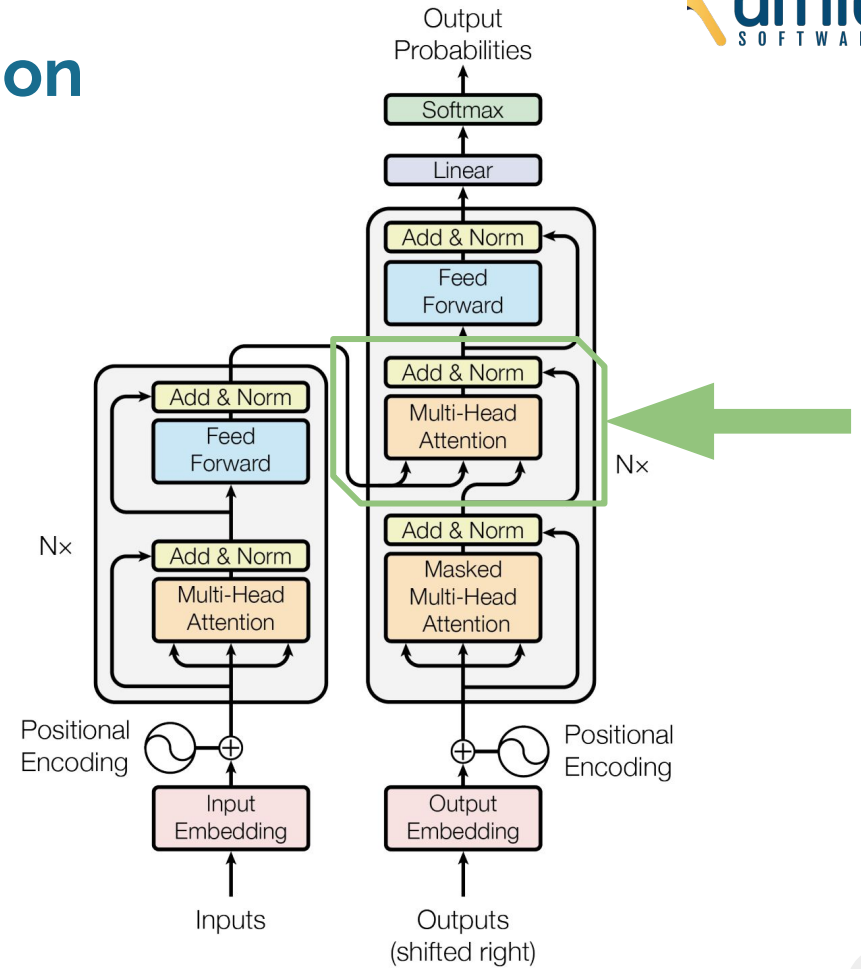


Sursa

Transformers - Cross-Attention

- Mixeaza 2 embedding-uri diferite
- Trebuie să aibă același shape, insa pot fi diferite (imagine, text, sunet)
- Una din secvențe "joacă" rolul de Q cealalta produce K, V

Tasks: Image-text classification, machine translation



Sursa: Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

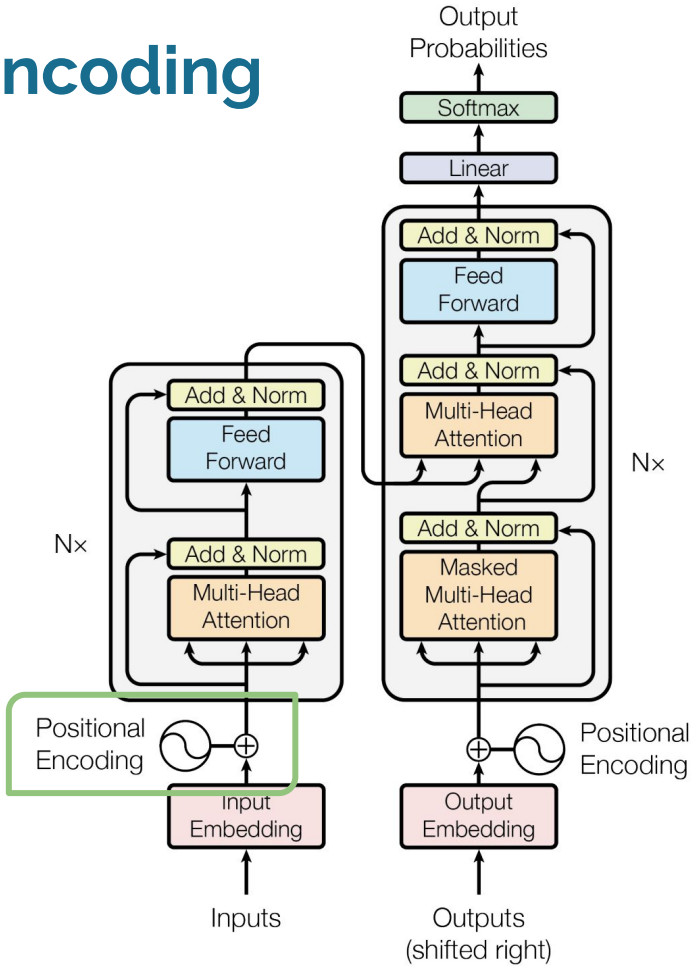
Cross-Attention

Lee, Kuang-Huei, et al. "Stacked cross attention for image-text matching." Proceedings of the European conference on computer vision (ECCV). 2018.

