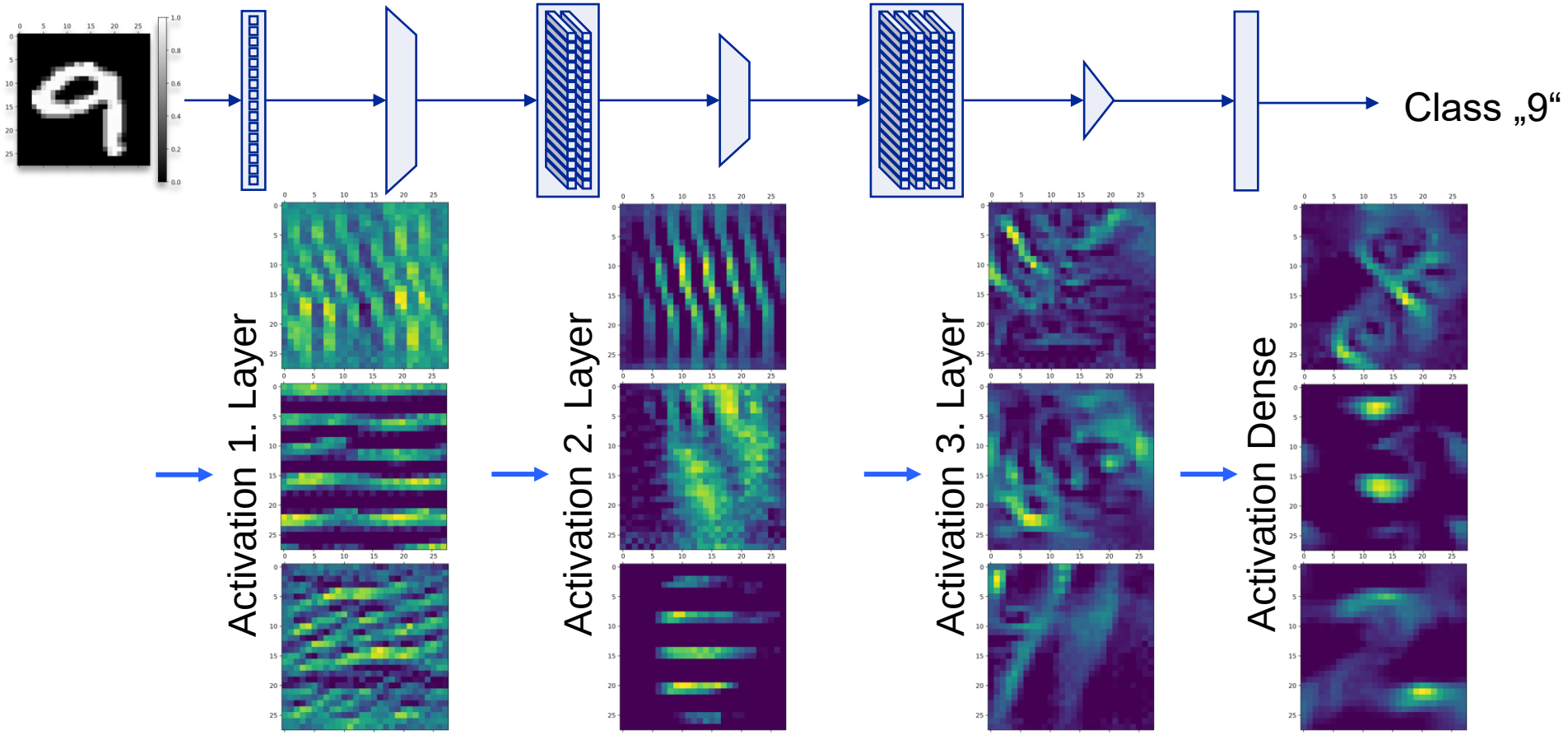
HEILBRONN UNIVERSITY
OF APPLIED SCIENCES

› VISION TRANSFORMERS: PREPROCESSING

Advanced Approaches for AI-Based Image Processing

Prof. Alexander Windberger - IT Faculty – AIB - SoSe 2024

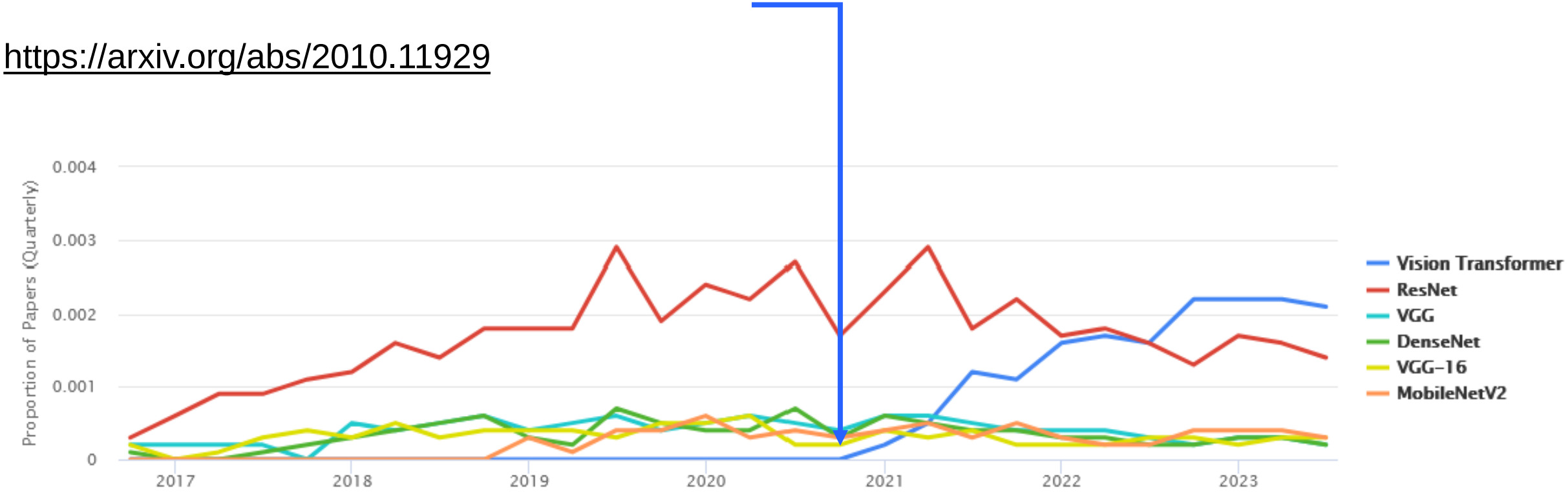
CNNS: HIERACHICAL PROCESSING



TRANSFORMERS: THE RISE OF THE BEASTS

An Image is Worth 16x16 Words: **Transformers** for Image Recognition at Scale

https://arxiv.org/abs/2010.11929



Numbers: https://www.cnbc.com/2023/03/13/chatgpt-and-generative-ai-are-booming-but-at-a-very-expensive-price.html (13.03.2023)
Graph: https://paperswithcode.com/method/vision-transformer (06.09.2023)

„ATTENTION IS ALL YOU NEED“

The giraffe doesn’t fit the suitcase because it’s too small.

The giraffe doesn’t fit the suitcase because it’s too big.

Attention Is All You Need, https://arxiv.org/abs/1706.03762

„ATTENTION IS ALL YOU NEED“

The giraffe doesn’t fit the suitcase because it’s too small.

The giraffe doesn’t fit the suitcase because it’s too big.

