



# Attention & Vision Transformers

Kartik Patwari  
Nov 25, 2024

# CNNs Recap

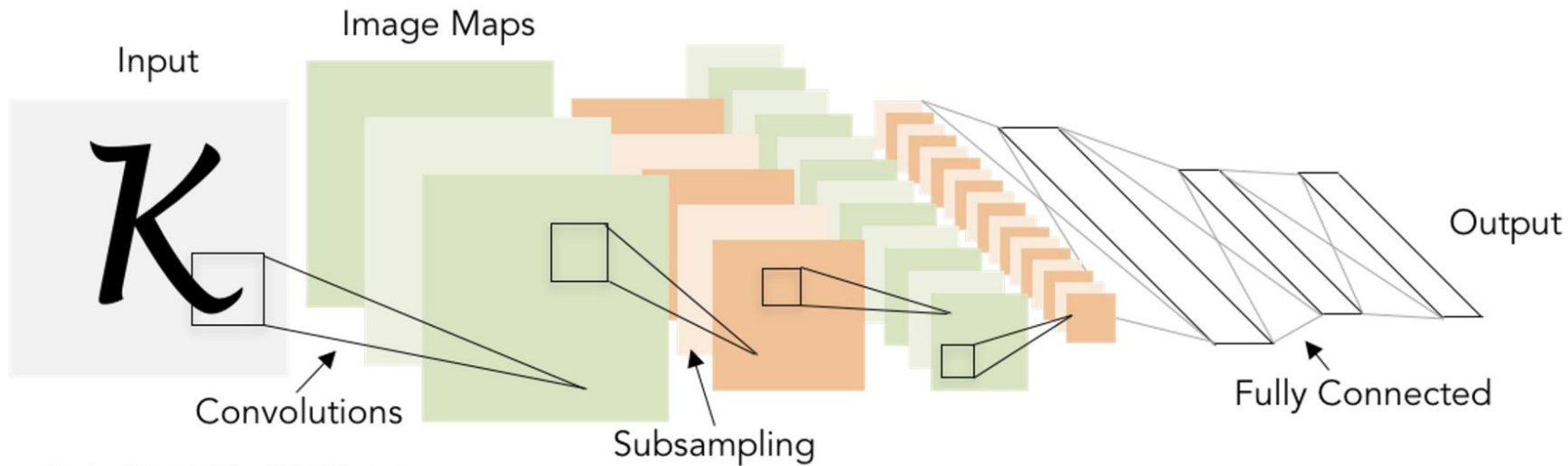
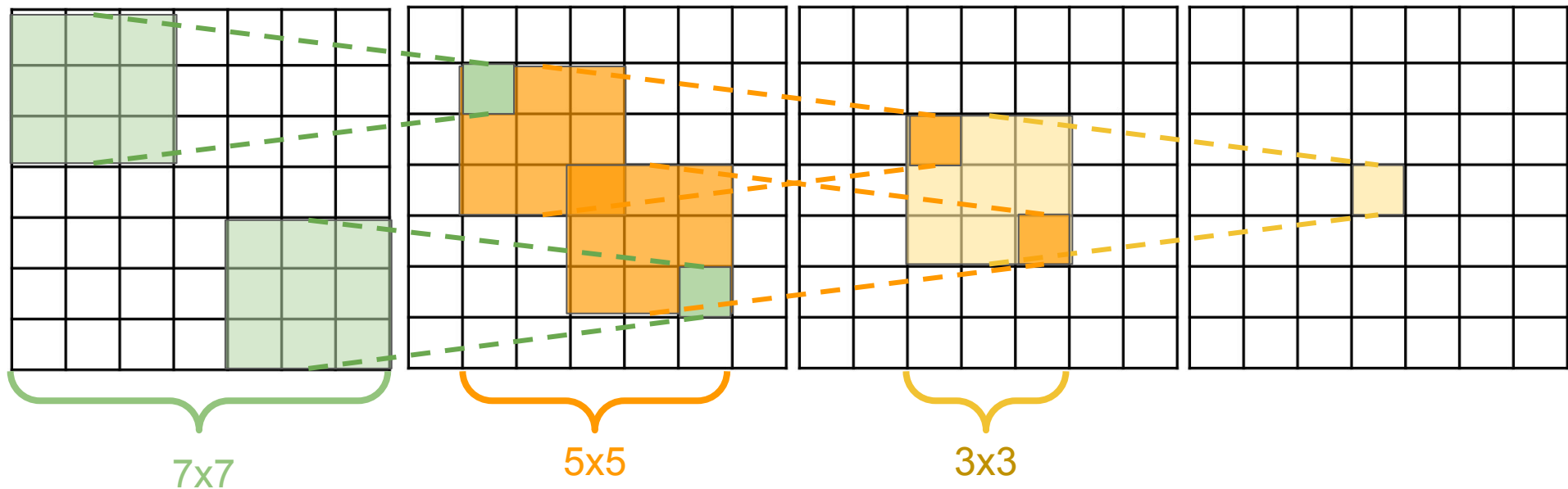
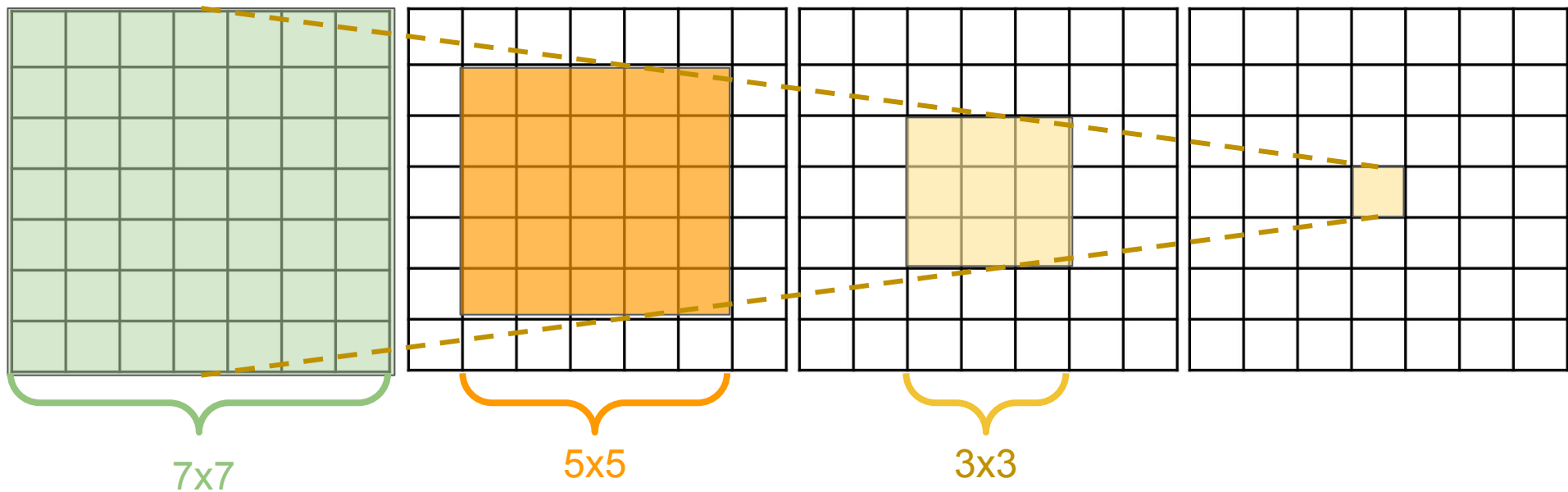


Illustration of LeCun et al. 1998 from CS231n 2017 Lecture 1

# CNN Limitations: Receptive Fields

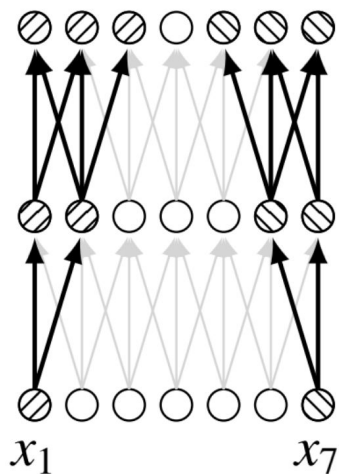


# CNN Limitations: Receptive Fields

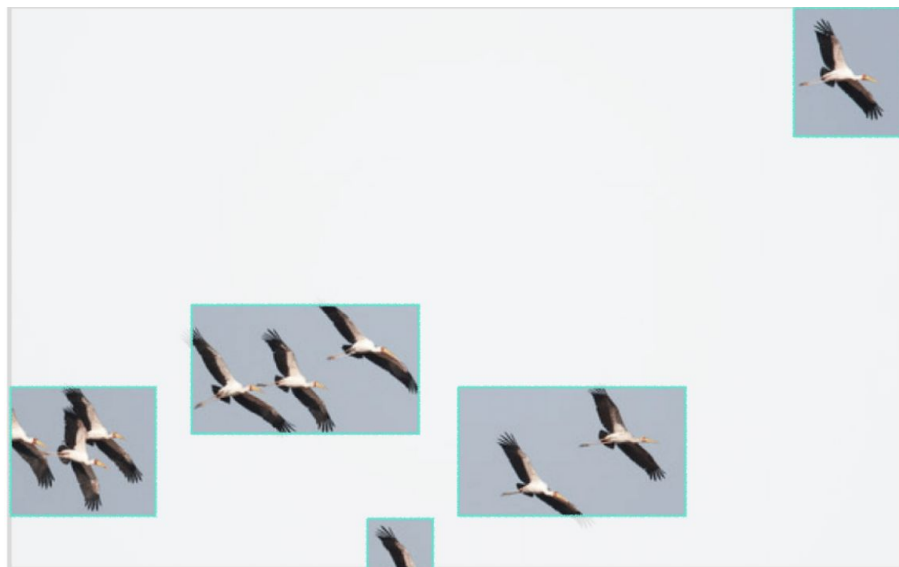
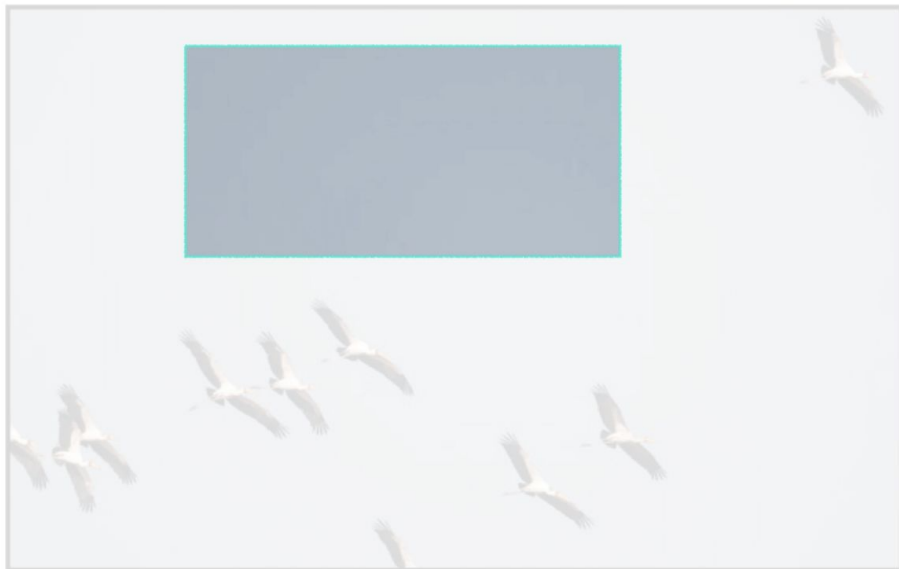


# CNN Limitations

- CNNs are built around **locality**
- Not well suited to modeling **long distance relationships**
- Far image regions do not interact



Can we focus on specific regions?