Transformers - Positional Encoding

Ordinea este integrată în reprezentarea cuvintelor

Furnizeaza informații despre poziția token-urilor în secvență

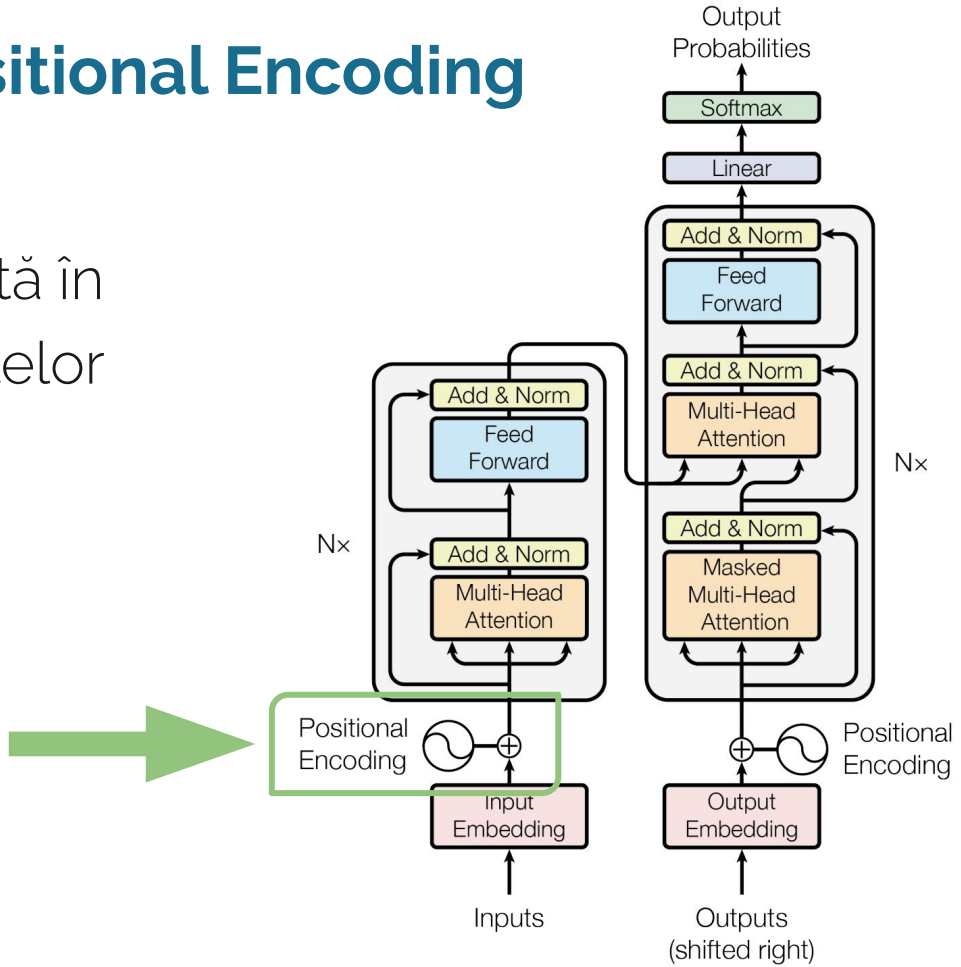


Sursa: Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

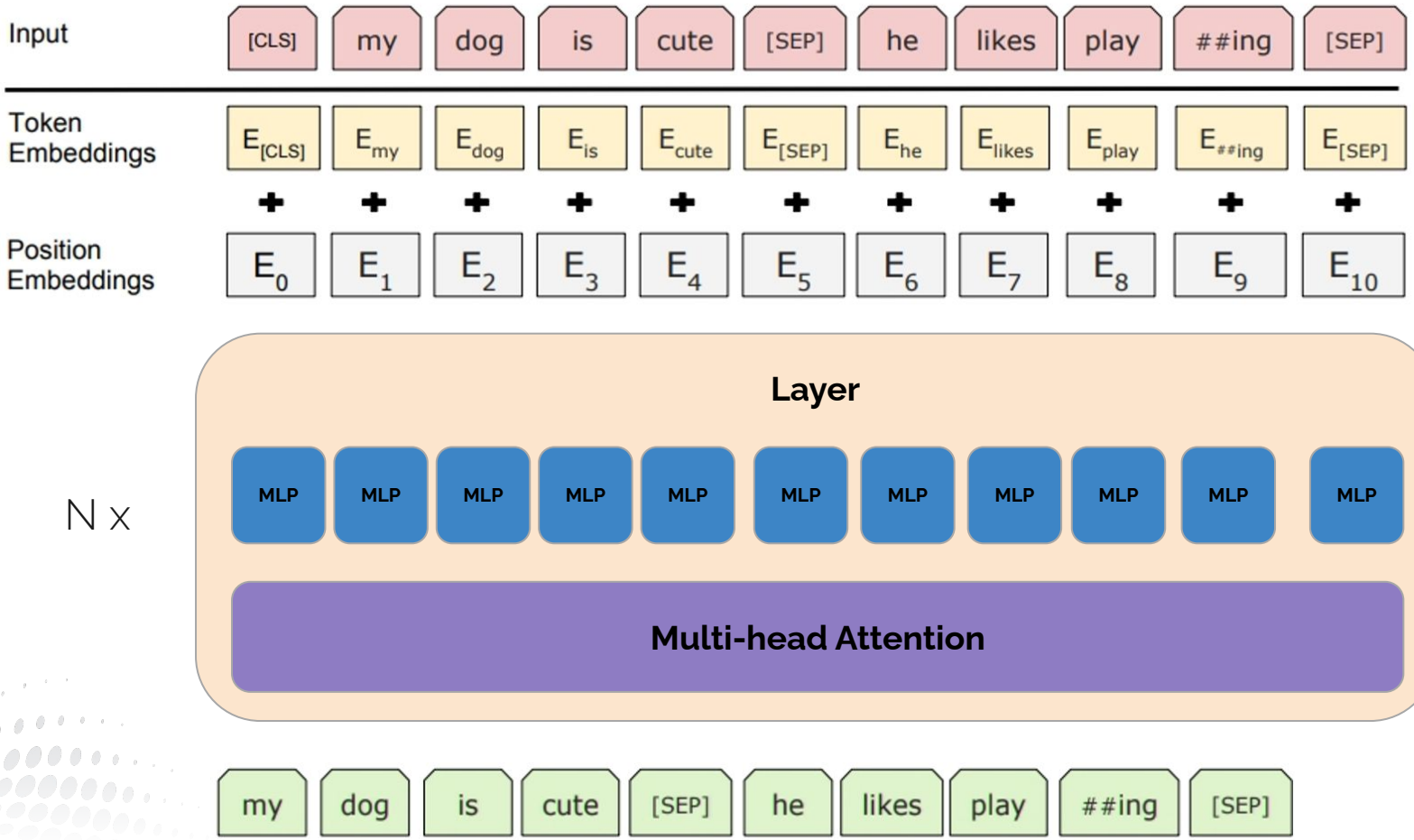
Transformers - Positional Encoding

Ordinea este integrată în reprezentarea cuvintelor

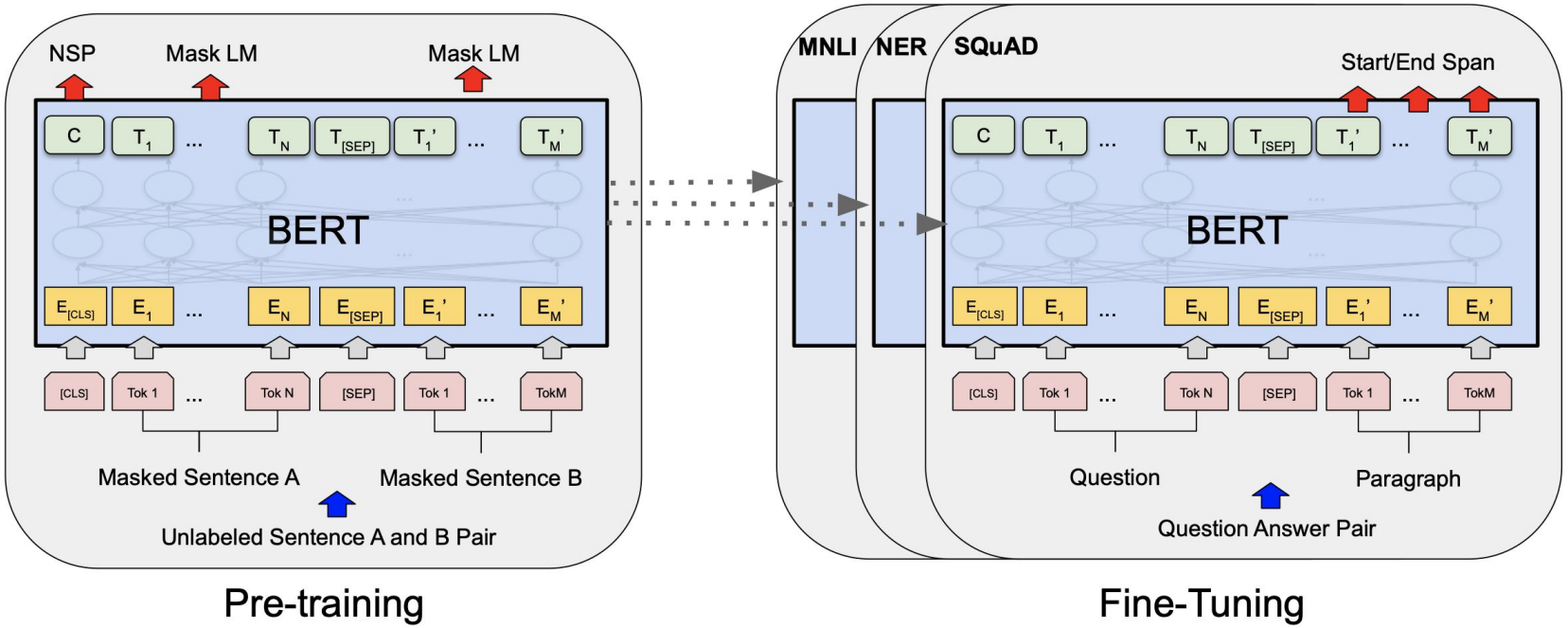
Putem paraleliza calculele, si procesa toti tokenii simultan