Attention Is All You Need, https://arxiv.org/abs/1706.03762

„ATTENTION IS ALL YOU NEED“

	giraffe …	suitcase …	it …	big
giraffe …				
suitcase …				
it …				
big				

Attention Is All You Need, https://arxiv.org/abs/1706.03762

„ATTENTION IS ALL YOU NEED“

Query

	giraffe …	suitcase …	it …	big
giraffe …	0.6	0.1	0.2	0.2
suitcase …				
it …				
big				

Attention Is All You Need, https://arxiv.org/abs/1706.03762

„ATTENTION IS ALL YOU NEED“

Query

	giraffe …	suitcase …	it …	big
giraffe …	0.6	0.1	0.2	0.2
suitcase …				
it …	0.4	0.0	0.5	0.1
big				

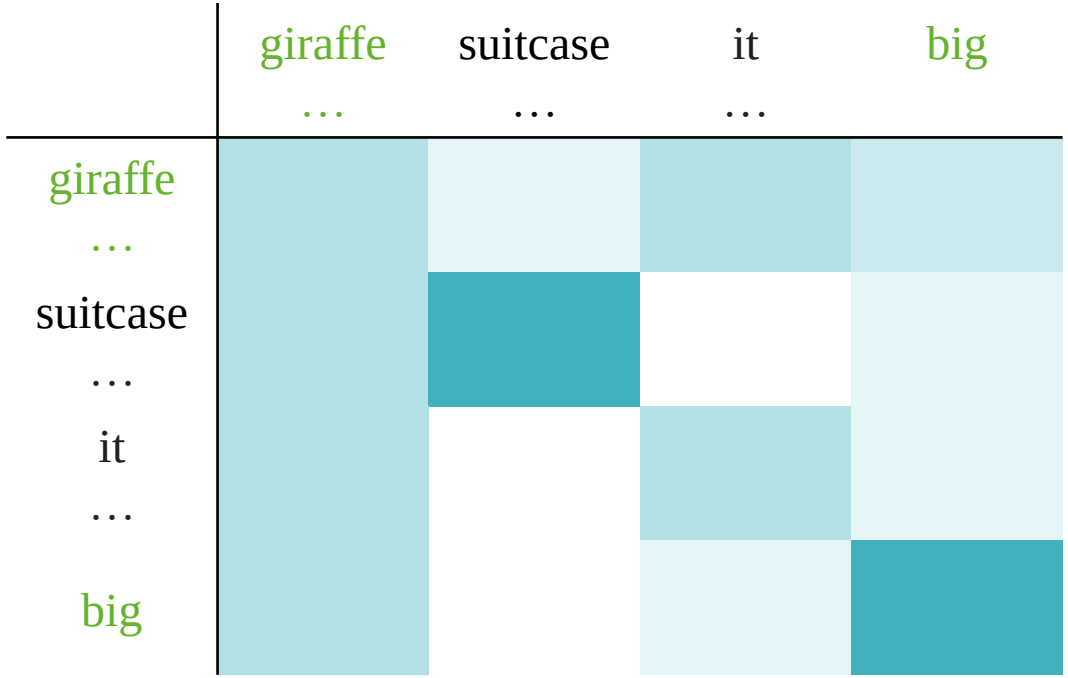
„ATTENTION IS ALL YOU NEED“

	giraffe …	suitcase …	it …	big
giraffe …	0.6	0.1	0.2	0.2
suitcase …	0.2	0.7	0.0	0.1
it …	0.4	0.0	0.5	0.1
big	0.3	0.0	0.1	0.6

Attention Is All You Need, https://arxiv.org/abs/1706.03762

„ATTENTION IS ALL YOU NEED“

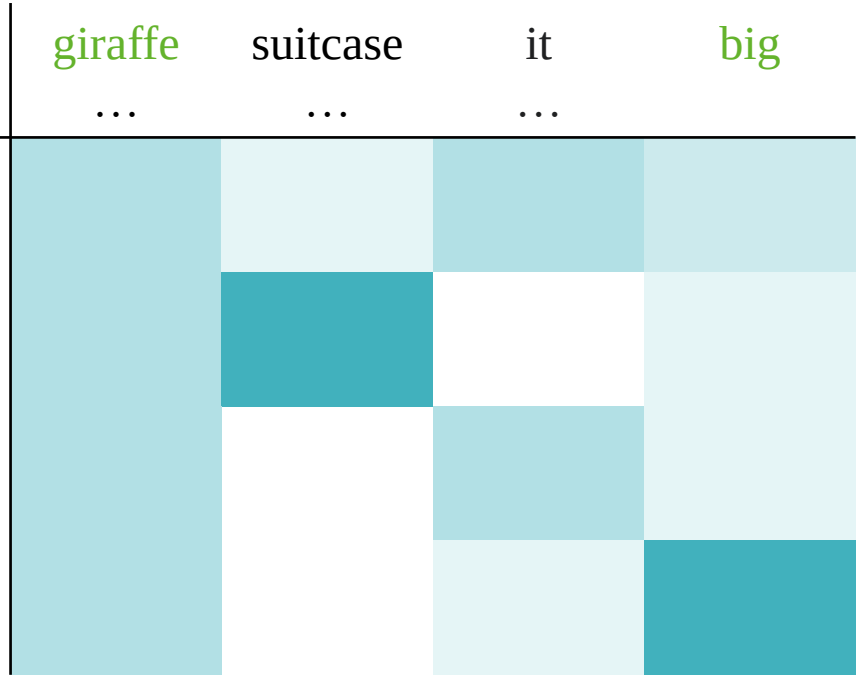
The giraffe doesn’t fit the suitcase because it’s too big.



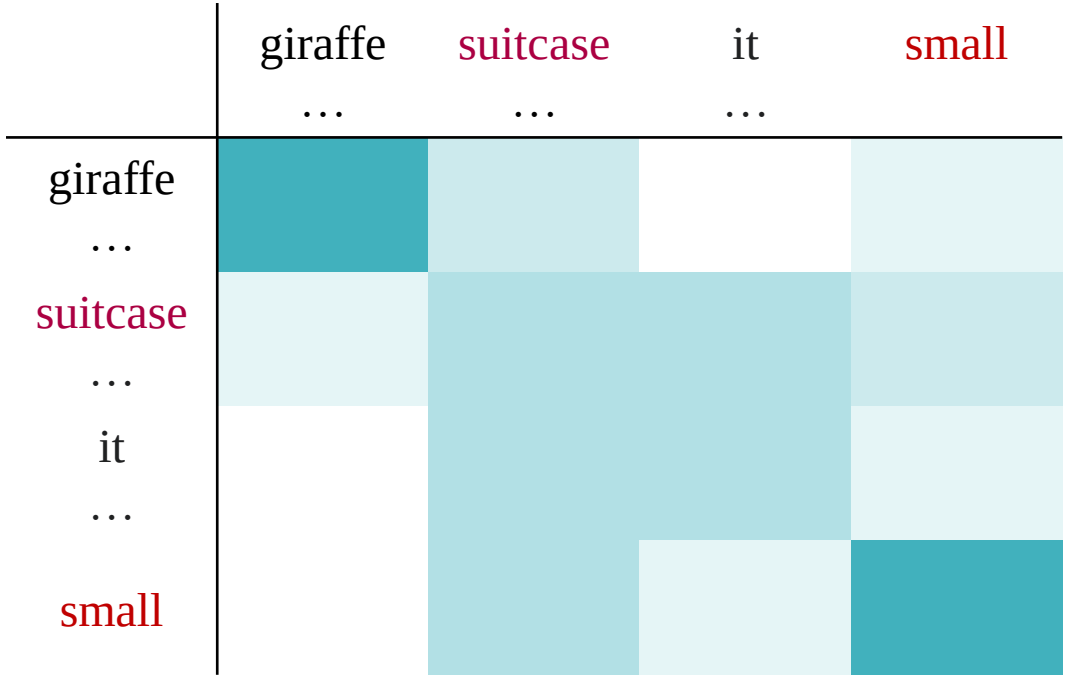
Attention Is All You Need, https://arxiv.org/abs/1706.03762

„ATTENTION IS ALL YOU NEED“

The giraffe doesn’t fit the suitcase because it’s too big.