# Attention Is All You Need

## 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

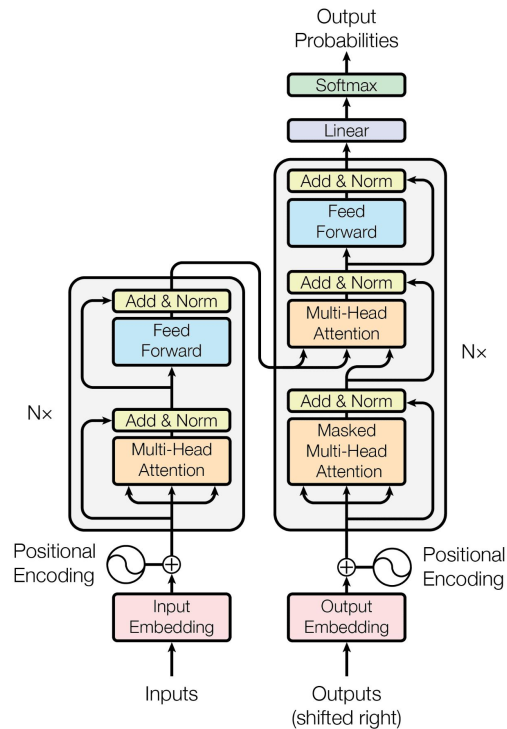
Łukasz Kaiser\*  
Google Brain  
lukasz.kaiser@google.com

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, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

arXiv:1706.03762v7 [cs.CL] 2 Aug 2023



# Attention Is All You Need: Attention, Transformers, ViTs

- Tokens
- Attention
- Self & Multi-head Attention
- Vision Transformers





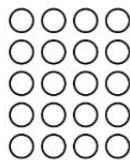
# Tokens

- Tokens can be seen as a vector of neurons
- Encapsulate bundle of information

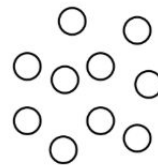
array of **neurons**



array of **neurons**



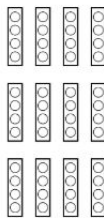
set of **neurons**



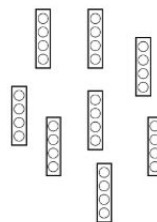
array of **tokens**



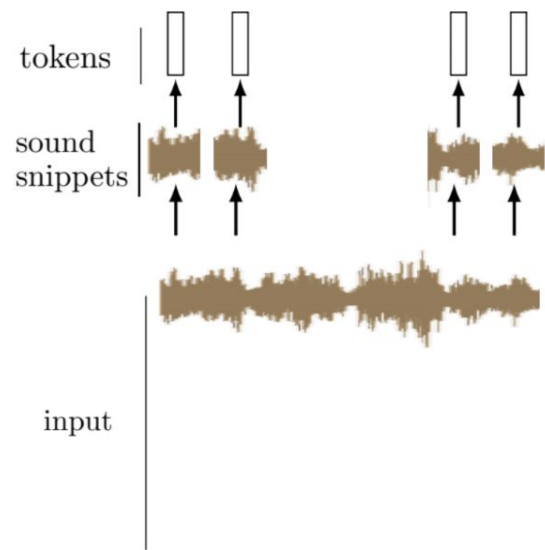
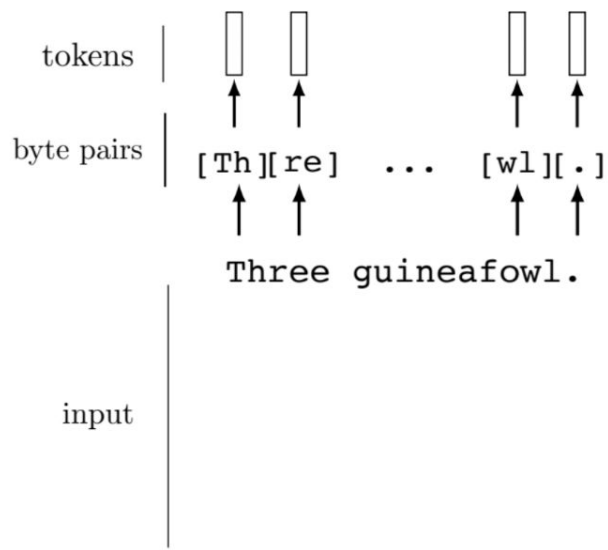
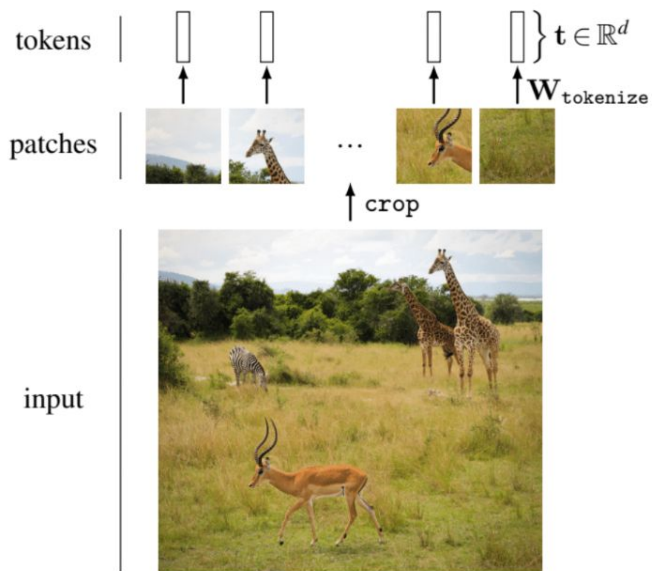
array of **tokens**



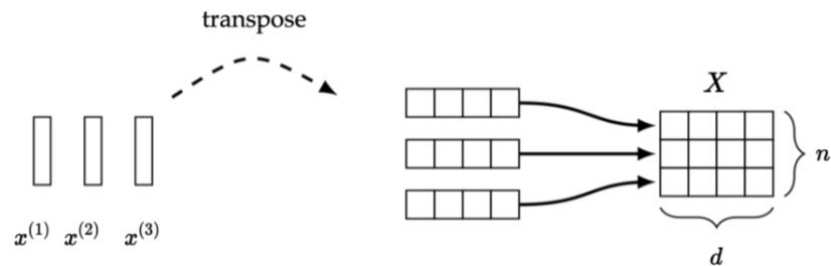
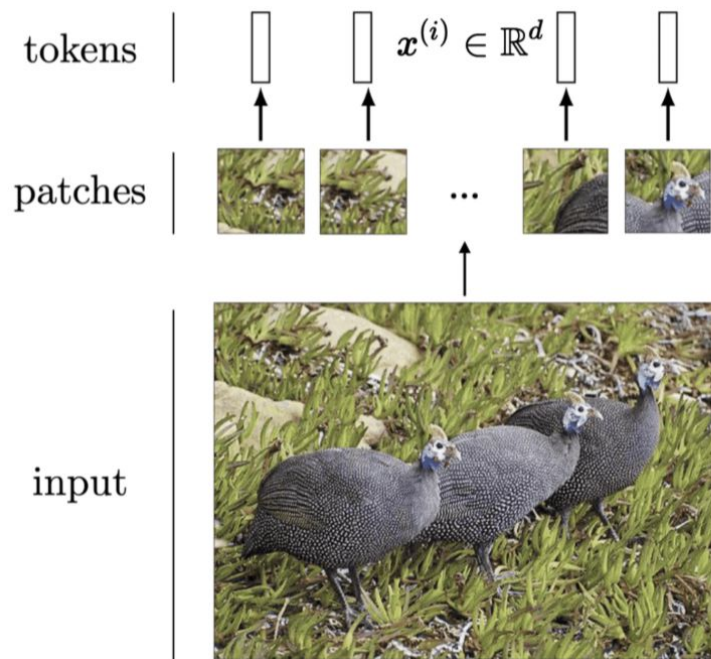
set of **tokens**



# Tokenization



# Tokenization



- $d$  is the size of each token ( $x^{(i)} \in \mathbb{R}^d$ )
- $n$  is the number of tokens

Attention: What should “goat” be?

The sheep and **goat** are grazing.



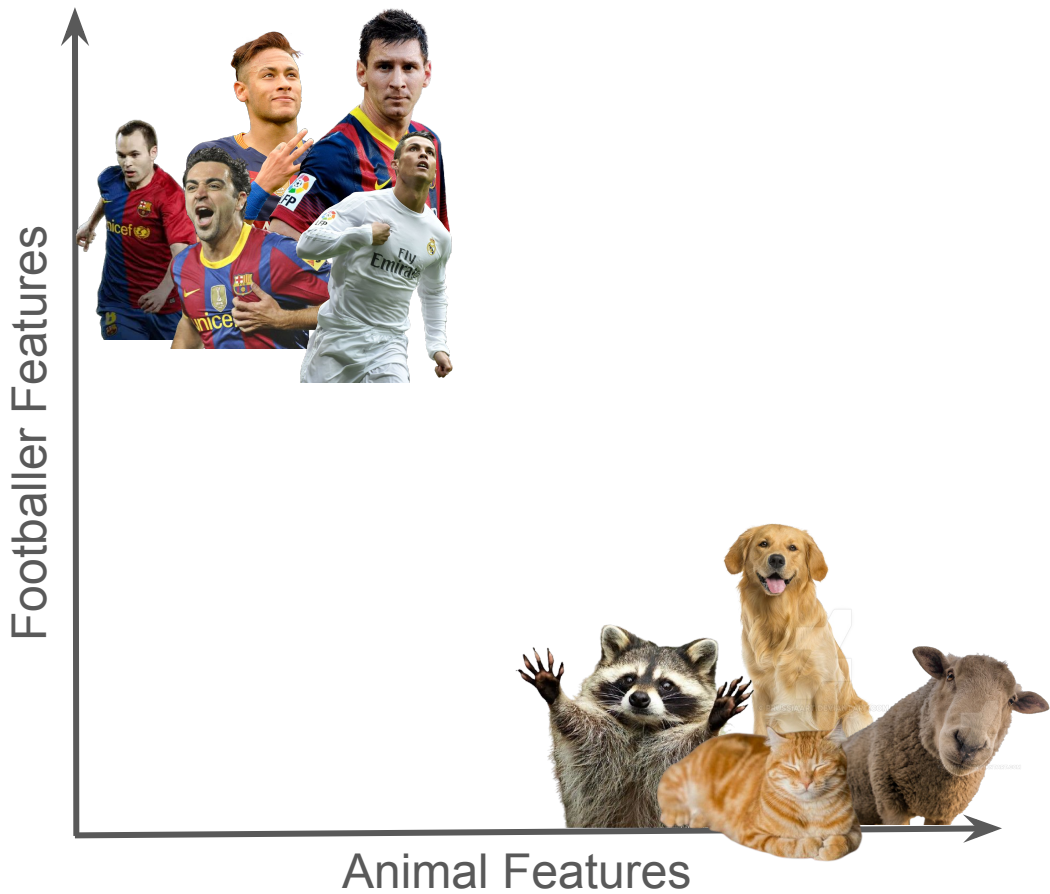
Q: What does **goat** mean here? **Animal**

Messi is the **goat**, not Ronaldo.



Q: What does **goat** mean here? **Footballer (Greatest of all time)**

# Embedding Space



$$E = [\textit{Footballer feature}, \textit{Animal feature}]$$

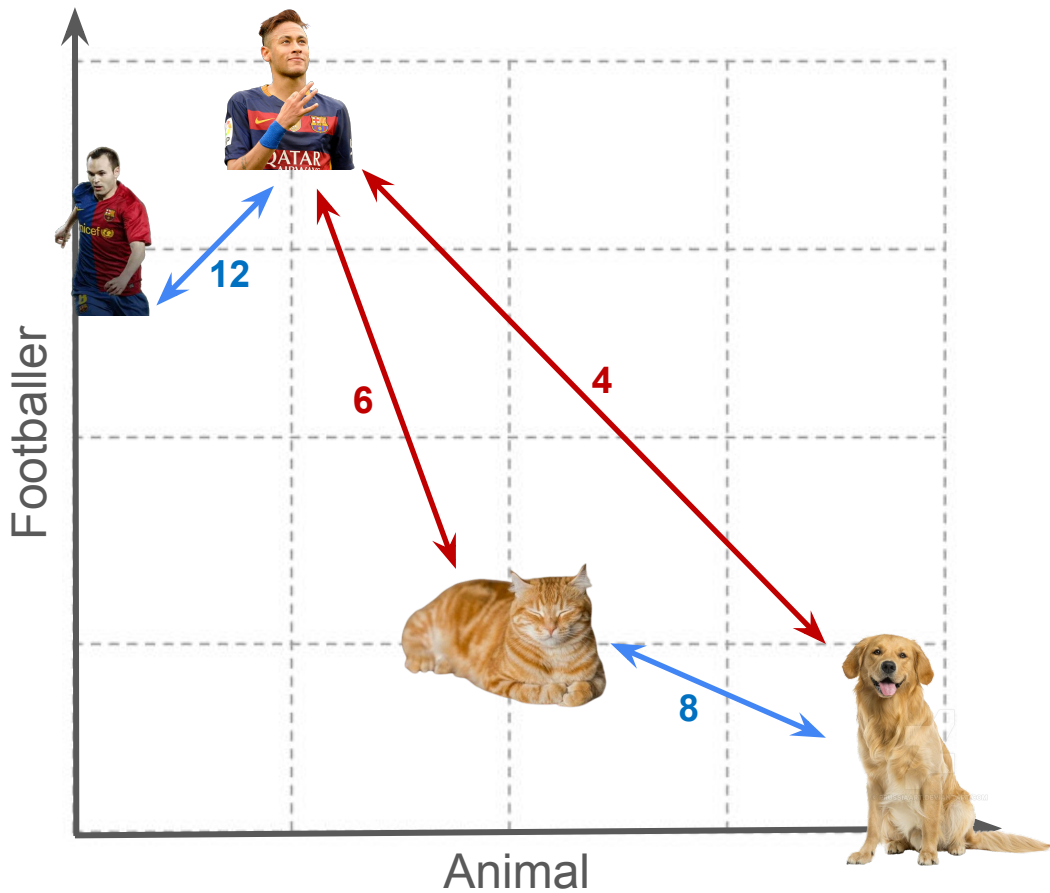


Messi is the **goat**, not Ronaldo.