Sursa: Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).


Input
[CLS] my dog is cute [SEP] he likes play ##ing [SEP]
Token Embeddings
E[CLS] Emy Edog Eis Ecute E[SEP] Ehe Elikes Eplay E##ing E[SEP]
Position Embeddings
E0 E1 E2 E3 E4 E5 E6 E7 E8 E9 E10
N x
Layer
MLP MLP MLP MLP MLP MLP MLP MLP MLP MLP MLP
Multi-head Attention
my dog is cute [SEP] he likes play ##ing [SEP]

BERT



Sursa: Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Transformers - Aplicatii

LLM

Compozitie
Muzicala

ChatGPT

Alpha Fold

Generare
de imagini

Visual Transformer (ViT)

Sursa: https://ai.googleblog.com/2020/12/transformers-for-image-recognition-at.html

II. CLIP: Connecting text and images

Sursa: https://openai.com/research/clip

Notiuni preliminare - N-shot learning

Scop: învățarea din date etichetate limitate.

N: se referă la numărul de exemple etichetate disponibile per clasă pentru training (1, 5,10)

Motiv:

- Date etichetate limitate: costisitor
- Generalizare
- Tehnici de meta-learning: capturam patterns si similiaritati per tasks

Chintesenta: Se bazează pe informațiile și conexiunile invatate dintr-o mulțime mare de exemple din alte clase.

Notiuni preliminare - N-shot learning

Zero-shot learning: recunoașterea și clasificarea de noi exemple, ale căror clase **nu** au fost “văzute” în timpul antrenarii, folosind informații auxiliare.

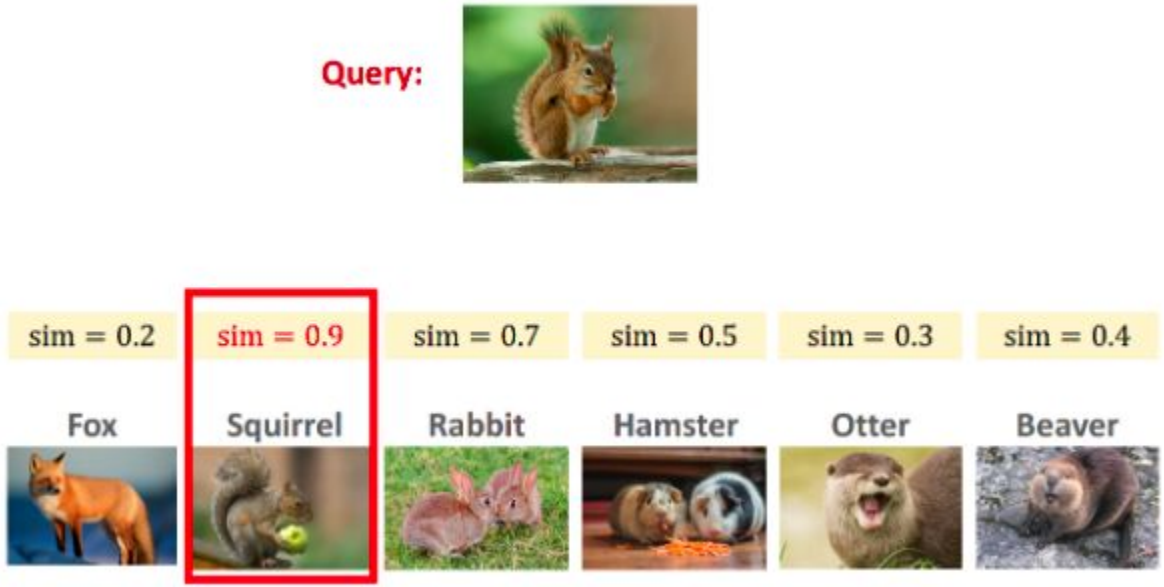
DALL·E mini by craiyon.com
Prompt: An Avocado Armchair

←run1

run2 →

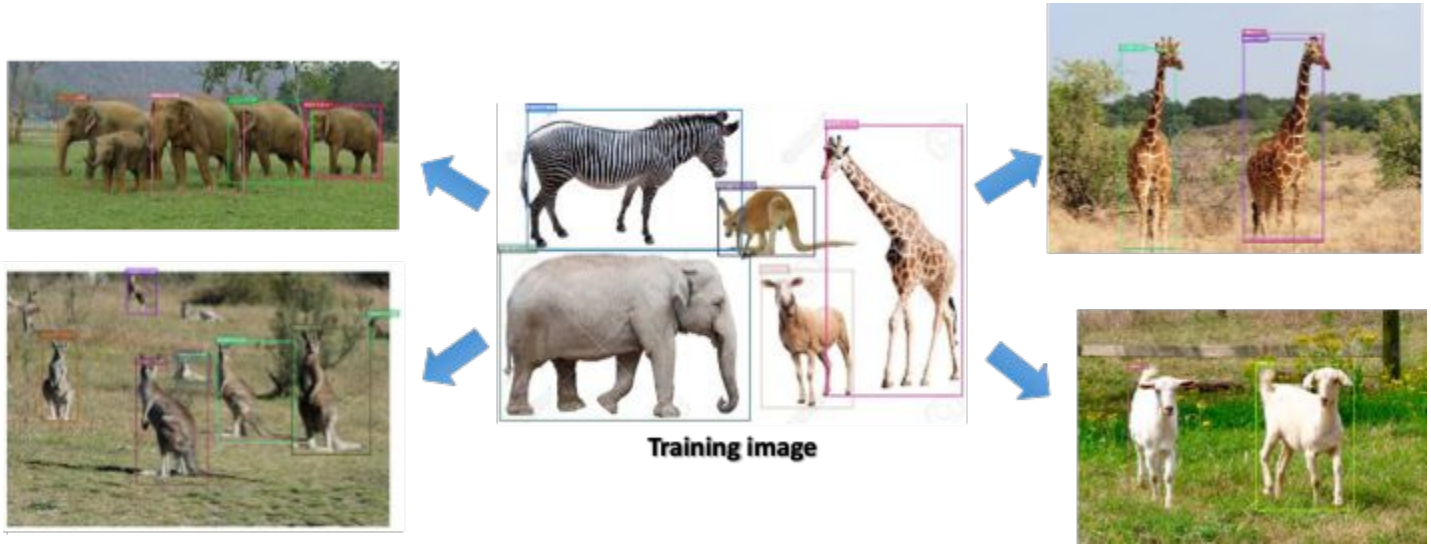
Notiuni preliminare - N-shot learning

One-shot learning: un singur exemplu etichetat al clasei.