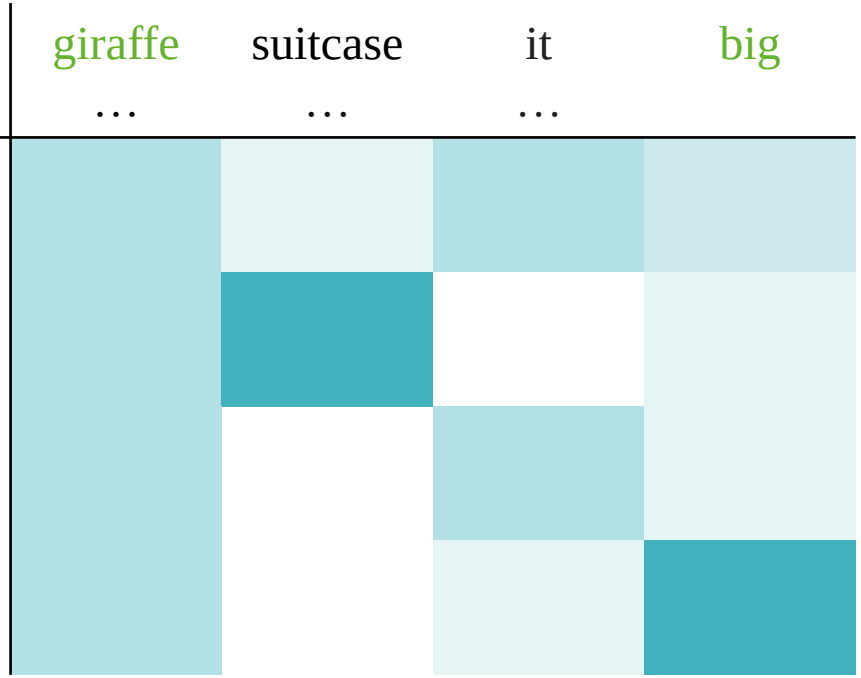
The giraffe doesn’t fit the suitcase because it’s too small.



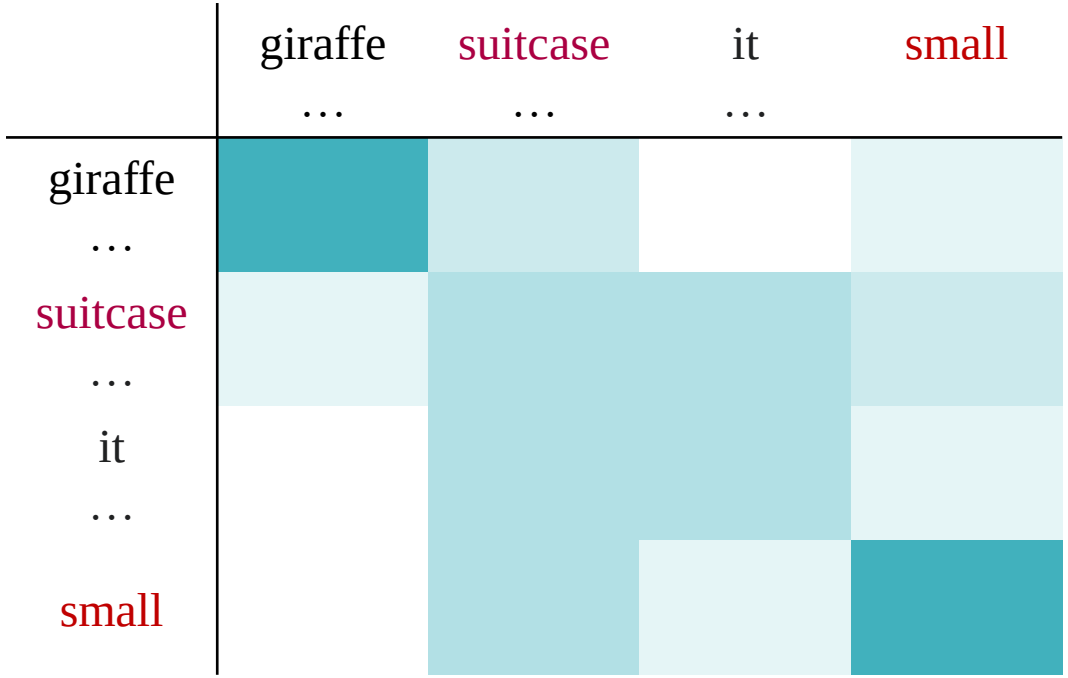
Attention Is All You Need, https://arxiv.org/abs/1706.03762

„ATTENTION IS ALL YOU NEED“

The giraffe doesn’t fit the suitcase because it’s too big.



The giraffe doesn’t fit the suitcase because it’s too small.