The sheep and **goat** are grazing.

# Similarity: Dot Product



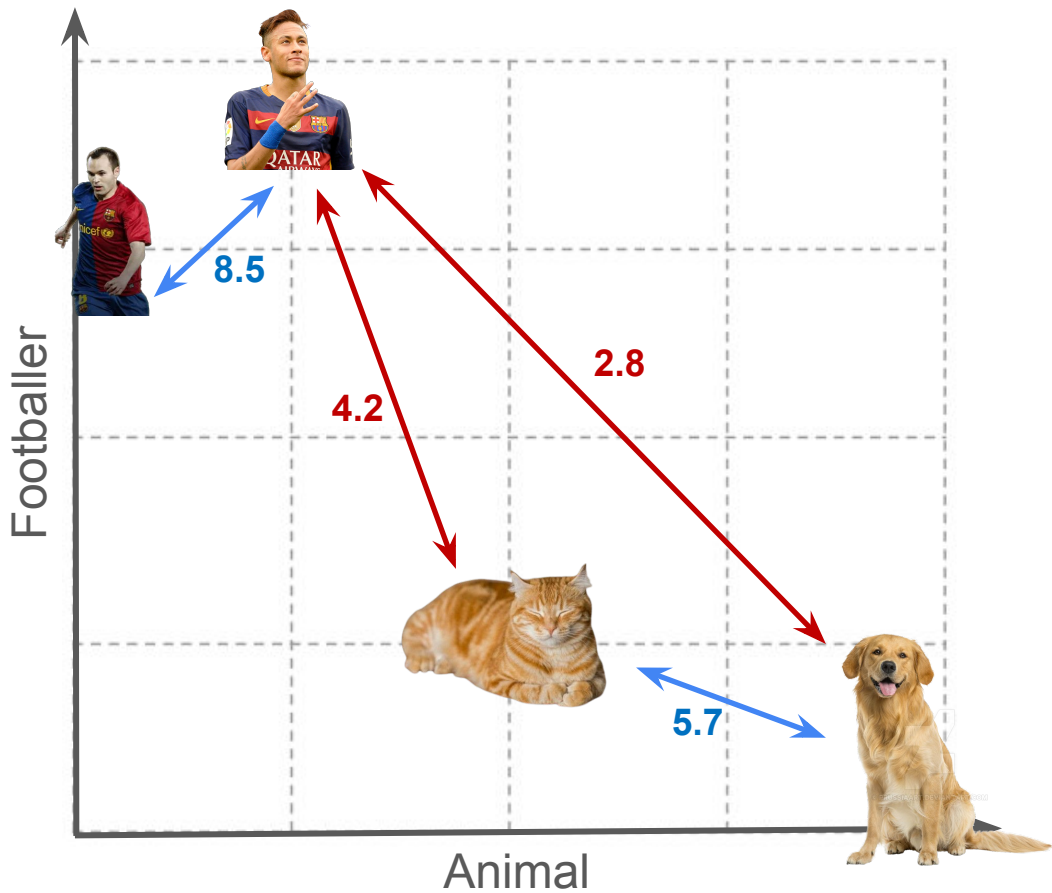
$$\text{sim}\left(\begin{array}{c} \text{Footballer 1} \\ \text{Footballer 2} \end{array}, \begin{array}{c} \text{Footballer 1} \\ \text{Footballer 2} \end{array}\right) = \begin{bmatrix} 1 \\ 4 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 3 \end{bmatrix} = 12$$

$$\text{sim}\left(\begin{array}{c} \text{Dog} \\ \text{Cat} \end{array}, \begin{array}{c} \text{Dog} \\ \text{Cat} \end{array}\right) = \begin{bmatrix} 4 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 1 \end{bmatrix} = 8$$

$$\text{sim}\left(\begin{array}{c} \text{Footballer 1} \\ \text{Footballer 2} \end{array}, \begin{array}{c} \text{Dog} \\ \text{Cat} \end{array}\right) = \begin{bmatrix} 1 \\ 4 \end{bmatrix} \cdot \begin{bmatrix} 4 \\ 0 \end{bmatrix} = 4$$

$$\text{sim}\left(\begin{array}{c} \text{Footballer 1} \\ \text{Footballer 2} \end{array}, \begin{array}{c} \text{Cat} \\ \text{Dog} \end{array}\right) = \begin{bmatrix} 1 \\ 4 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 1 \end{bmatrix} = 6$$

# Similarity: Scaled Dot Product



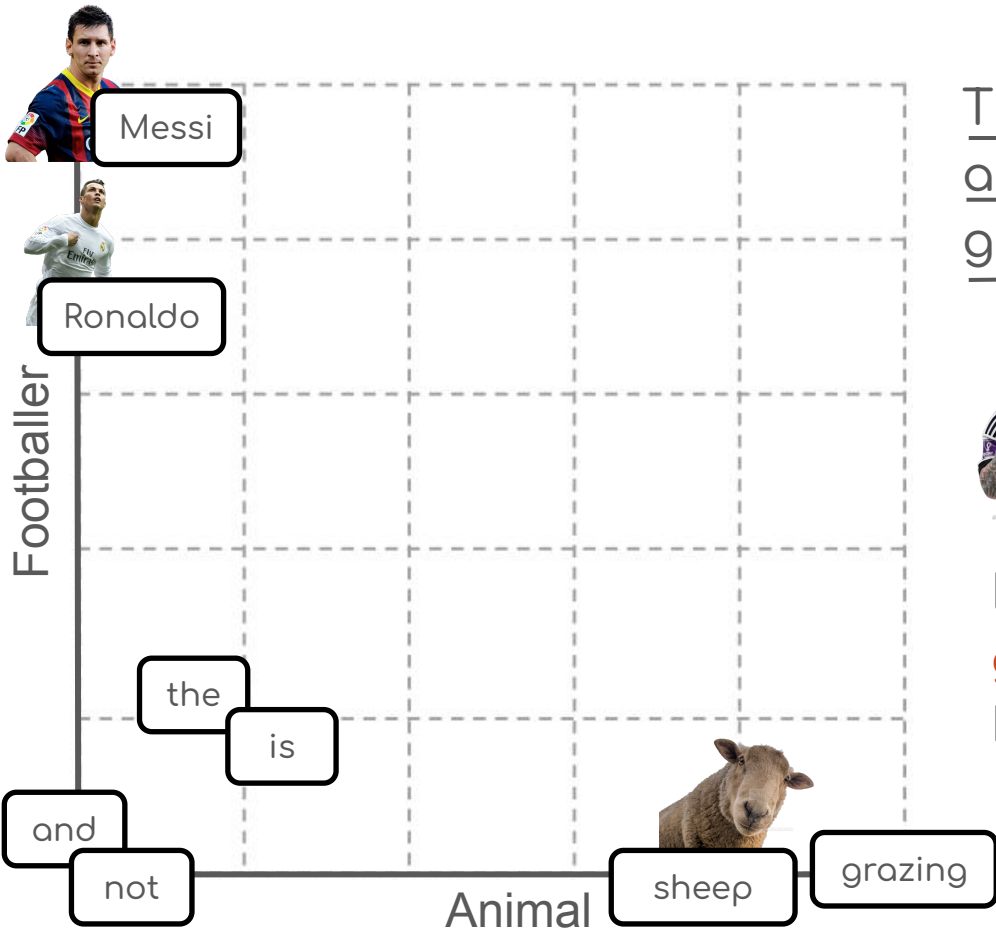
$$\text{sim}\left(\begin{array}{c} \text{Footballer 1} \\ \text{Footballer 2} \end{array}, \begin{array}{c} \text{Footballer 1} \\ \text{Footballer 2} \end{array}\right) = \frac{\begin{bmatrix} 1 \\ 4 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 3 \end{bmatrix}}{\sqrt{2}} \sim 8.5$$

$$\text{sim}\left(\begin{array}{c} \text{Dog} \\ \text{Cat} \end{array}, \begin{array}{c} \text{Dog} \\ \text{Cat} \end{array}\right) = \frac{\begin{bmatrix} 4 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 1 \end{bmatrix}}{\sqrt{2}} \sim 5.7$$

$$\text{sim}\left(\begin{array}{c} \text{Footballer 1} \\ \text{Footballer 2} \end{array}, \begin{array}{c} \text{Dog} \\ \text{Cat} \end{array}\right) = \frac{\begin{bmatrix} 1 \\ 4 \end{bmatrix} \cdot \begin{bmatrix} 4 \\ 0 \end{bmatrix}}{\sqrt{2}} \sim 2.8$$

$$\text{sim}\left(\begin{array}{c} \text{Footballer 1} \\ \text{Footballer 2} \end{array}, \begin{array}{c} \text{Cat} \\ \text{Dog} \end{array}\right) = \frac{\begin{bmatrix} 1 \\ 4 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 1 \end{bmatrix}}{\sqrt{2}} \sim 4.2$$

# What should “goat” be?



The sheep  
and **goat** are  
grazing.

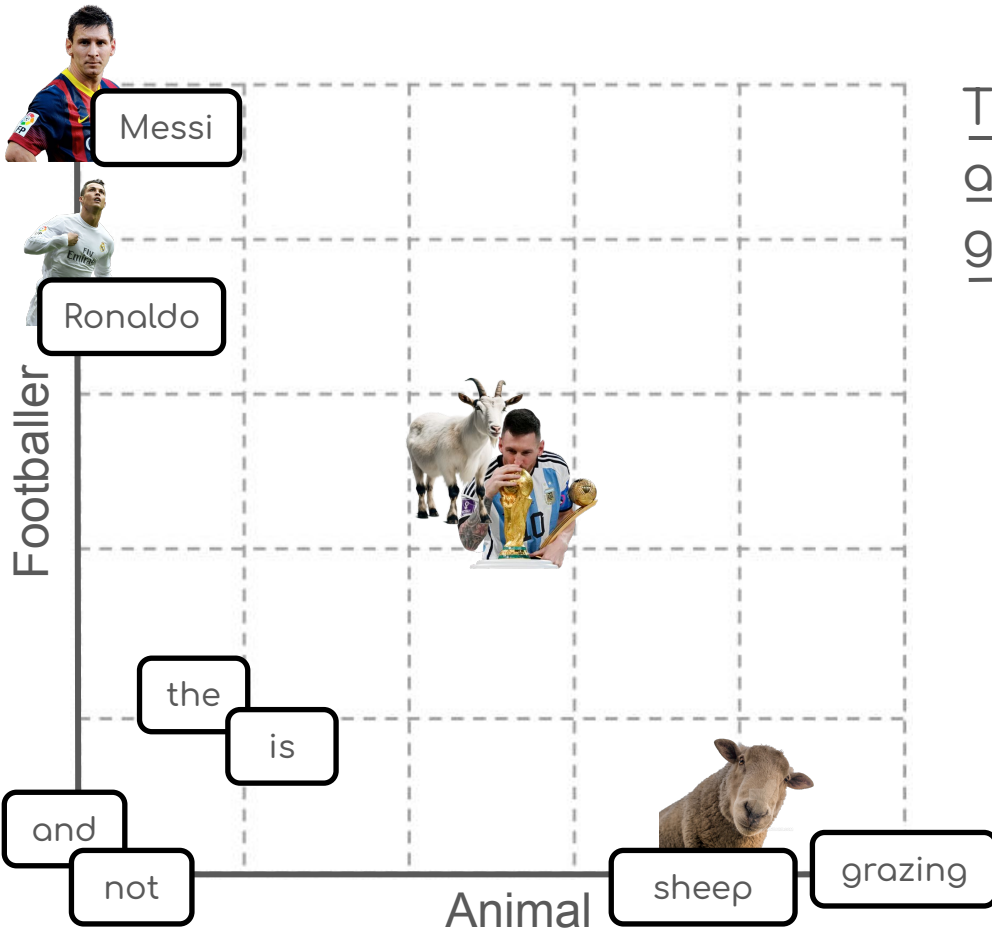


Messi is the  
**goat**, not  
Ronaldo.