Notiuni preliminare - N-shot learning

Few-shot learning: cateva exemple etichetate ale clasei.



Contrastive Language–Image Pre-training (CLIP)

Contrastive learning - maximizarea similitudinii între perechile similare și minimizarea similitudinii între perechile diferite

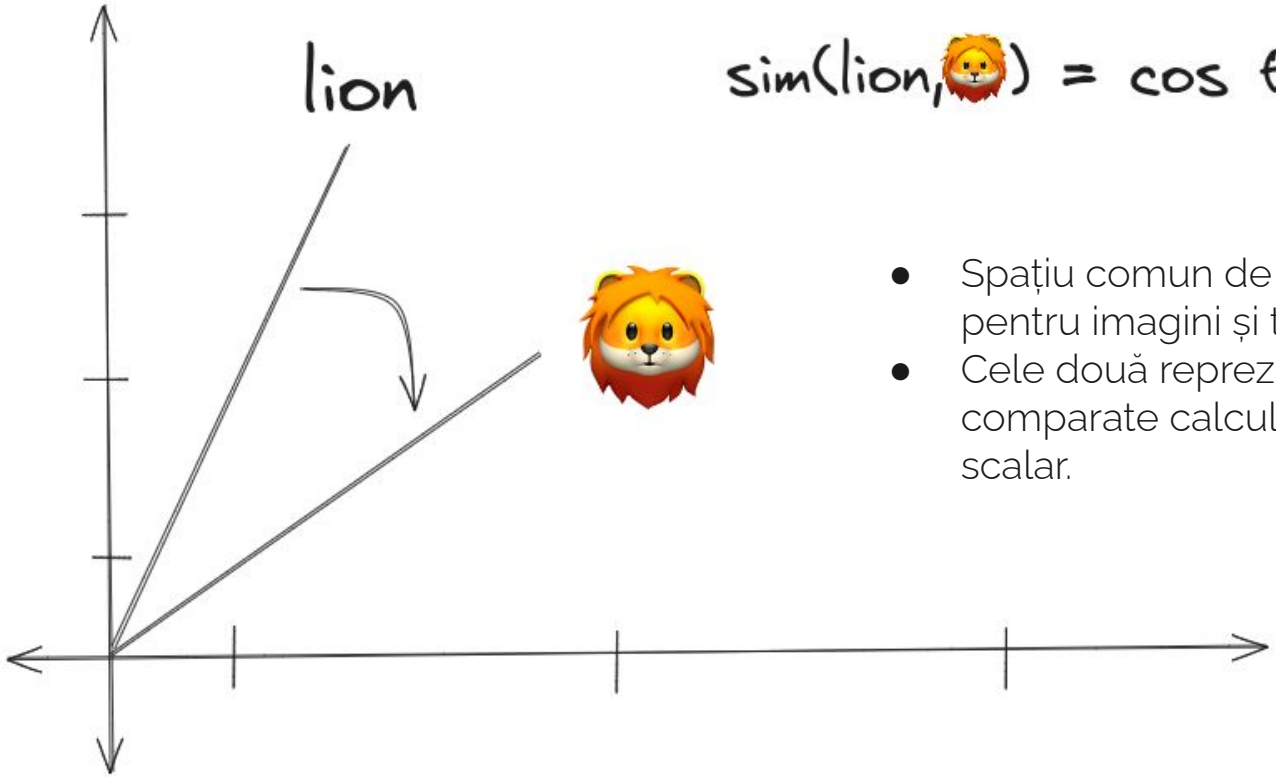
Date: Antrenat pe 400M de perechi de imagini de pe internet și descrierile asociate (e.g. titluri și etichete). Date diverse și cu mult zgomot.

Task-ul proxy: Pentru o imagine, sa se prezica, dintr-o multime de 32,768 de descrieri eșantionate aleatoriu, care este descrierea pereche.

Intuitiv: ar trebui sa invete supervizat concepte vizuale din imagini și asocierea cu numele acestora în limbaj natural.

Asocierea: Similitudinea dintre toate cele N x N perechi posibile din batch

Contrastive Language–Image Pre-training (CLIP)



- Spațiu comun de reprezentare pentru imagini și text.
- Cele două reprezentări sunt comparate calculand produsul scalar.