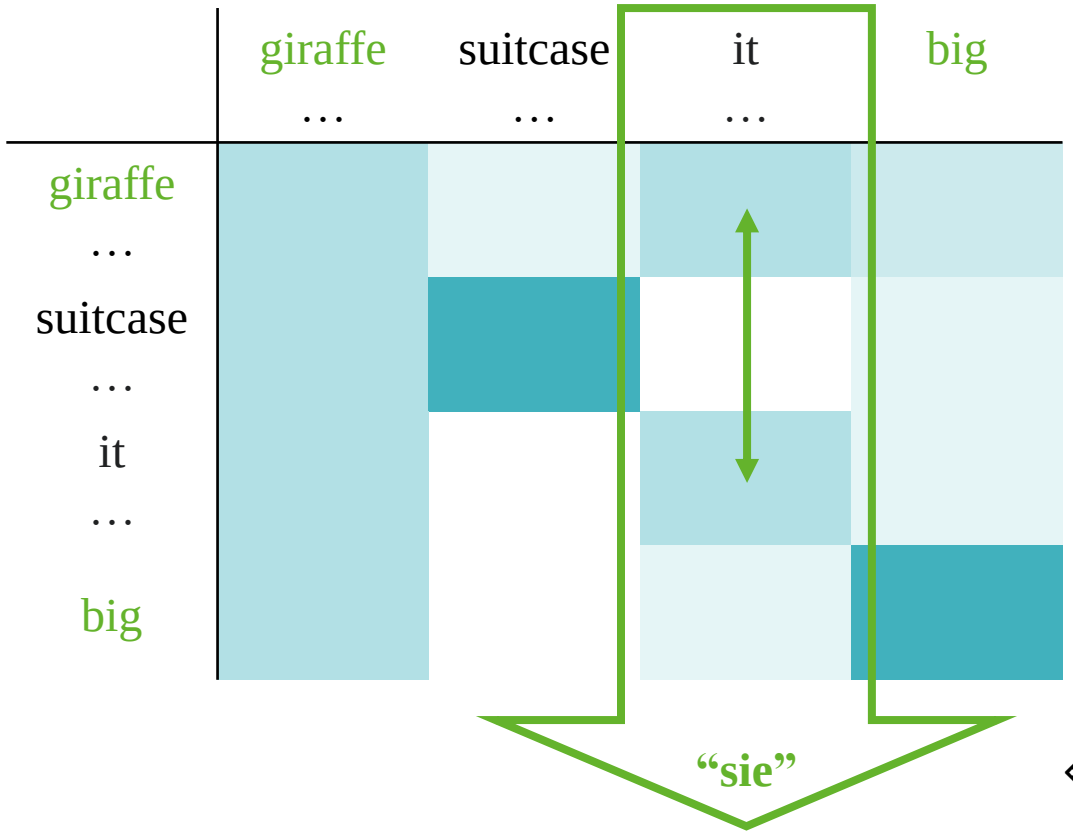


⇐ **“Translate to German”** ⇒

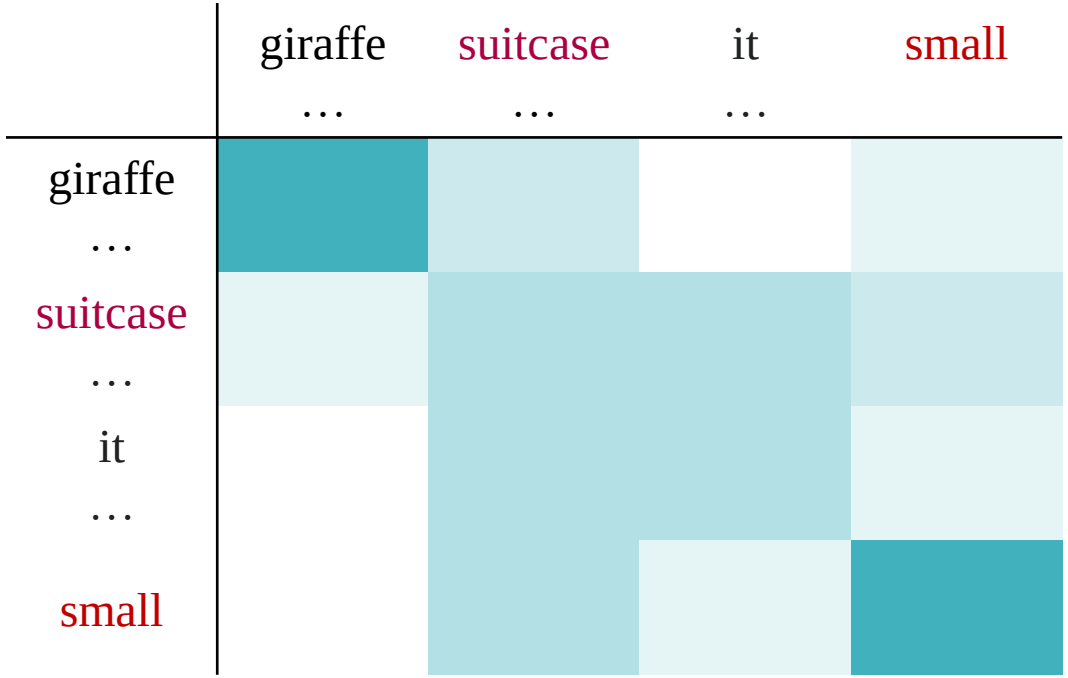
Attention Is All You Need, https://arxiv.org/abs/1706.03762

„ATTENTION IS ALL YOU NEED“

The giraffe doesn’t fit the suitcase because it’s too big.



The giraffe doesn’t fit the suitcase because it’s too small.