Word/Token	Embedding
the	[1,1]
sheep	[4,0]
and	[0,0]
<b>goat</b>	<b>???</b>
grazing	[5,0]
Messi	[0,5]
is	[1,1]
not	[0,0]
Ronaldo	[0,4]



# What should “goat” be?



The sheep  
and **goat** are  
grazing.

Messi is the  
**goat**, not  
Ronaldo.

Word	Embedding
the	[1,1]
sheep	[4,0]
and	[0,0]
<b>goat</b>	<b>[2.5,2.5]</b>
grazing	[5,0]
Messi	[0,5]
is	[1,1]
not	[0,0]
Ronaldo	[0,4]

# What should “goat” be?

$S_1$ : The sheep and **goat** are grazing.

$$Q, K, V = \begin{bmatrix} \text{the} & [1,1] \\ \text{sheep} & [4,0] \\ \text{and} & [0,0] \\ \text{goat} & [2.5,2.5] \\ \text{grazing} & [5,0] \end{bmatrix}$$

$S_2$ : Messi is the **goat**, not Ronaldo.

$$Q, K, V = \begin{bmatrix} \text{Messi} & [0,5] \\ \text{is} & [1,1] \\ \text{goat} & [2.5,2.5] \\ \text{not} & [0,0] \\ \text{Ronaldo} & [0,4] \end{bmatrix}$$

Word	Q,K,V
the	[1,1]
sheep	[4,0]
and	[0,0]
<b>goat</b>	<b>[2.5,2.5]</b>
grazing	[5,0]
Messi	[0,5]
is	[1,1]
not	[0,0]
Ronaldo	[0,4]

\*Note: We take Q,K,V as same for this simple illustrative example

# Attention: Calculate Scores

Attention scores are computed as the dot product of the **Query (Q)** of the target word (**goat**) with the **Keys (K)** of all words in the sentence.

$$Score = Q_{goat} \cdot K^T$$

$$Scores(S_1) = [2.5, 2.5] \cdot \begin{bmatrix} [1,1] \\ [4,0] \\ [0,0] \\ [2.5,2.5] \\ [5,0] \end{bmatrix} = \begin{bmatrix} 5 \\ 10 \\ 0 \\ 12.5 \\ 12.5 \end{bmatrix}$$

$$Scores(S_2) = [2.5, 2.5] \cdot \begin{bmatrix} [0,5] \\ [1,1] \\ [2.5,2.5] \\ [0,0] \\ [0,4] \end{bmatrix} = \begin{bmatrix} 12.5 \\ 5 \\ 12.5 \\ 0 \\ 10 \end{bmatrix}$$

$S_1$ :

The sheep and **goat** are grazing.

$S_2$ :

Messi is the **goat**, not Ronaldo.

Word	Q,K,V
the	[1,1]
sheep	[4,0]
and	[0,0]
<b>goat</b>	<b>[2.5,2.5]</b>
grazing	[5,0]
Messi	[0,5]
is	[1,1]
not	[0,0]
Ronaldo	[0,4]

# Attention: Scale Scores

We scale the scores by dividing by  $\sqrt{d_k}$ ,  
where  $d_k = 2$  (embedding dimension):

$$\text{Scaled Score} = \frac{Q_{goat} \cdot K^T}{\sqrt{d_k}}$$

$$\text{Scaled Scores}(S_1) = \begin{bmatrix} 5 \\ 10 \\ 0 \\ 12.5 \\ 12.5 \end{bmatrix} / \sqrt{2} = \begin{bmatrix} 3.55 \\ 7.08 \\ 0 \\ 8.85 \\ 8.85 \end{bmatrix}$$

$S_1$ :

The sheep and **goat**  
are grazing.

$$\text{Scaled Scores}(S_2) = \begin{bmatrix} 12.5 \\ 5 \\ 12.5 \\ 0 \\ 10 \end{bmatrix} / \sqrt{2} = \begin{bmatrix} 8.85 \\ 3.55 \\ 8.85 \\ 0 \\ 7.08 \end{bmatrix}$$

$S_2$ :

Messi is the **goat**,  
not Ronaldo.

Word	Q,K,V
the	[1,1]
sheep	[4,0]
and	[0,0]
<b>goat</b>	<b>[2.5,2.5]</b>
grazing	[5,0]
Messi	[0,5]
is	[1,1]
not	[0,0]
Ronaldo	[0,4]

# Attention: Softmax

We convert the scaled scores into probabilities using the **Softmax**.

$$\text{Attention Score} = \text{softmax} \left( \frac{Q_{goat} \cdot K^T}{\sqrt{d_k}} \right)$$

$$\text{Attention Scores}(S_1) = \text{softmax} \left( \begin{bmatrix} 3.55 \\ 7.08 \\ 0 \\ 8.85 \\ 8.85 \end{bmatrix} \right) = \begin{bmatrix} 0.00 \\ 0.08 \\ 0.00 \\ 0.46 \\ 0.46 \end{bmatrix}$$

$S_1$ :

The sheep and **goat** are grazing.

$$\text{Attention Scores}(S_2) = \text{softmax} \left( \begin{bmatrix} 8.85 \\ 3.55 \\ 8.85 \\ 0 \\ 7.08 \end{bmatrix} \right) = \begin{bmatrix} 0.43 \\ 0 \\ 0.43 \\ 0 \\ 0.14 \end{bmatrix}$$

$S_2$ :

Messi is the **goat**, not Ronaldo.

Word	Q,K,V
the	[1,1]
sheep	[4,0]
and	[0,0]
<b>goat</b>	<b>[2.5,2.5]</b>
grazing	[5,0]
Messi	[0,5]
is	[1,1]
not	[0,0]
Ronaldo	[0,4]

# Attention: Weighted Sum

The output for **goat** is the weighted sum of the **Value (V)** vectors.

$$Attention(Q, K, V) = \sum softmax\left(\frac{Q_{goat} \cdot K^T}{\sqrt{d_k}}\right) \cdot V$$

$$Attention(S_1) = \sum \begin{bmatrix} 0.00 \\ 0.08 \\ 0.00 \\ 0.46 \\ 0.46 \end{bmatrix} \cdot \begin{bmatrix} [1,1] \\ [4,0] \\ [0,0] \\ [2.5,2.5] \\ [5,0] \end{bmatrix} = \begin{bmatrix} 3.77 \\ 1.15 \end{bmatrix}$$

$S_1$ :

The sheep and **goat** are grazing.

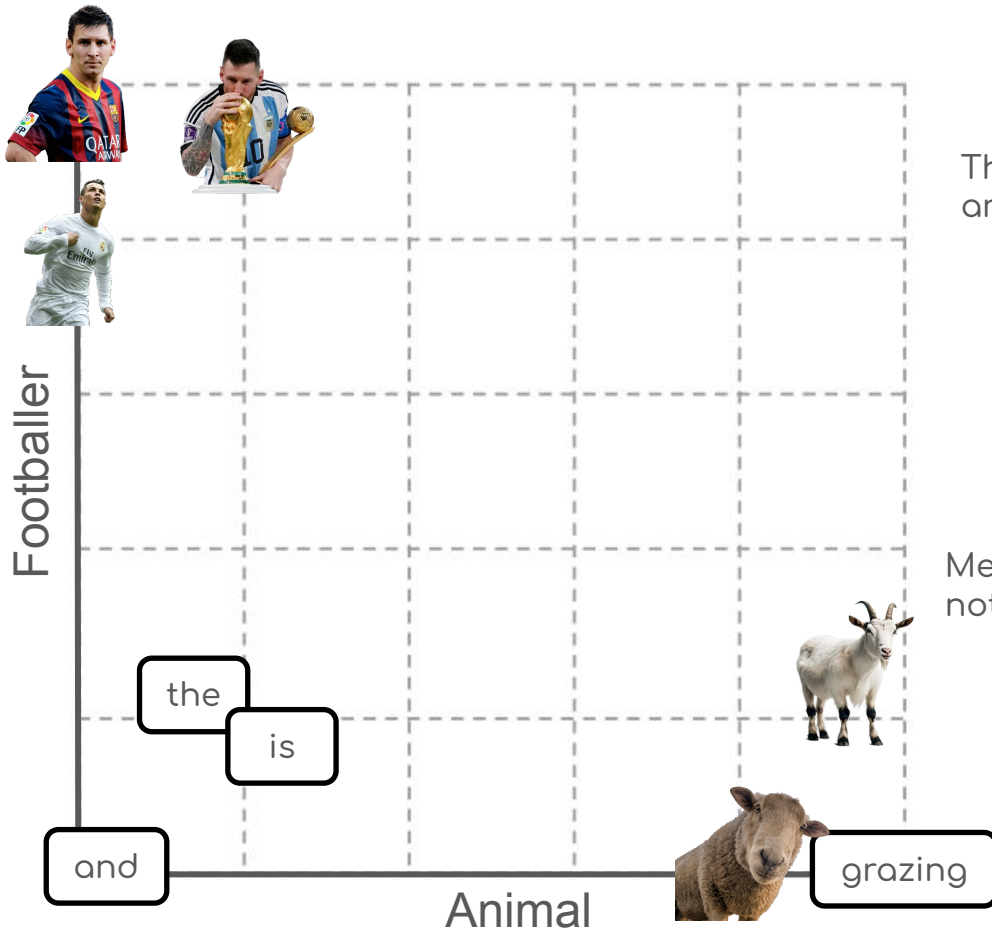
$$Attention(S_2) = \sum \begin{bmatrix} 0.43 \\ 0 \\ 0.43 \\ 0 \\ 0.14 \end{bmatrix} \cdot \begin{bmatrix} [0,5] \\ [1,1] \\ [2.5,2.5] \\ [0,0] \\ [0,4] \end{bmatrix} = \begin{bmatrix} 1.08 \\ 3.79 \end{bmatrix}$$

$S_2$ :

Messi is the **goat**, not Ronaldo.

Word	Q,K,V
the	[1,1]
sheep	[4,0]
and	[0,0]
<b>goat</b>	<b>[2.5,2.5]</b>
grazing	[5,0]
Messi	[0,5]
is	[1,1]
not	[0,0]
Ronaldo	[0,4]

# What should “goat” be?



$S_1$ :

The sheep and **goat** are grazing.

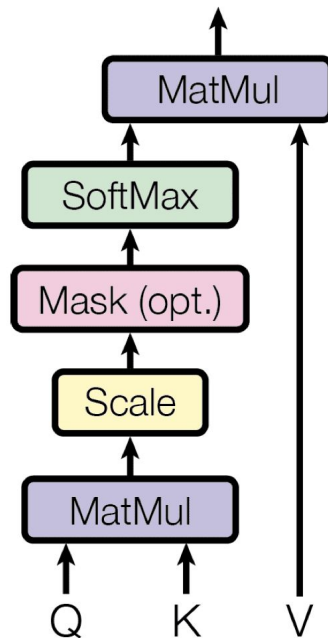
$S_2$ :

Messi is the **goat**, not Ronaldo.