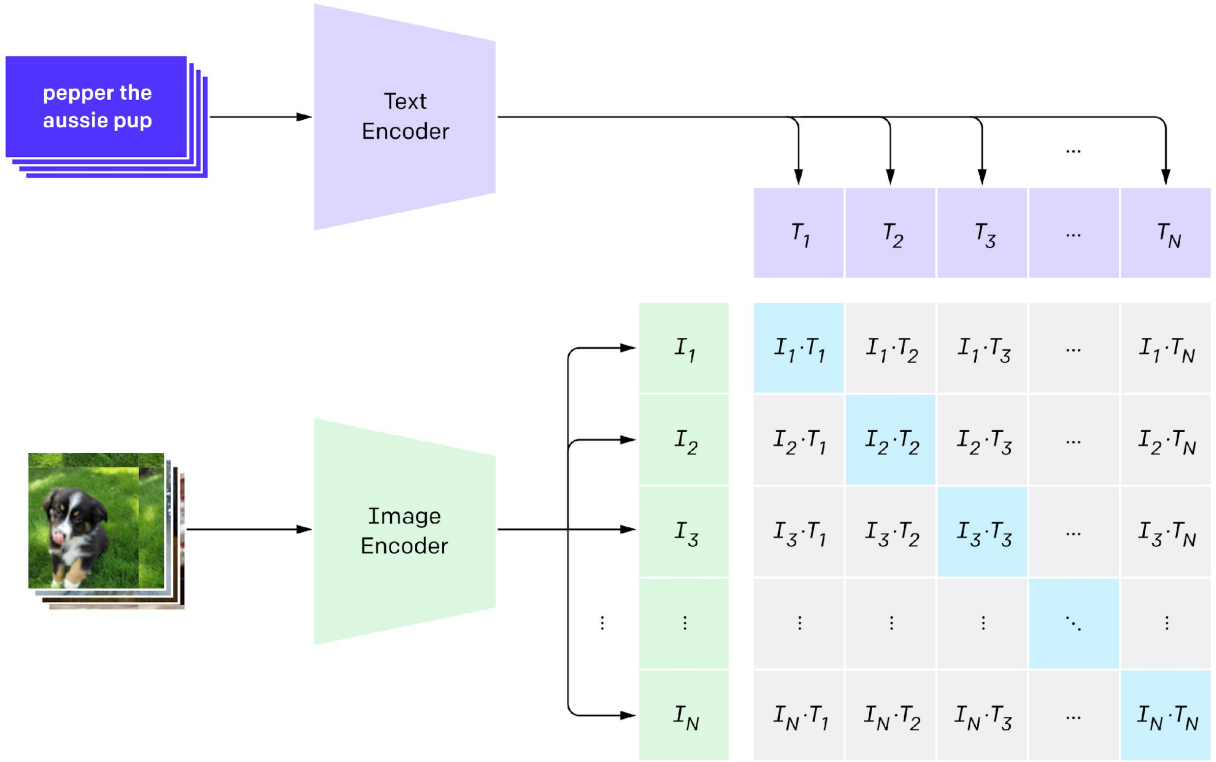
CLIP - Pseudocode

```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

Figure 3. Numpy-like pseudocode for the core of an implementation of CLIP.

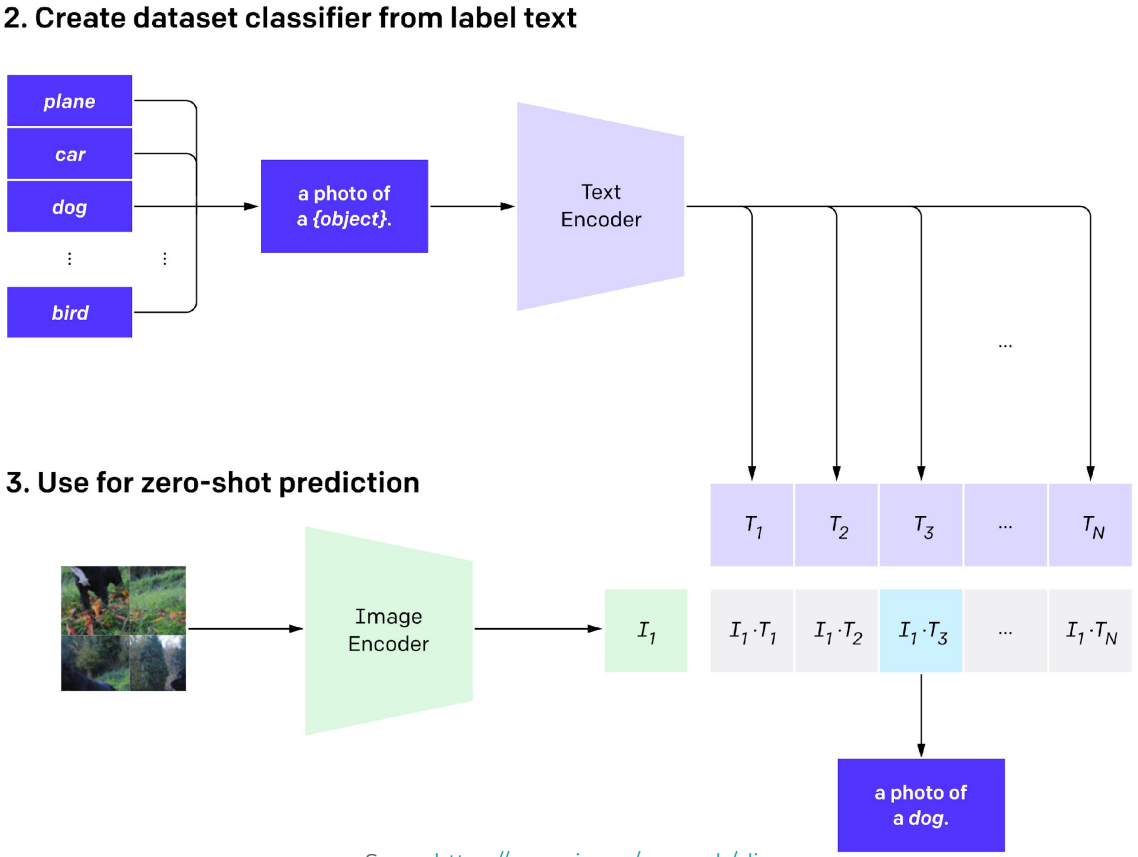
Sursa: Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

CLIP - Contrastive pre-training



Sursa: https://openai.com/research/clip

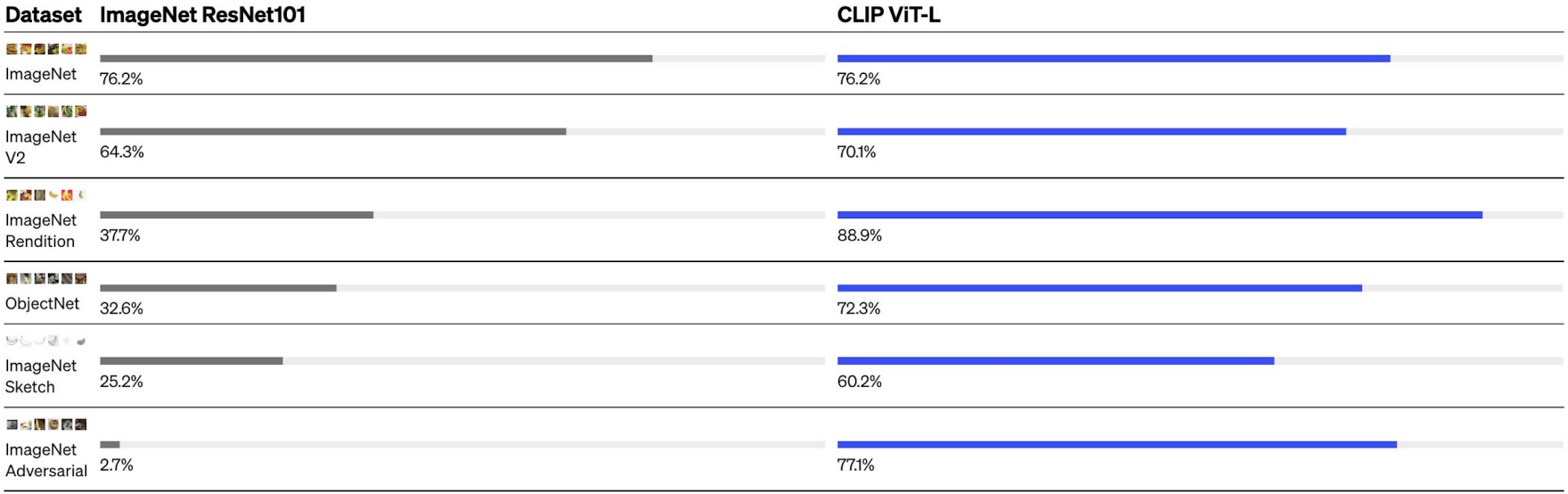
CLIP - Clasificare zero-shot



Sursa: https://openai.com/research/clip

CLIP - Performanta zero-shot

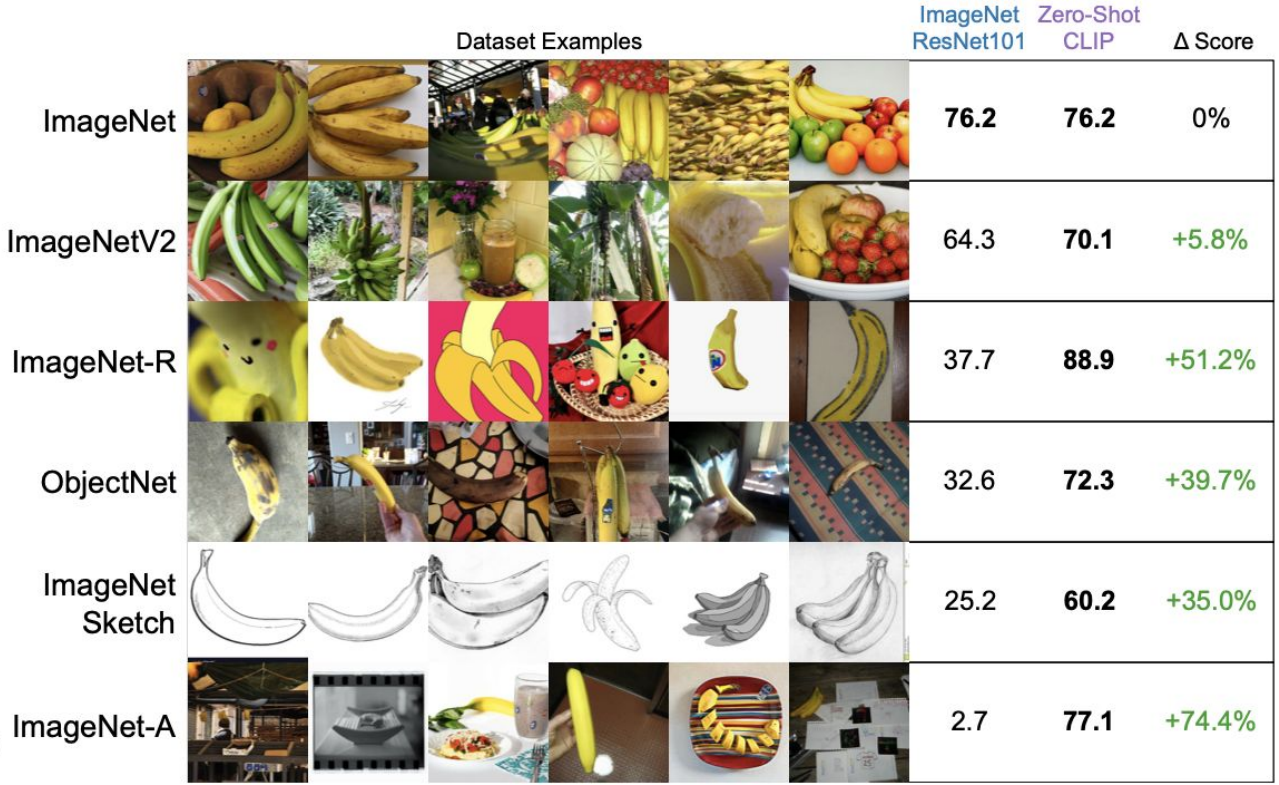
Fără niciunul din cele 1.28M de exemple originale etichete din ImageNet



Although both models have the same accuracy on the ImageNet test set, CLIP's performance is much more representative of how it will fare on datasets that measure accuracy in different, non-ImageNet settings. For instance, ObjectNet checks a model's ability to recognize objects in many different poses and with many different backgrounds inside homes while ImageNet Rendition and ImageNet Sketch check a model's ability to recognize more abstract depictions of objects.

Sursa: https://openai.com/research/clip

CLIP - Performanta zero-shot



Sursa: Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

CLIP - Probleme adresate

Costul colectării datelor:

- ImageNet-21K: 25000+ oameni, 14M+ imagini etichetate, 21k+ clase
- CLIP învață din perechi text-imagine deja disponibile public pe internet.

Modele inguste:

- Pentru a refolosi un model antrenat pe ImageNet-1K pentru un task nou, este nevoie de: modificarea retelei, colectarea de date noi si fine-tunning;
- CLIP are nevoie doar de descrierile in limbaj natural ale noilor concepte.

Performante scazute in lumea reala:

- Discrepanta intre rezultatele pe benchmark si “in the wild”;