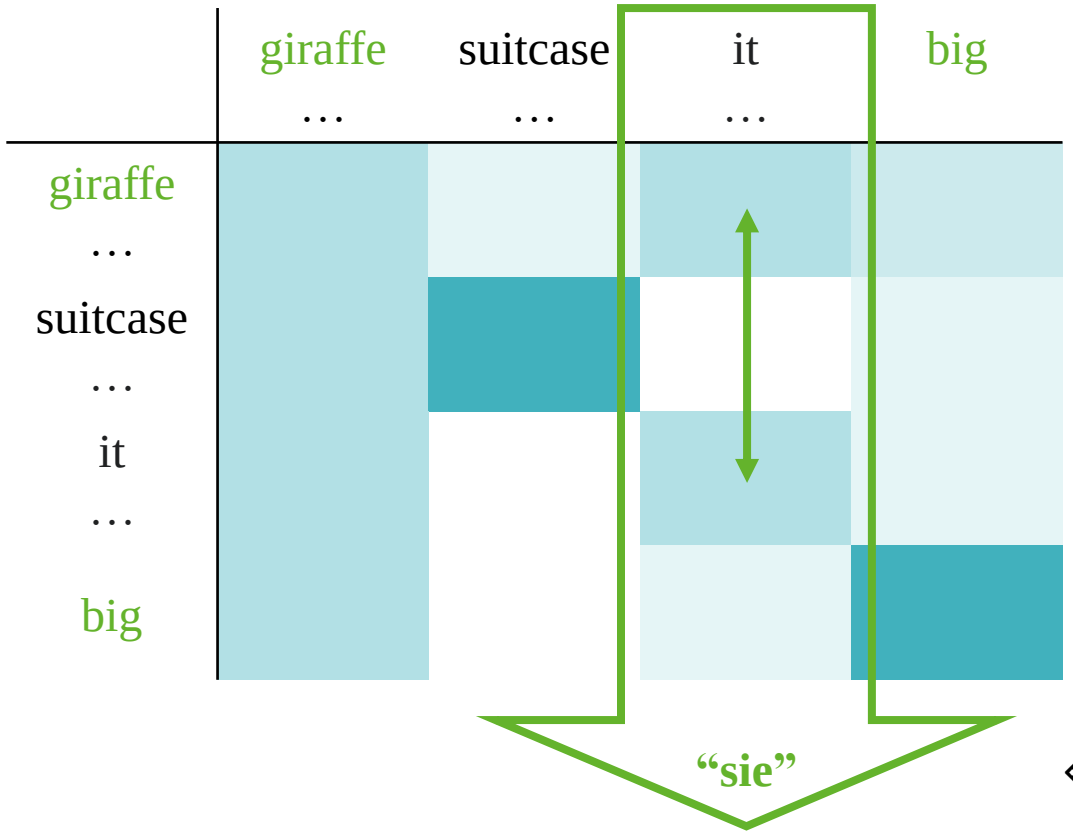


⇐ **“Translate to German”** ⇒

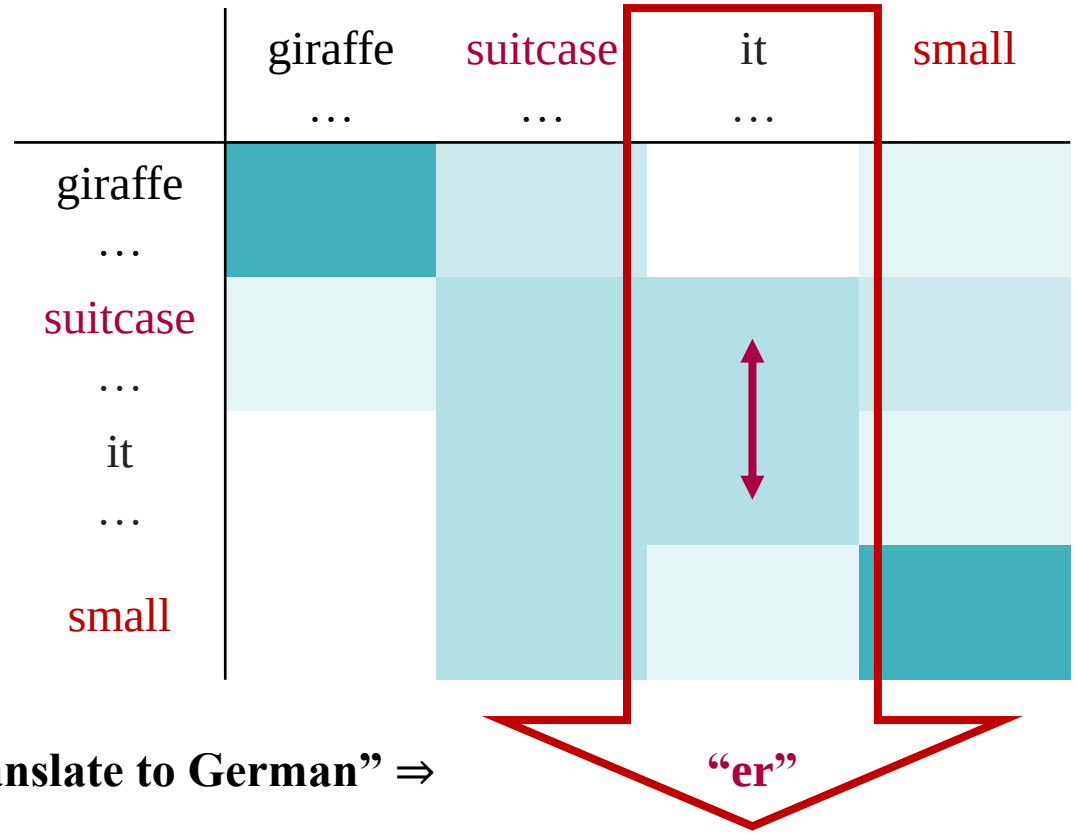
Attention Is All You Need, https://arxiv.org/abs/1706.03762

„ATTENTION IS ALL YOU NEED“

The giraffe doesn’t fit the suitcase because it’s too big.



The giraffe doesn’t fit the suitcase because it’s too small.



⇐ **“Translate to German”** ⇒

Attention Is All You Need, https://arxiv.org/abs/1706.03762

HEILBRONN UNIVERSITY
OF APPLIED SCIENCES

AN IMAGE IS WORTH 16X16 WORDS

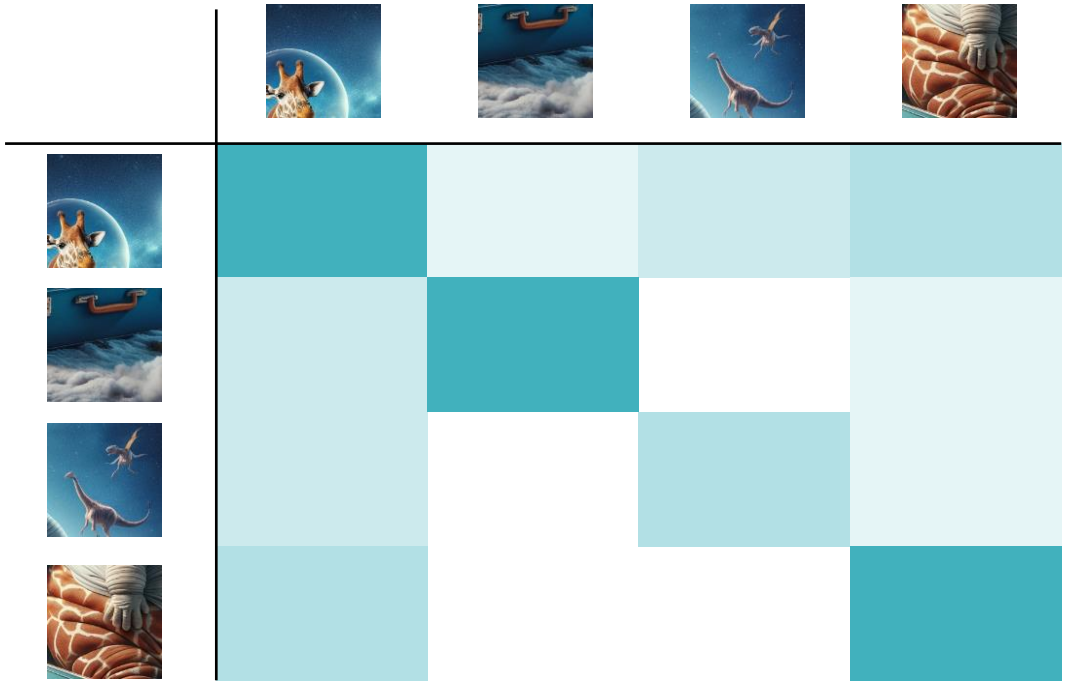
An Image is Worth 16x16 Words, https://arxiv.org/abs/2010.11929

HEILBRONN UNIVERSITY
OF APPLIED SCIENCES

AN IMAGE IS WORTH 16X16 WORDS

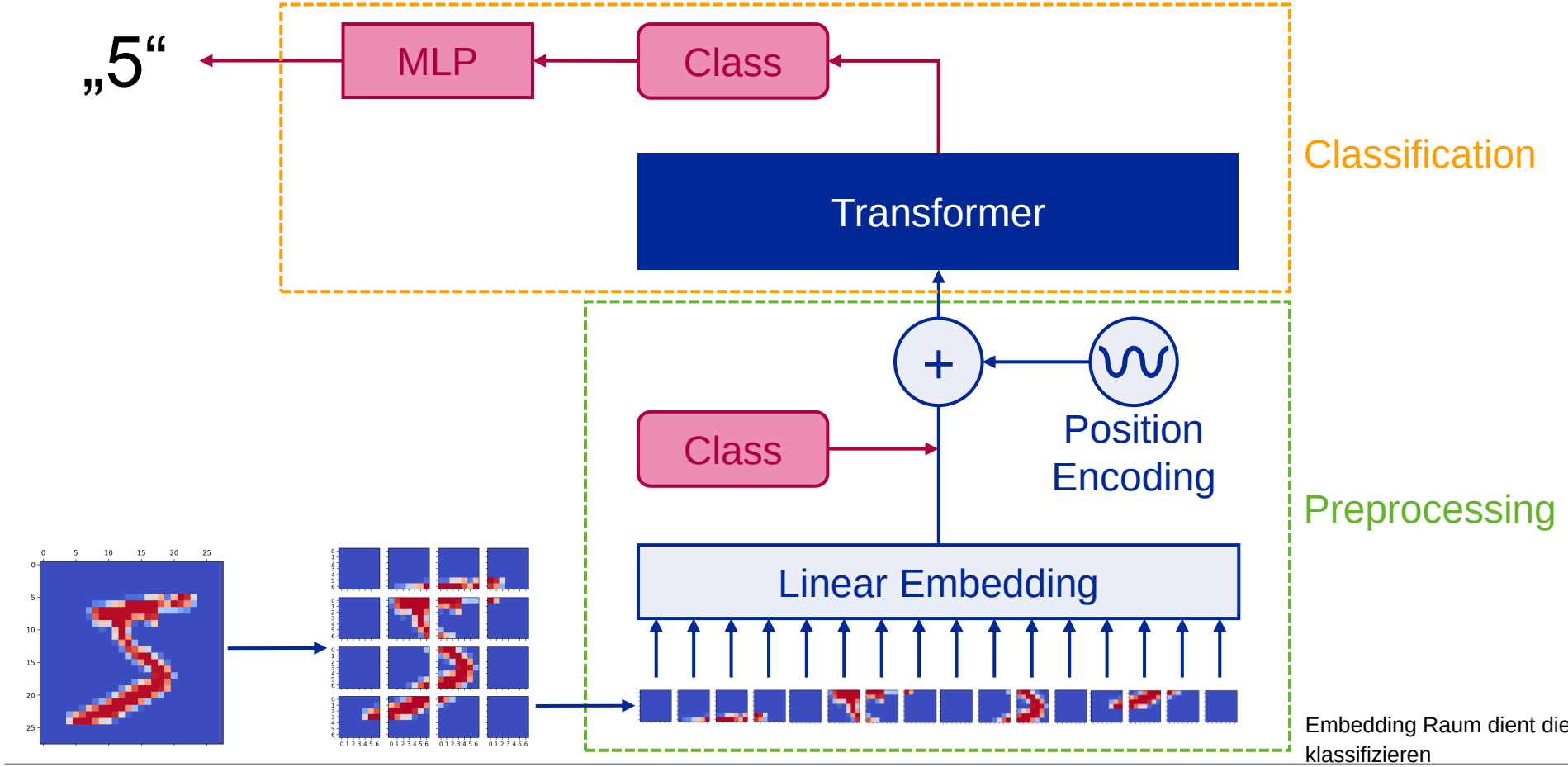
An Image is Worth 16x16 Words, https://arxiv.org/abs/2010.11929