Word	Q,K,V
the	[1,1]
sheep	[4,0]
and	[0,0]
goat ( $S_1$ )	[3.77, 1.15]
goat ( $S_2$ )	[1.08, 3.79]
grazing	[5,0]
Messi	[0,5]
is	[1,1]
not	[0,0]
Ronaldo	[0,4]

# Scaled Dot-Product Attention

- Query (Q)
  - The current word/token's focus or what it's "asking for."
  - A **vector representing the intent of the query word.**
- Key (K)
  - The **identity** or "role" of each word in the context of the sentence.
  - A matrix where each row corresponds to a word and holds its **contextual representation.**
- Value (V)
  - The **information or content** associated with each word.
  - A matrix where each row corresponds to a word's **content to aggregate if it's attended to.**





# Attention: Single Query

1. Similarity score with key ( $k_j$ ):

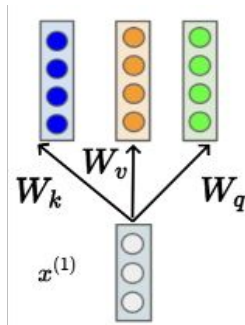
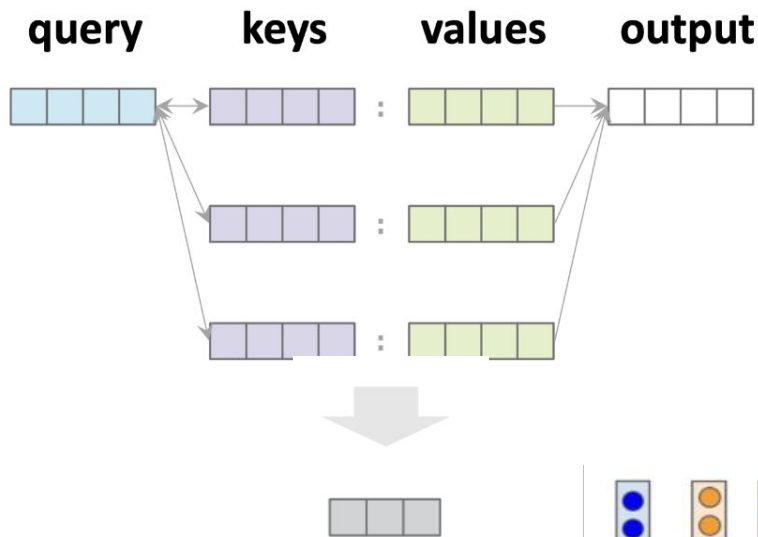
$$s_j = qk_j^T / \sqrt{d_k}$$

2. Attention weights:

$$a = \text{softmax}(qk_j^T / \sqrt{d_k})$$

3. Output: Attention-weighted sum:

$$y = \sum_j a_j v_i$$



# Attention: Multiple Query

1. Similarity score with query ( $q_i$ ) and key ( $k_j$ ):

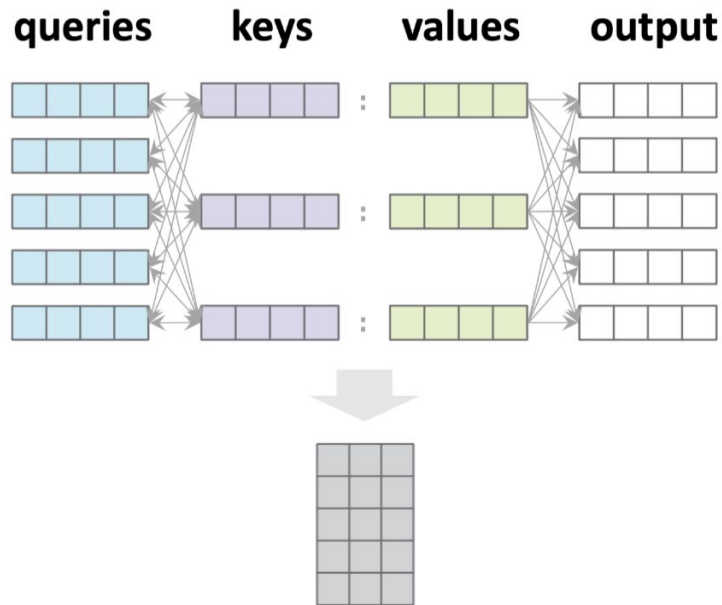
$$s_{ij} = q_i k_j^T / \sqrt{d_k}$$

2. Attention weights:

$$a_i = \text{softmax}(q_i k_j^T / \sqrt{d_k})$$

3. Output: Attention-weighted sum:

$$y_i = \sum_j a_{ij} v_i$$



# Self-Attention

Let us take following sentence with new embeddings:  
We will use all words as Q, K, and V

Messi is the goat, not Ronaldo.

$$a_i = \text{softmax}(q_i k_j^T / \sqrt{d_k})$$

Word	Q,K,V
Messi	[6,2]
is	[1,1]
the	[1,1]
goat	[5,3]
,	[0,0]
not	[1,0]
Ronaldo	[4,2]

# Self-Attention

Messi is the goat, not Ronaldo.

Step 1:  $s_{ij} = q_i k_j^T / \sqrt{d_k}$

	Messi	is	the	goat	,	not	Ronaldo
Messi	28.37	5.67	5.67	25.53	0	4.25	19.86

$$s_{messi,messi} = q_i k_j^T / \sqrt{d_k} = (6 \times 6 + 2 \times 2) / \sqrt{2} = 28.37$$

Word	Q,K,V
Messi	[6,2]
is	[1,1]
the	[1,1]
goat	[5,3]
,	[0,0]
not	[1,0]
Ronaldo	[4,2]

# Self-Attention

Messi is the goat, not Ronaldo.

Step 2:  $a_i = \text{softmax}(s_{ij})$

	Messi	is	the	goat	,	not	Ronaldo
Messi	0.94	0.0	0.0	0.06	0	0.0	0.0

$$a_{messi} = e^{28.37} / (e^{28.37} + e^{5.67} + e^{5.67} + e^{25.53} + e^0 + e^{4.25} + e^{19.85}) \approx 0.94$$

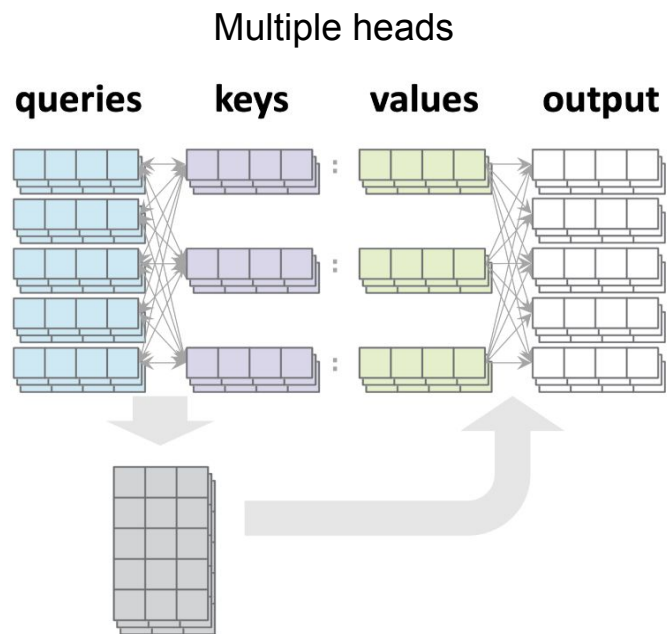
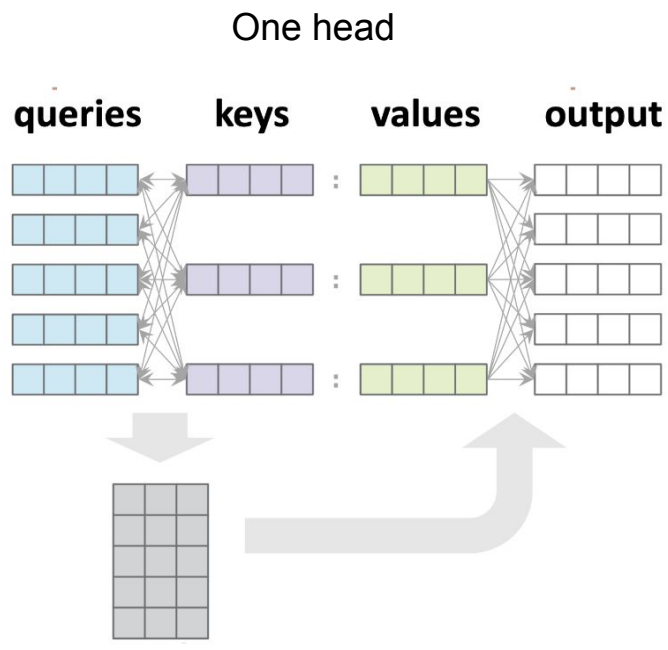
Word	$s_{ij}$
Messi	28.37
is	5.67
the	5.67
goat	25.53
,	0.00
not	4.25
Ronaldo	19.86

# Self-Attention

Messi is the goat, not Ronaldo.

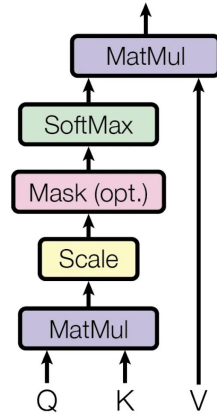
	Messi	is	the	goat	,	not	Ronaldo
Messi	0.94	0.0	0.0	0.06	0.0	0.0	0.0
is	0.44	0.01	0.01	0.44	0.0	0.0	0.11
the	0.44	0.01	0.01	0.44	0.0	0.0	0.11
goat	0.8	0.0	0.0	0.2	0.0	0.0	0.0
,	0.14	0.14	0.14	0.14	0.14	0.14	0.14
not	0.54	0.02	0.02	0.27	0.01	0.02	0.13
Ronaldo	0.0	0.0	0.0	0.2	0.0	0.0	0.8

# Multi-head Attention



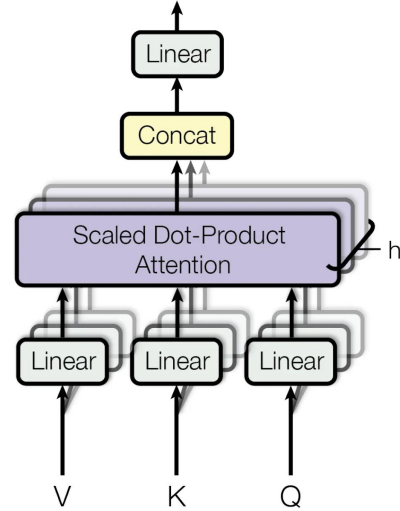
# Multi-head Attention

Scaled Dot-Product Attention



$$\text{Attention}(Q, K, V) = \sum \text{softmax}\left(\frac{Q_{\text{goat}} \cdot K^T}{\sqrt{d_k}}\right) \cdot V$$

Multi-Head Attention





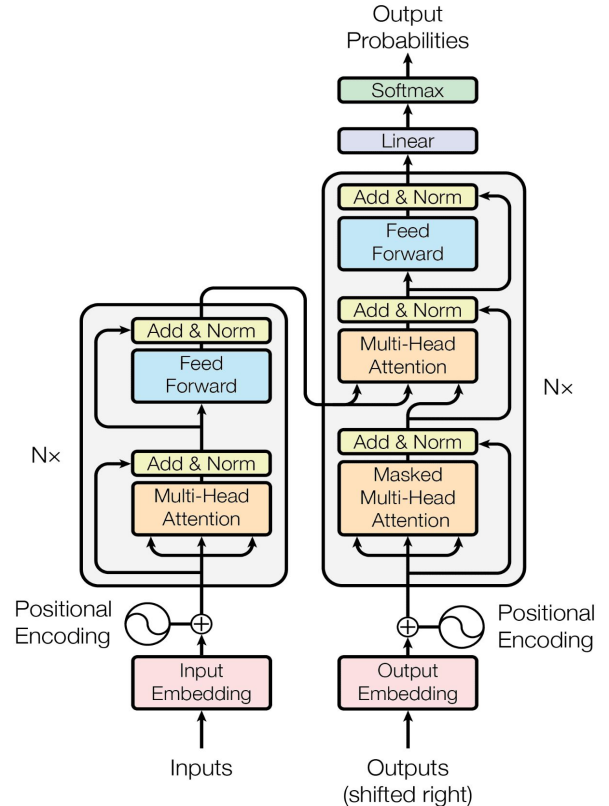
# Using Attention: Transformers

Messi is the goat, not Ronaldo.

Messi est la chèvre, pas Ronaldo.