- CLIP nu “triseaza” prin optimizarea pentru benchmark.

CLIP - Exemple

Youtube-BB

airplane, person (89.0%) Ranked 1 out of 23 labels

✓ a photo of a **airplane**.

× a photo of a **bird**.

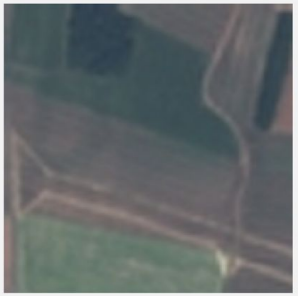
× a photo of a **bear**.

× a photo of a **giraffe**.

× a photo of a **car**.

EuroSAT

annual crop land (46.5%) Ranked 4 out of 10 labels



× a centered satellite photo of **permanent crop land**.

× a centered satellite photo of **pasture land**.

× a centered satellite photo of **highway or road**.

✓ a centered satellite photo of **annual crop land**.

× a centered satellite photo of **brushland or shrubland**.

Food101

guacamole (90.1%) Ranked 1 out of 101 labels

✓ a photo of **guacamole**, a type of food.

× a photo of **ceviche**, a type of food.

× a photo of **edamame**, a type of food.

× a photo of **tuna tartare**, a type of food.

× a photo of **hummus**, a type of food.

SUN397

television studio (90.2%) Ranked 1 out of 397 labels

✓ a photo of a **television studio**.

× a photo of a **podium indoor**.

× a photo of a **conference room**.

× a photo of a **lecture room**.

× a photo of a **control room**.

Sursa: https://openai.com/research/clip

CLIP - Limitari

- Task-uri abstracte (e.g. numarare);
- Task-uri complexe (e.g. estimarea distanțelor);
- Task-uri foarte specifice (e.g. clasificarea unor specii de flori);

III. Multimodal Deep Learning

Context

- Subset al DL care se ocupă cu fuziunea și analiza datelor: text, imagini, video, audio, senzori.
- Multiple “strengths” + reprezentare cat mai completa a datelor → Multiple tasks

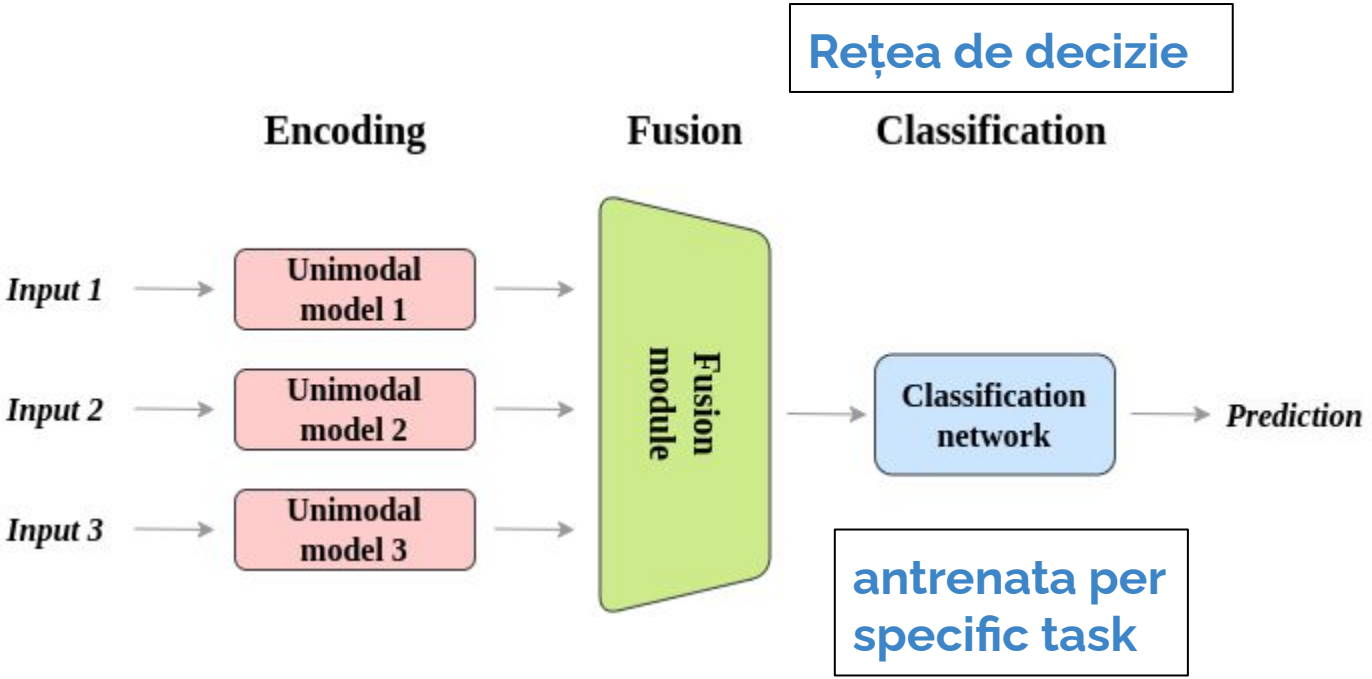
Scop: spațiu de reprezentare comun.

Arhitectura:

- Multiple NNs, fiecare avand *specializarea* sa.
- Output-ul acestora este combinat folosind diferite tipuri de fuzionare (e.g. early fusion, late fusion, hybrid fusion).

Tasks: image captioning, speech recognition.

Workflow general



Sursa: https://blog.roboflow.com/multimodal-models/

StudyCase: CogVLM paper

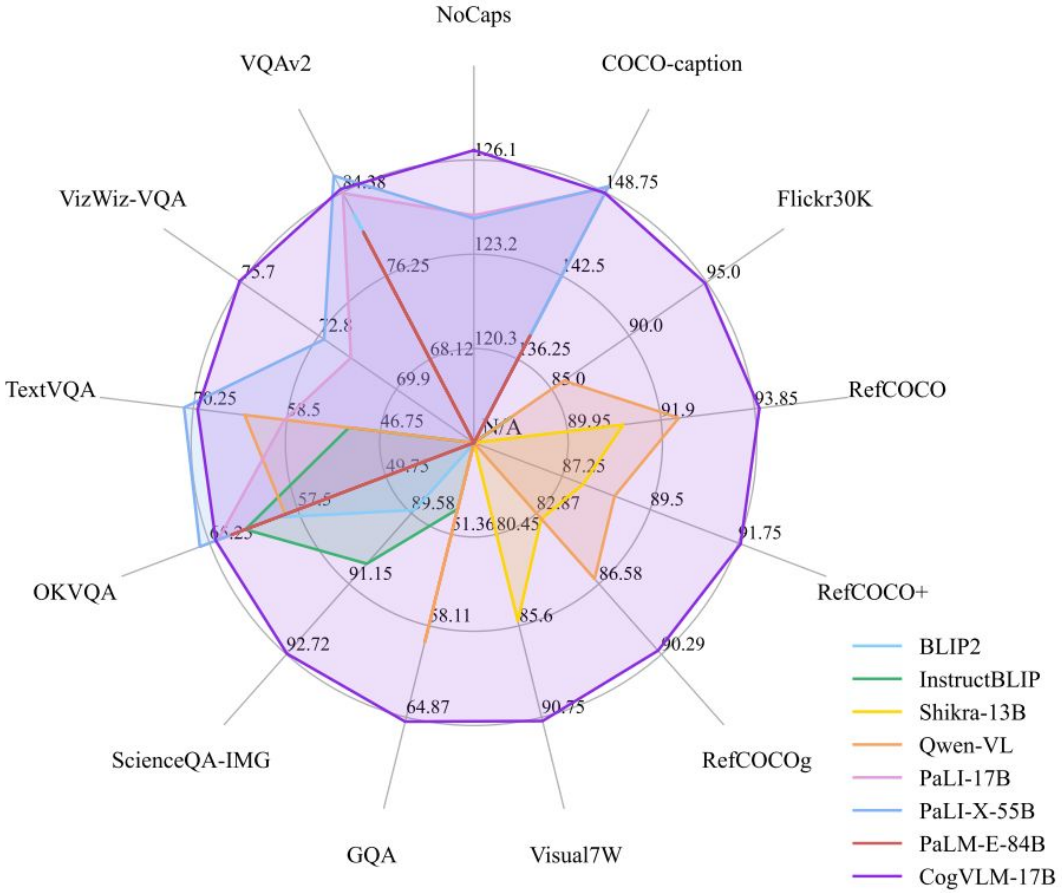
Main and important ideas:

- It sets itself apart from traditional methods by avoiding the *shallow alignment* technique, which directly maps image features into the input space of the language model.

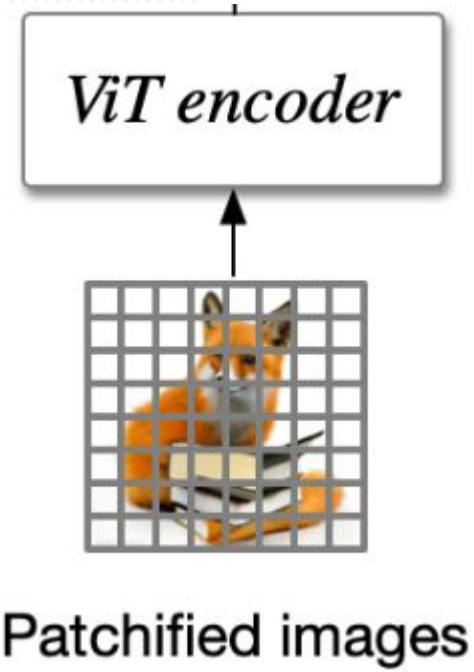
- Instead, CogVLM integrates a **trainable visual expert module into the attention** and feed-forward neural network (**FFN**) **layers**, bridging the gap between the frozen pretrained language model and the image encoder.

- This integration enables **deep fusion of vision** and **language features** while maintaining performance on natural language processing (NLP) tasks.

StudyCase: CogVLM paper



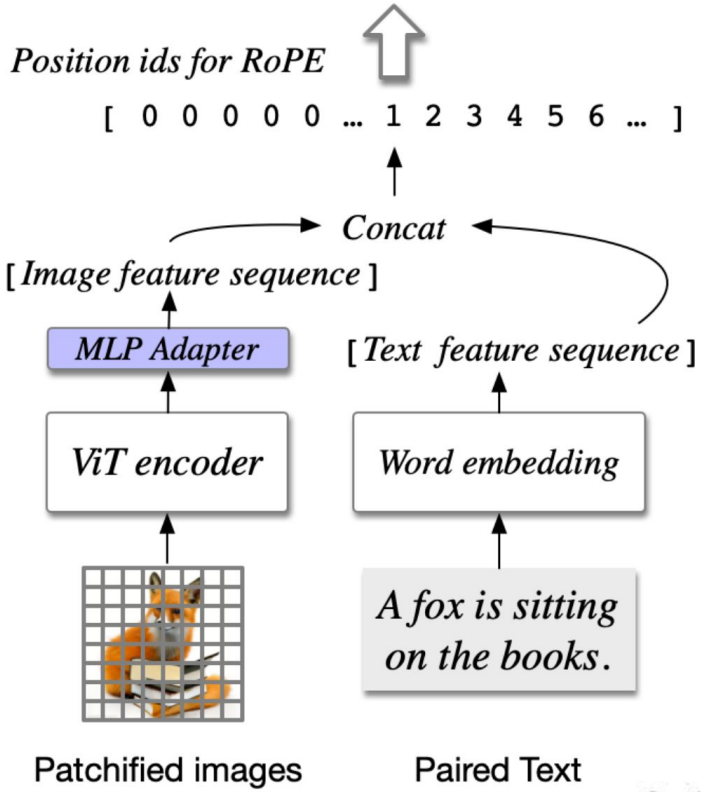
StudyCase: CogVLM paper



1st Component (ViT Encoder) - Highlights:

- The authors employ the pretrained EVA2-CLIP-E model in CogVLM-17B;
- In the process, the final layer of the Vision Transformer (ViT) encoder is removed.
- This modification is made because the final layer specializes in aggregating the [CLS] features for contrastive learning.

StudyCase: CogVLM paper

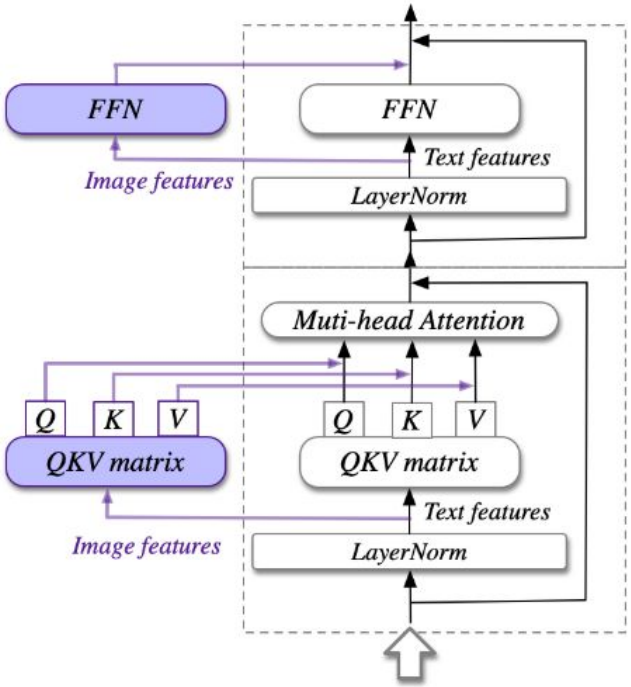


2nd Component (Combining Features) - Highlights:

- Word embeddings for text and MLP projection for image features
- We keep all the information via concatenation
- The entire image shares the same position id across all its features.

. Only the purple parts are trainable.

StudyCase: CogVLM paper