AN IMAGE IS WORTH 16X16 WORDS



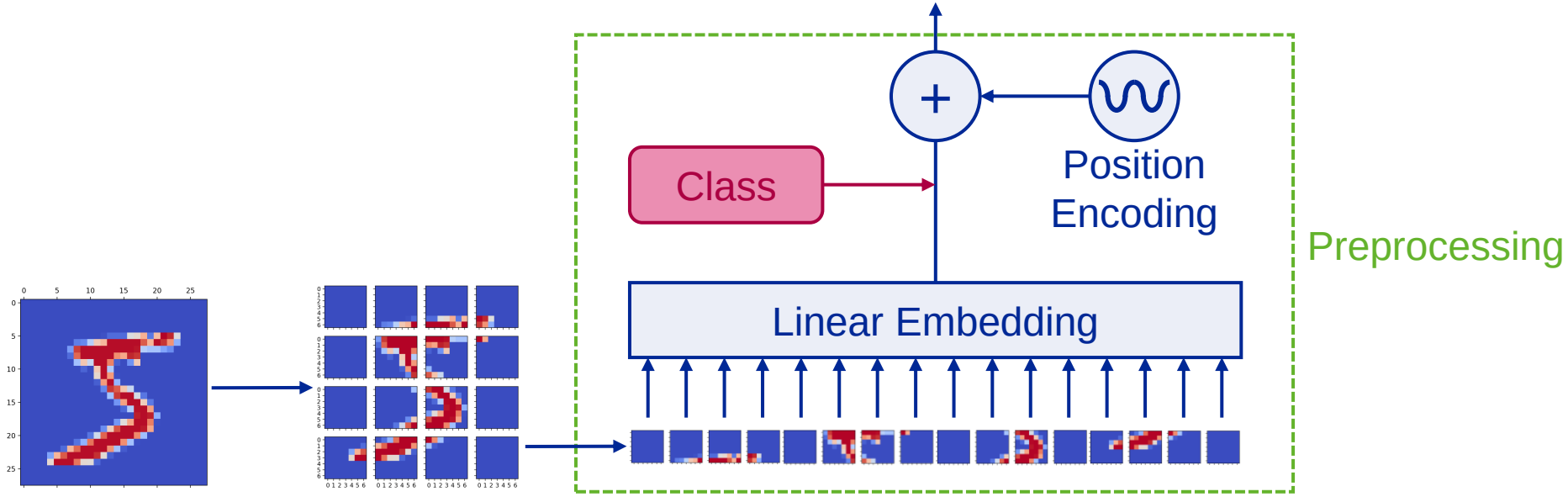
An Image is Worth 16x16 Words, https://arxiv.org/abs/2010.11929

VISION TRANSFORMER FROM SCRATCH



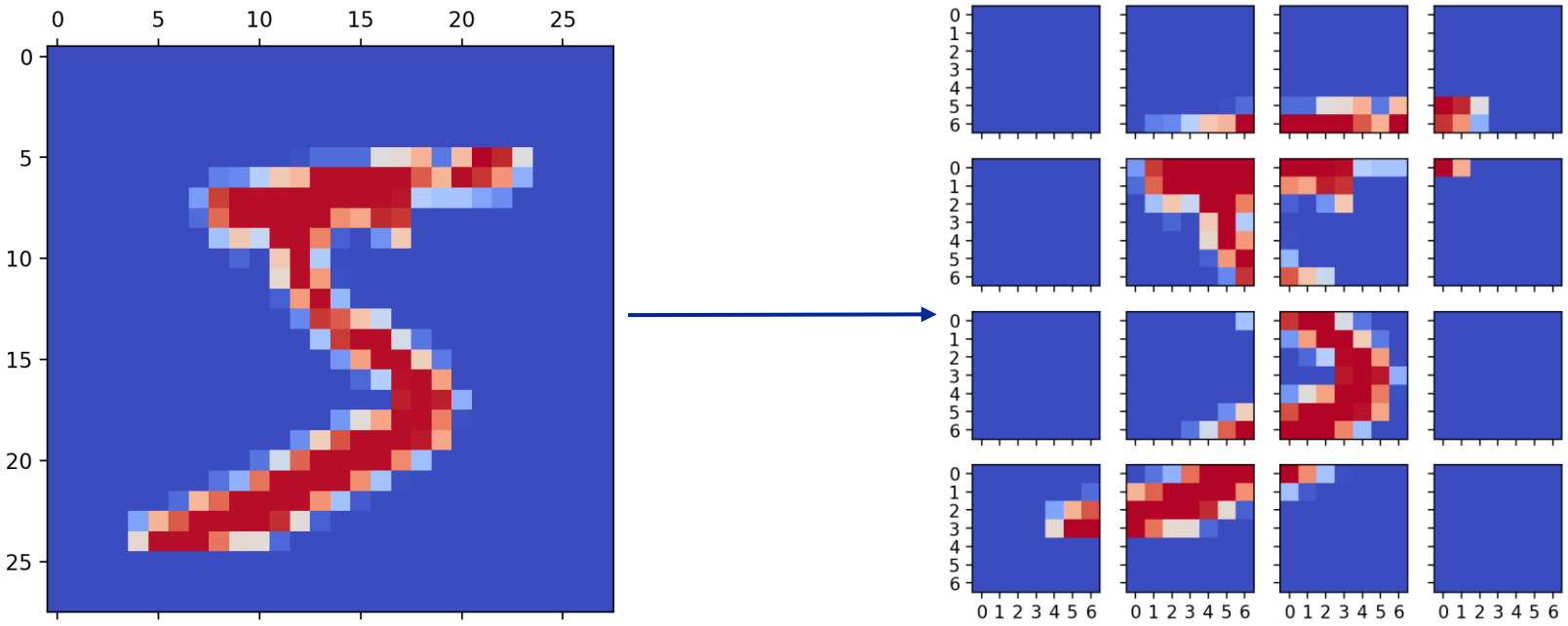
Embedding Raum dient die Pixel strukturiet ein ordnen , um dann zu klassifizieren