メッシはゴートで、ロナウドではありません。

梅西是山羊，不是罗纳尔多。



# Vision Transformers

Published as a conference paper at ICLR 2021

## AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy<sup>\*†</sup>, Lucas Beyer<sup>\*</sup>, Alexander Kolesnikov<sup>\*</sup>, Dirk Weissenborn<sup>\*</sup>,  
Xiaohua Zhai<sup>\*</sup>, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer,  
Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby<sup>\*†</sup>

<sup>\*</sup>equal technical contribution, <sup>†</sup>equal advising  
Google Research, Brain Team  
{adosovitskiy, neilhoulby}@google.com

### ABSTRACT

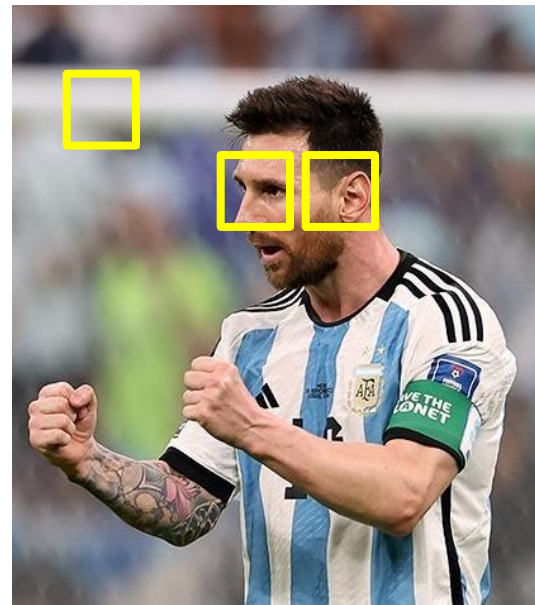
While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.<sup>1</sup>

### 1 INTRODUCTION

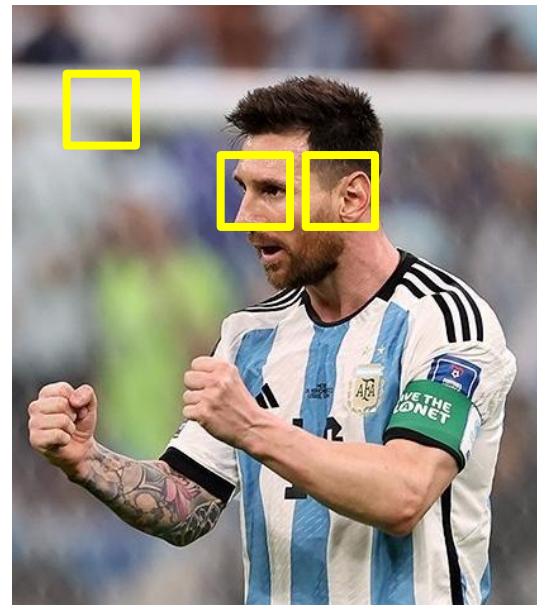
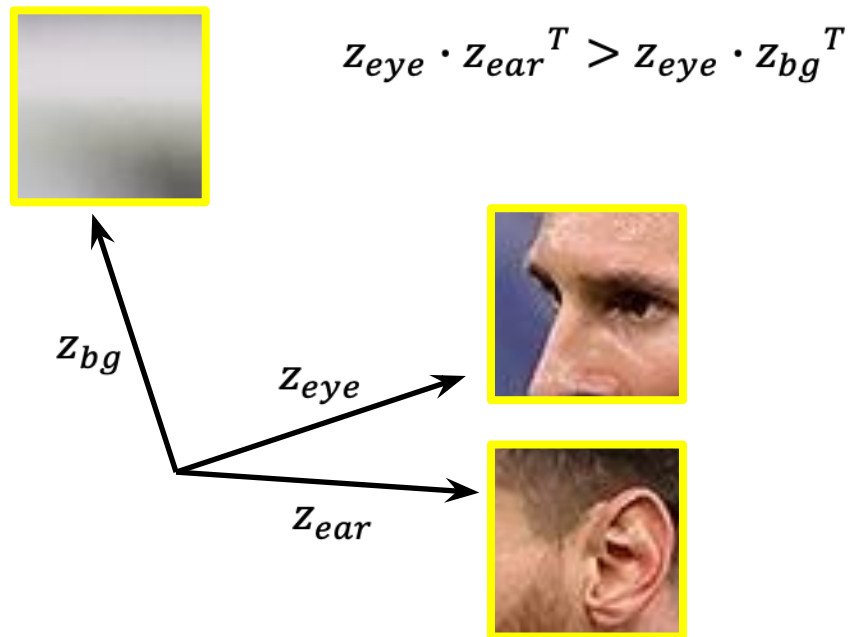
Self-attention-based architectures, in particular Transformers (Vaswani et al., 2017), have become the model of choice in natural language processing (NLP). The dominant approach is to pre-train on a large text corpus and then fine-tune on a smaller task-specific dataset (Devlin et al., 2019). Thanks to Transformers' computational efficiency and scalability, it has become possible to train models of unprecedented size, with over 100B parameters (Brown et al., 2020; Lepikhin et al., 2020). With the models and datasets growing, there is still no sign of saturating performance.