(b) The visual expert built on the language model Only the purple parts are trainable.

3rd Component (Visual Expert Module) - Highlights:

- **Deep Fusion:** Trainable visual expert enables the transformation of image features to align with the various attention heads in the language model.
- The visual expert module exists within each layer → **deep visual-language feature alignment.**
- At its core, the visual expert module consists of a QKV Matrix and a MLP layer.

StudyCase: CogVLM paper

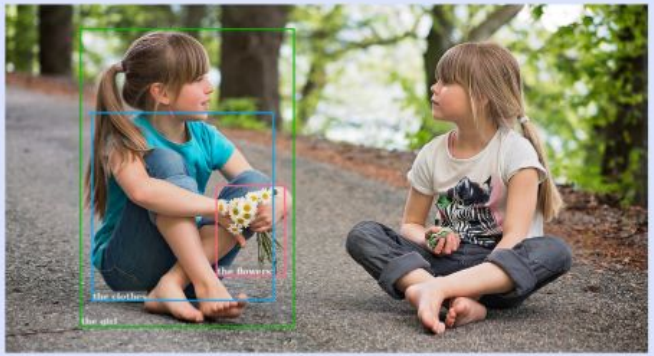
Multi-modal benchmarks:

- Image Captioning
- Visual Question Answering
- Visual Grounding (links between a textual mention & a specific region)

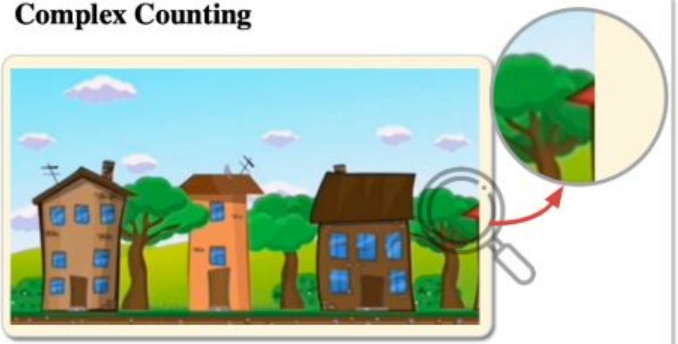
Grounding Visual Question Answering

Q1: What color are the clothes of the girl whose hands are holding flowers? Let's think step by step.

CogVLM: Find the flowers [[328,519,436,780]]. Check the girl holding it, got the girl [[116,069,450,927]]. Check the clothes of it, got the clothes [[134,309,418,852]]. The clothes [[134,309,418,852]] are blue. So the answer is blue.



Complex Counting



Q1: How many houses are there in this cartoon?

CogVLM: There are four houses in this cartoon.

Q2: Why do I only notice three?

CogVLM: The reason you only see three houses is because the fourth house, which should be located on the right side of the picture, cannot be seen clearly.

Q3: How many houses which can be seen completly in this image?

CogVLM: In this image, only three houses can be seen completely. The fourth house is partially visible and cannot be fully identified.

Resurse

- https://poloclub.github.io/transformer-explainer/
- https://www.youtube.com/watch?v=wjZofJX0v4M
- https://www.youtube.com/watch?v=eMlx5fFNoYc