VISION TRANSFORMER FROM SCRATCH



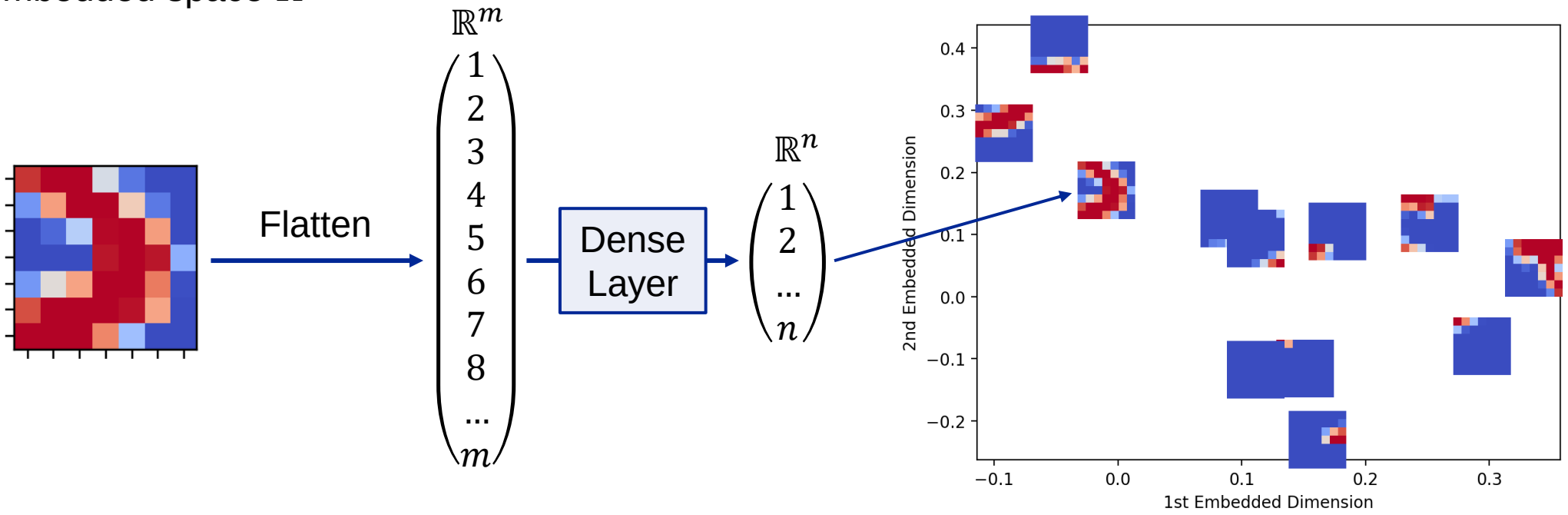
VISION TRANSFORMER PREPROCESSING PATCHING

- ViTs divide an image into small fixed-size patches.
- Common sizes include 16x16 or 32x32 pixels. For this example, it'll be 7x7 pixels resulting in 4x4 patches



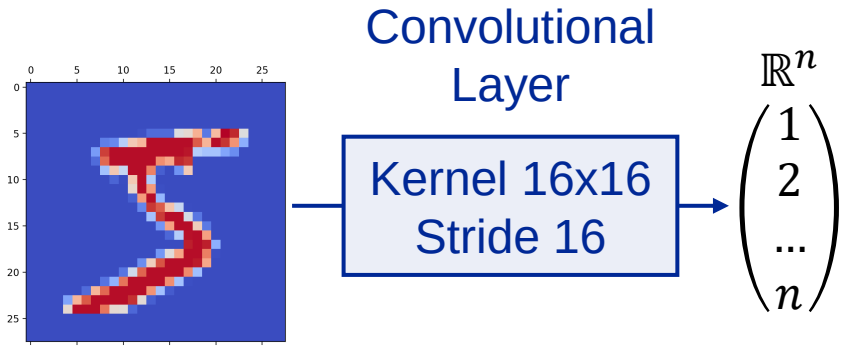
VISION TRANSFORMER PREPROCESSING
LINEAR EMBEDDING

- Maps the pixel values of the patch into a lower (or higher)-dimensional space
- Linear embedding can be achieved by a Dense Layer (without activation) projecting from patch space $\mathbb{R}^m$ to embedded space $\mathbb{R}^n$

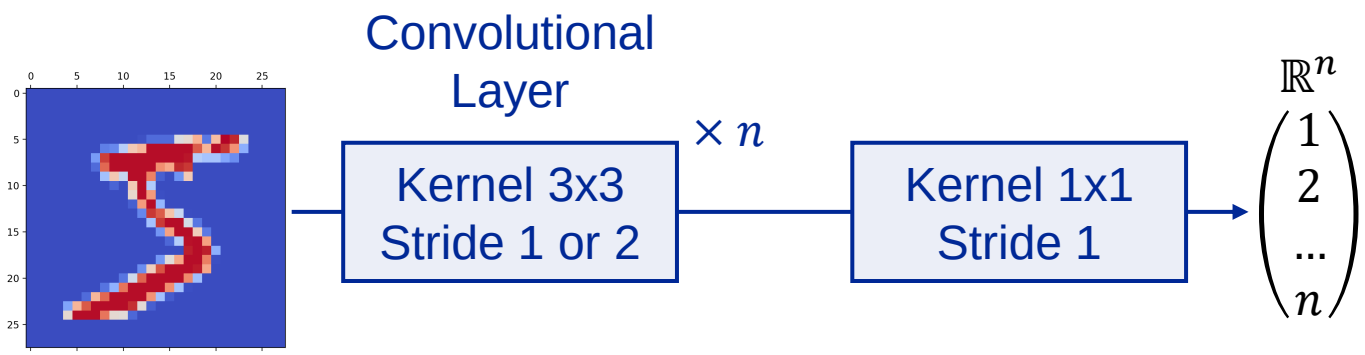


VISION TRANSFORMER PREPROCESSING LINEAR EMBEDDING

- Later attempts suggest other techniques involving concolutional layers for patching AND embedding.



- More robust, quicker, less parameters, outperforms CNNs first time on Image Net



Early Convolutions Help Transformers See Better, https://arxiv.org/abs/2106.14881

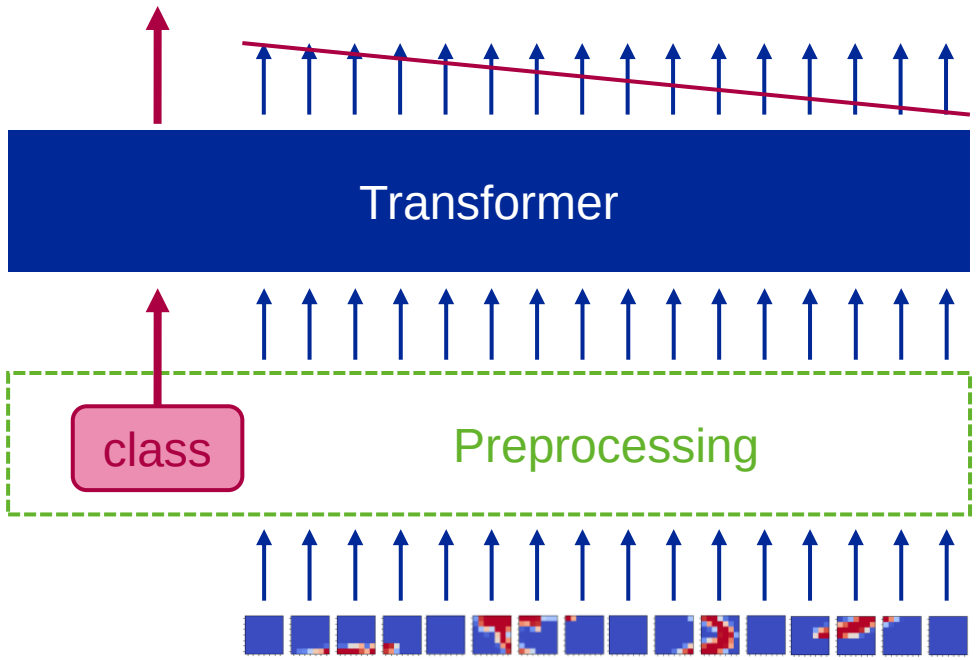
Der class token , soll dabei helfen welche bild inhalt wichtig für die klassiificationn ist !