2010.11929v2 [cs.CV] 3 Jun 2021

# Visual Attention



# Visual Attention: Dot-product Similarity



# Image to Patches

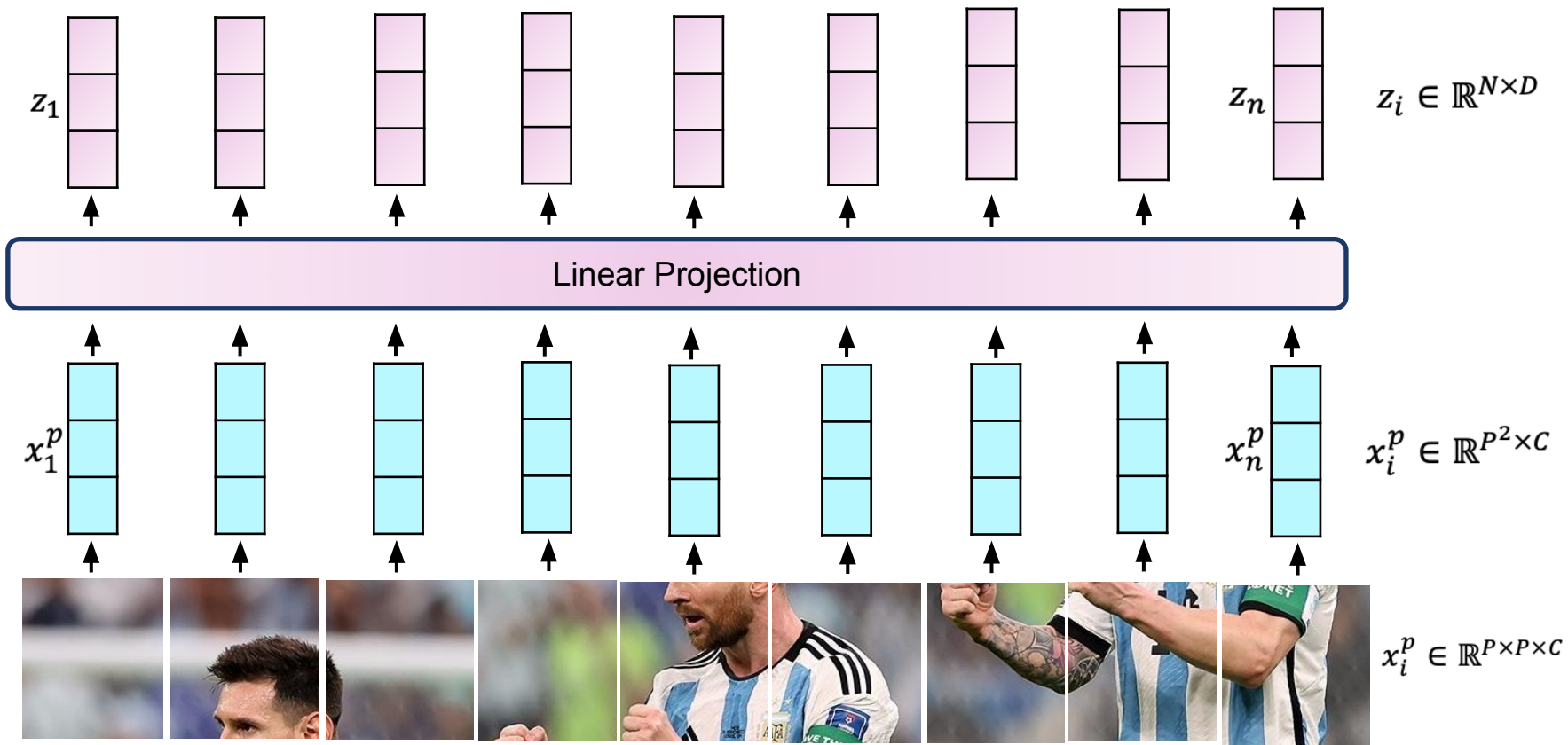


$$x \in \mathbb{R}^{W \times H \times C}$$

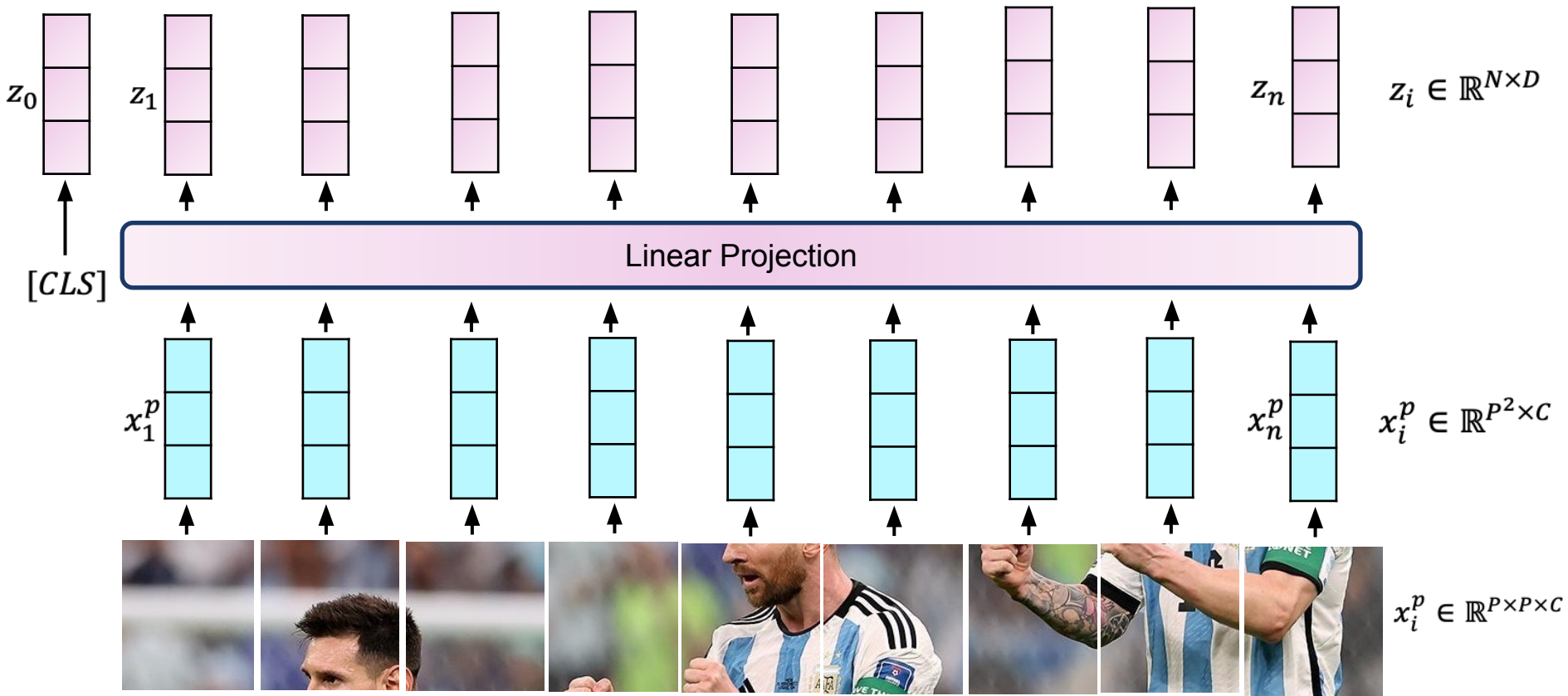


$$x_i^p \in \mathbb{R}^{P \times P \times C}, i = 1 \dots N$$

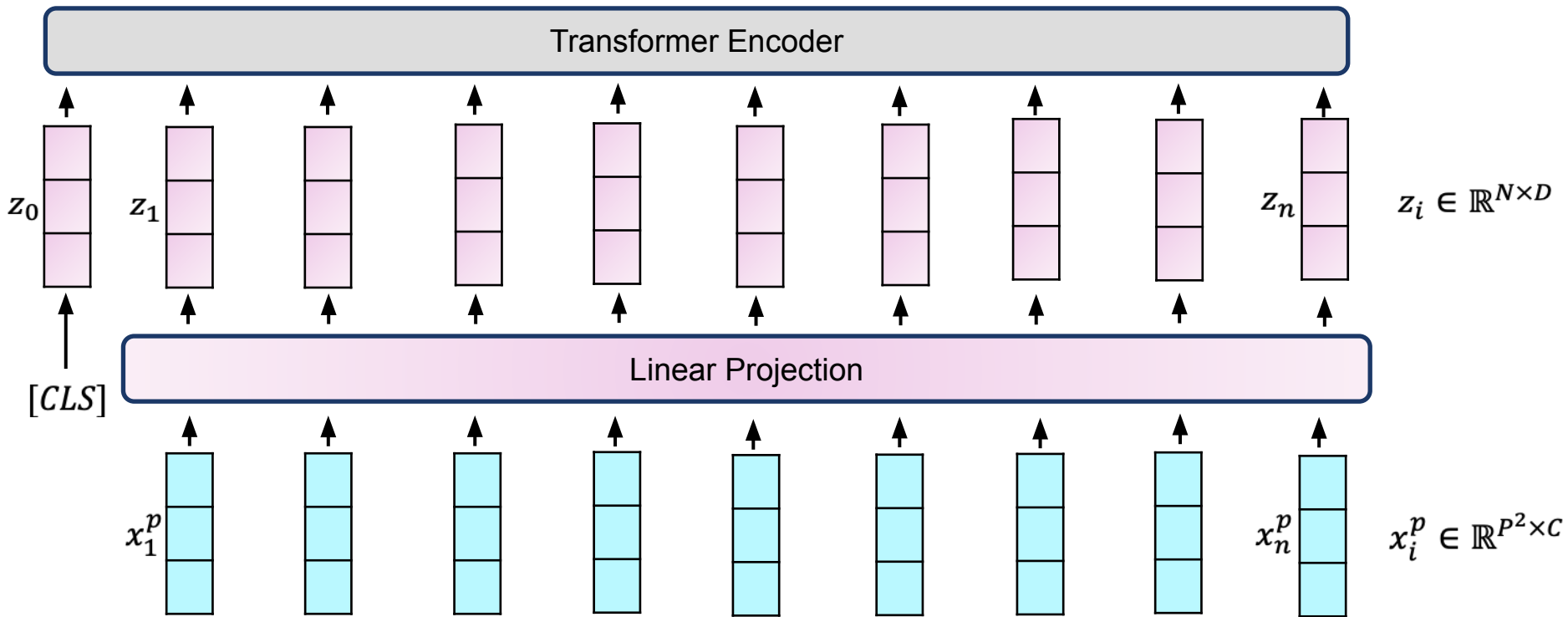
# Linear Projection



# Positional Encoding

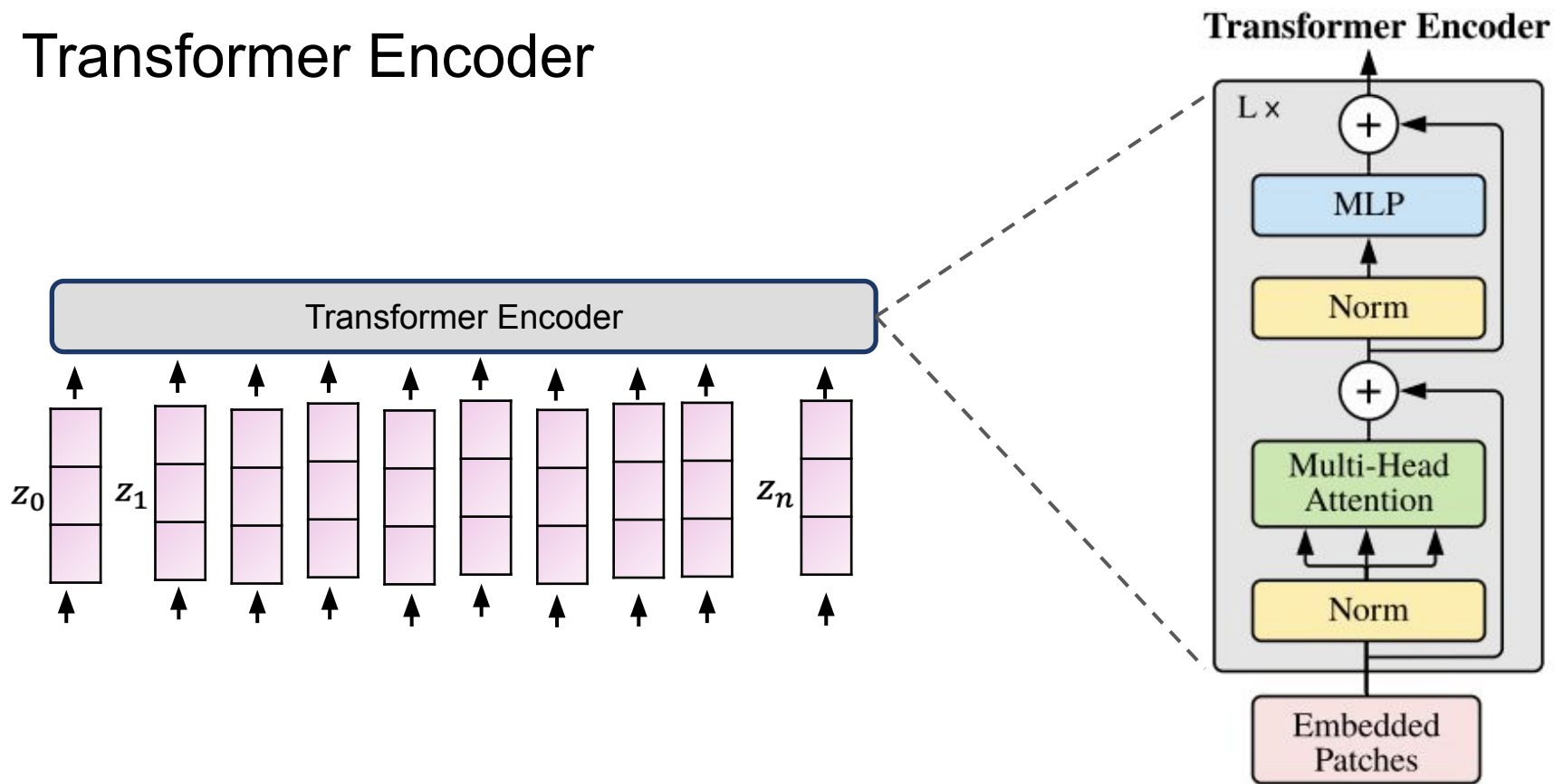


# Transformer Encoder

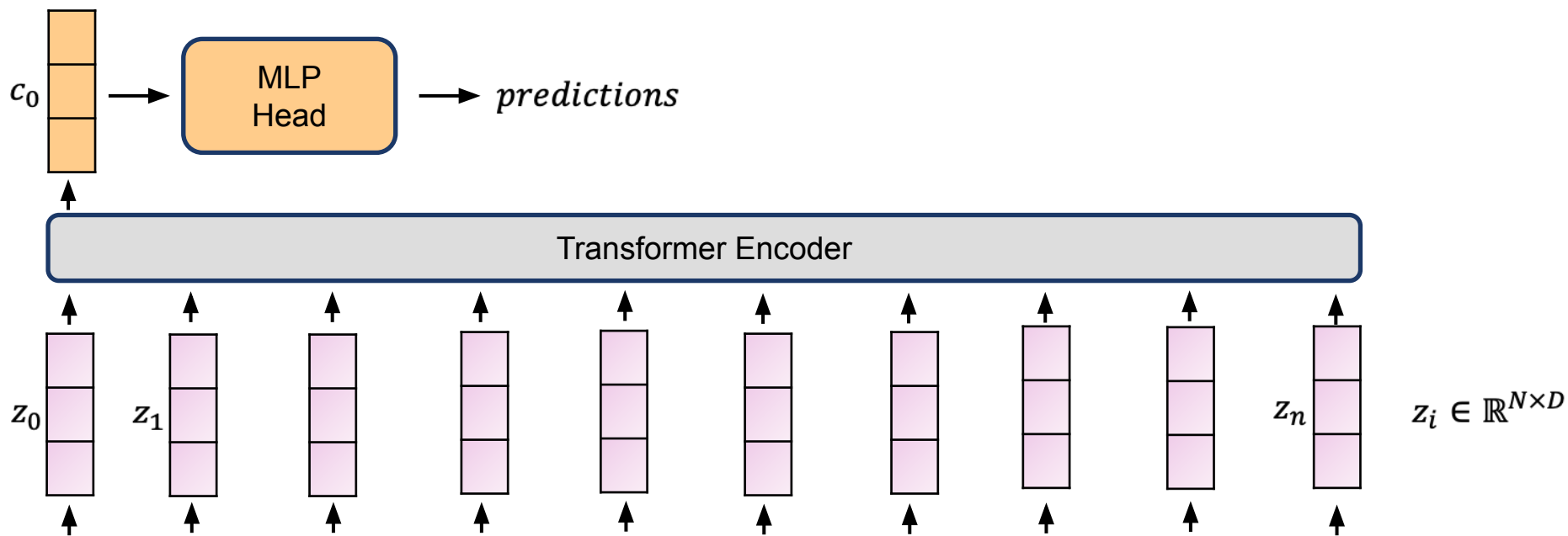




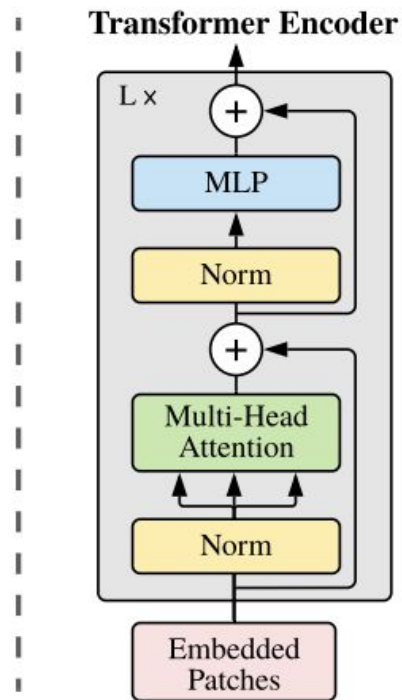
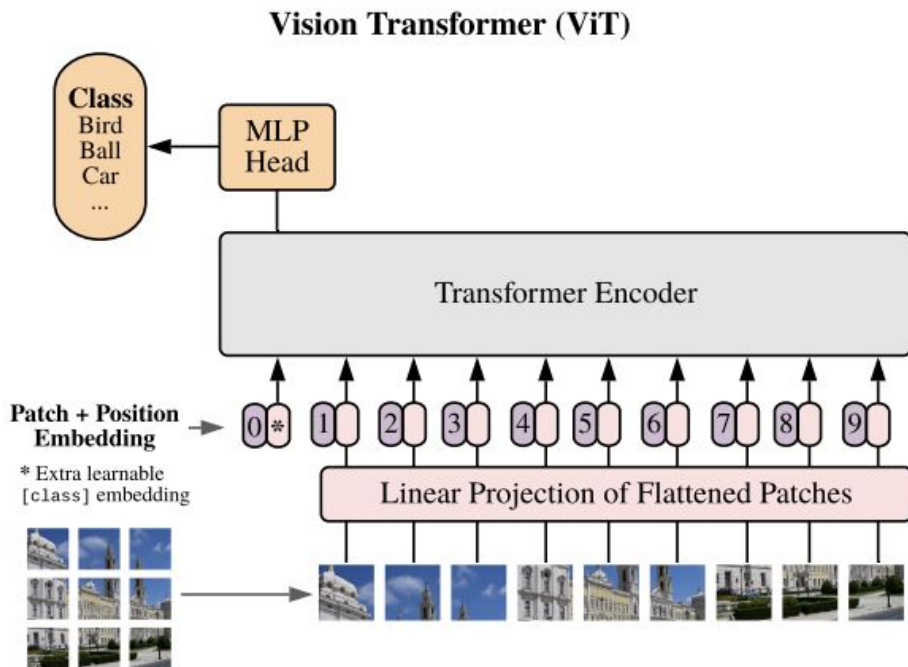
# Transformer Encoder



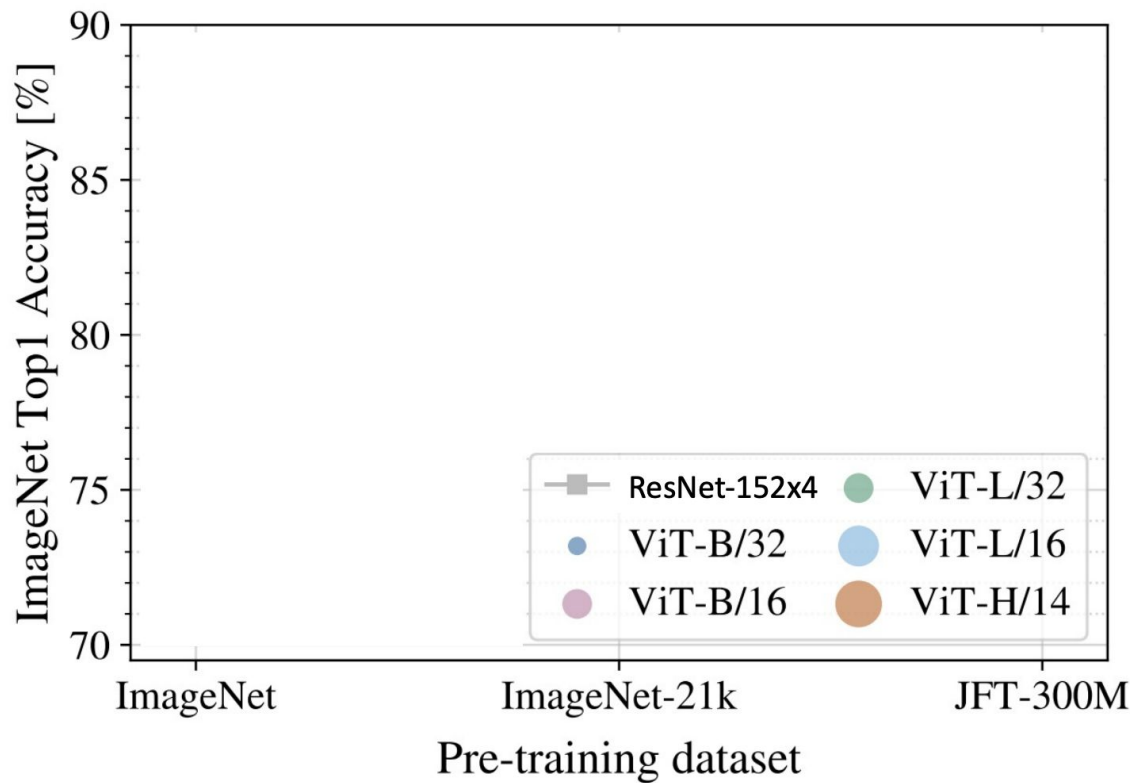
# Transformer Encoder



# Vision Transformer (ViT)



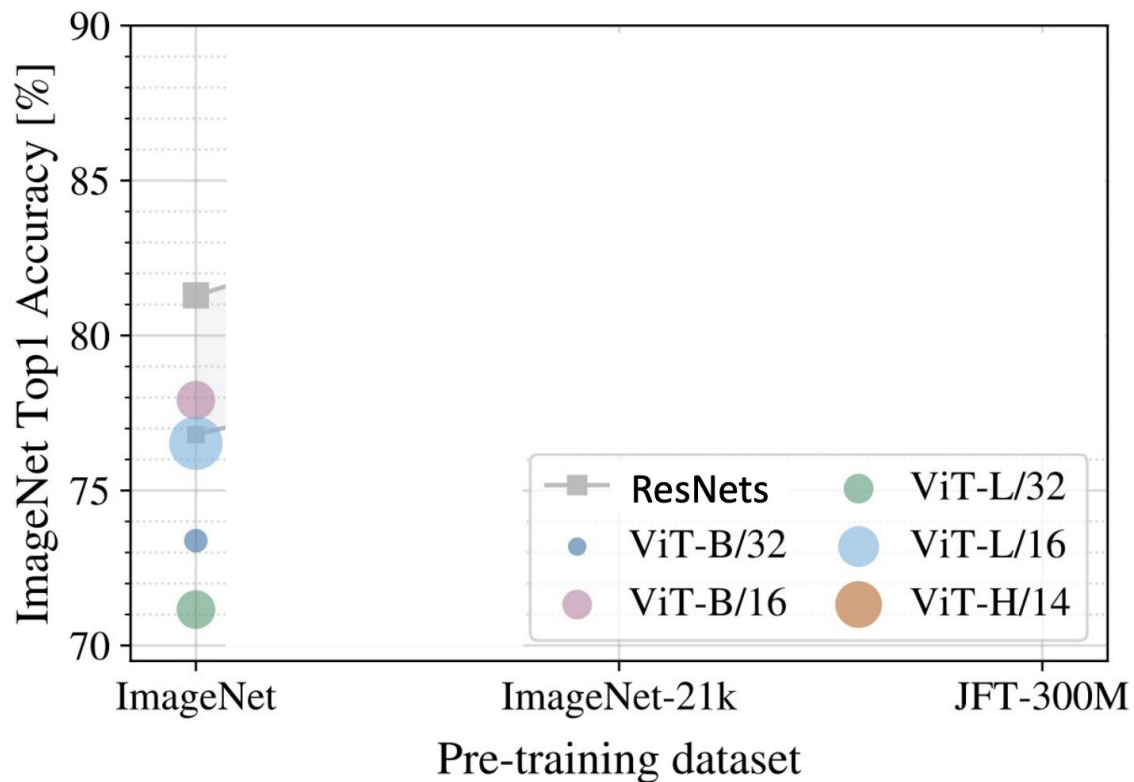
# ViTs vs ResNet



# ViTs vs ResNet

ImageNet:

- 1k Classes
- 1.2M images



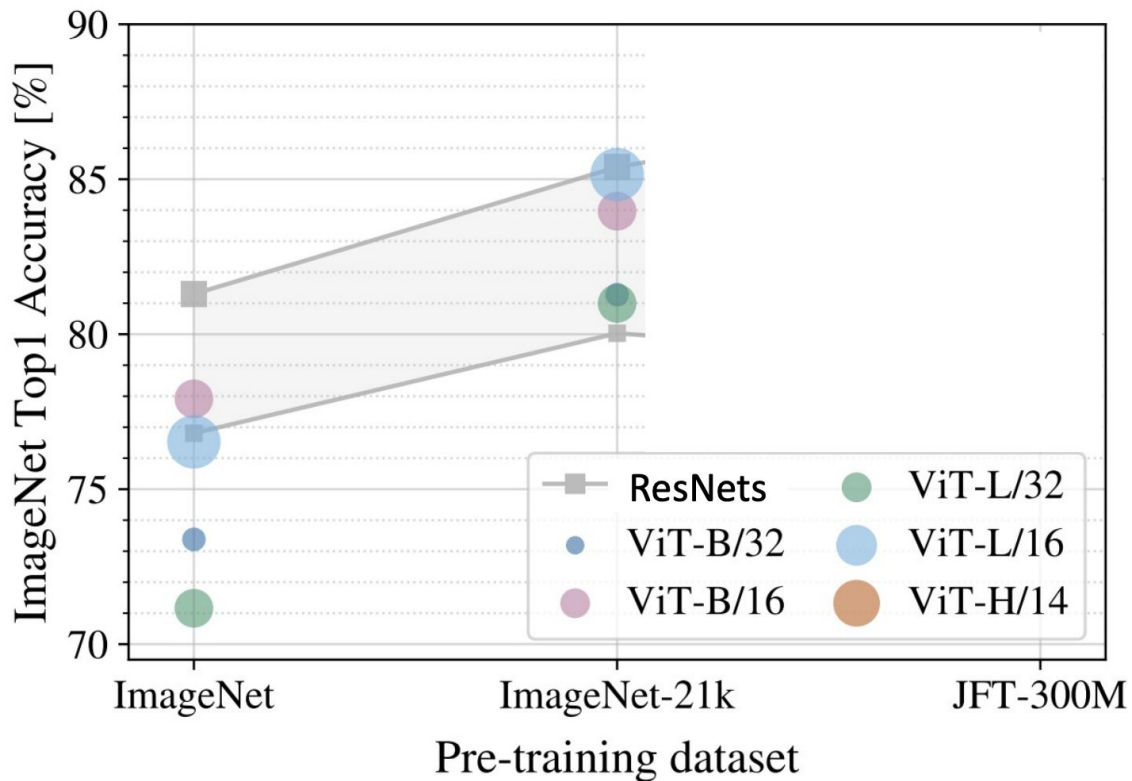
# ViTs vs ResNet

ImageNet:

- 1k Classes
- 1.2M images

ImageNet-21k

- 21k classes
- 14M images



# ViTs vs ResNet

ImageNet:

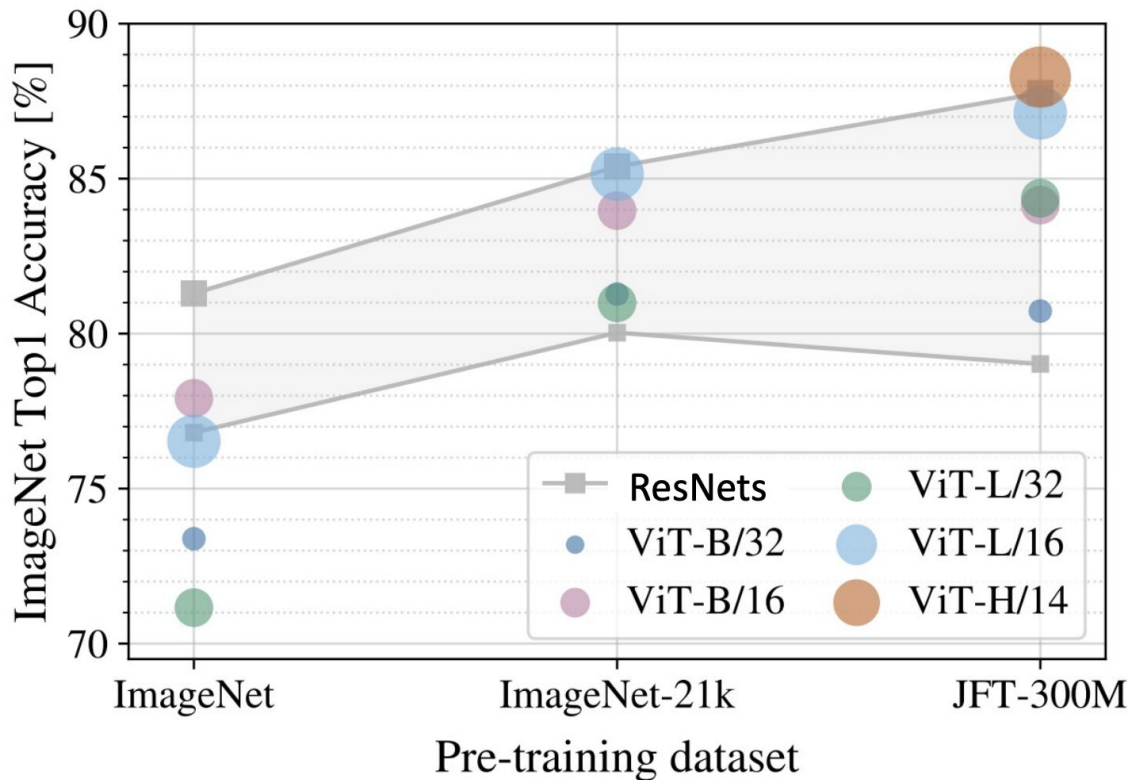
- 1k Classes
- 1.2M images

ImageNet-21k

- 21k classes
- 14M images

JFT-300M

- 300M images



Thank you  
for your attention :)