VISION TRANSFORMER PREPROCESSING THE CLASS TOKEN

- After the embedding, image date can be treated **similarly to language data**. Hence, the embedded patches are often referred to as **tokens**, just like the word embeddings in Natural Language Processing (NLP) Transfomers

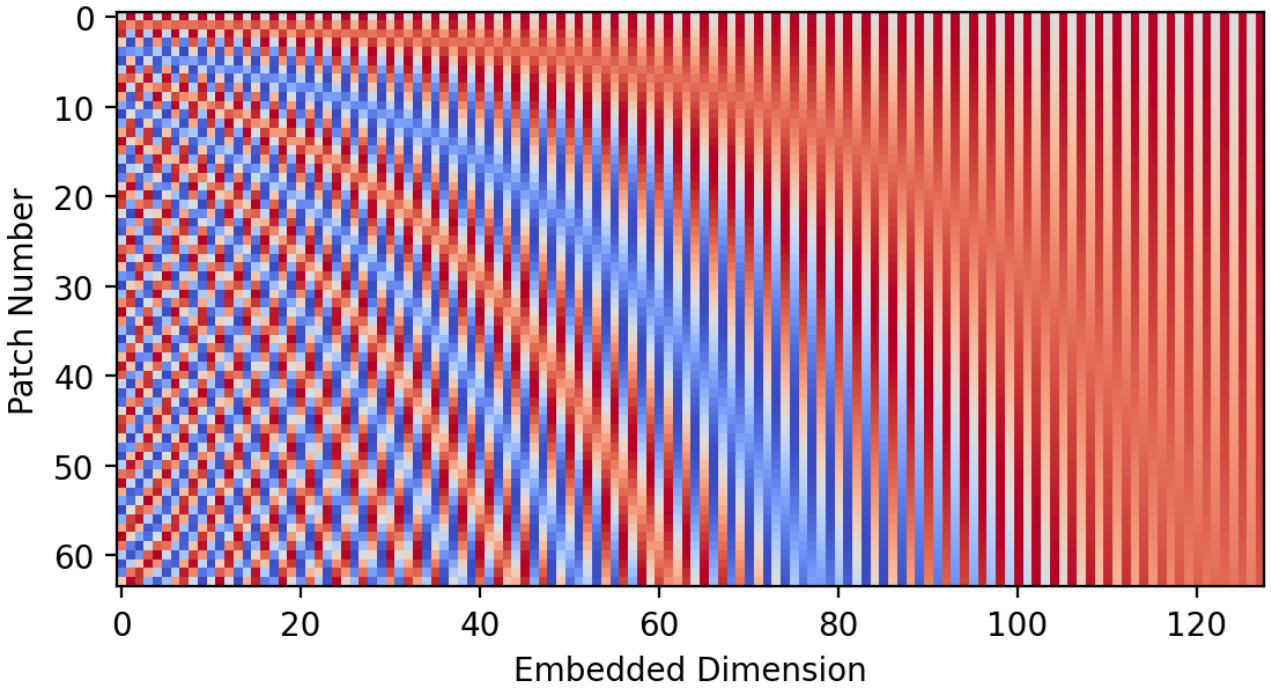
- Since, the NLP Transofmer we are using (see Attention Is All You Need, https://arxiv.org/abs/1706.03762) is originally designed as **Sequence-to-Sequence** model, we‘ll use a trick to apply it to **classification tasks**

- We‘ll add a **classifikation token** to our image tokens of the same dimension. This token will be trained to contain **the relevant class information at the transformer output**



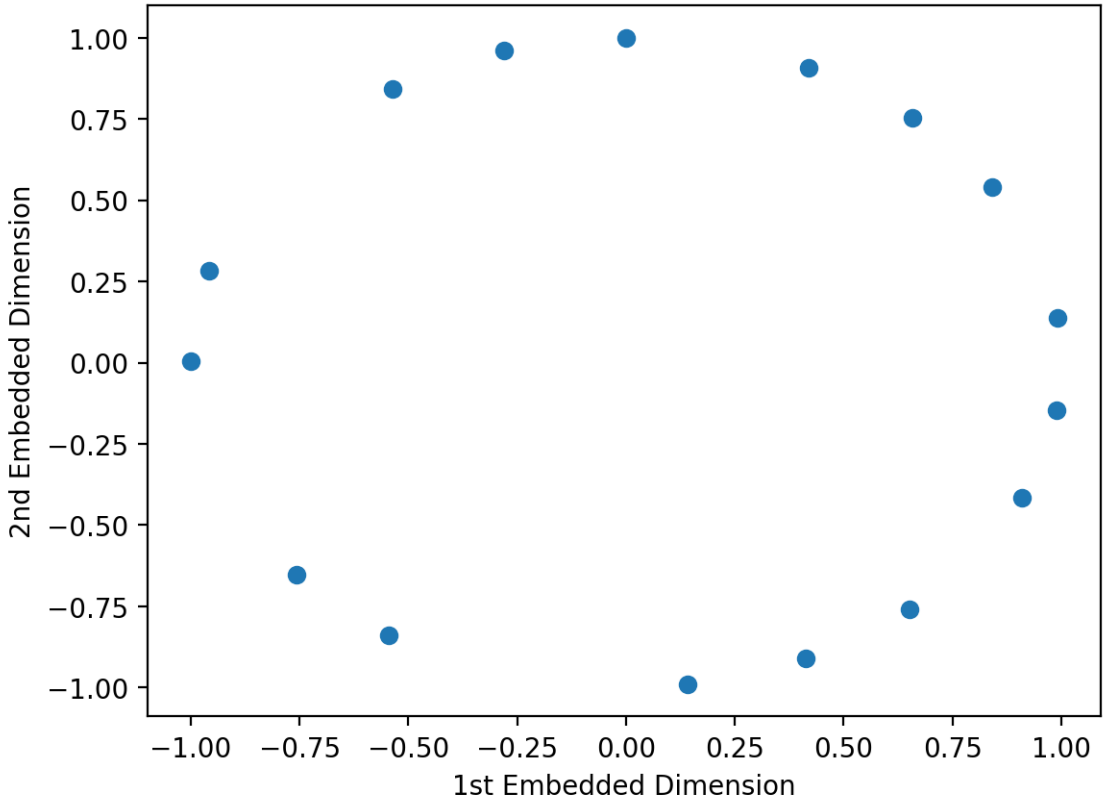
VISION TRANSFORMER PREPROCESSING POSITIONAL ENCODING

- Provides the model with information about the position of the patches.
- Essential since the transformer architecture does not inherently account for the order of input data.
- Fixed and learned positional encoding possible
- E.g. by adding $\begin{cases} \sin\left(\frac{i}{10\,000^{j/D_{\text{emb}}}}\right) & \text{for even } j \\ \cos\left(\frac{i}{10\,000^{(j-1)/D_{\text{emb}}}}\right) & \text{for odd } j \end{cases}$
- Provides a unique “bar code” for each patch i at each embedded coordinate in D_{emb} dimensions



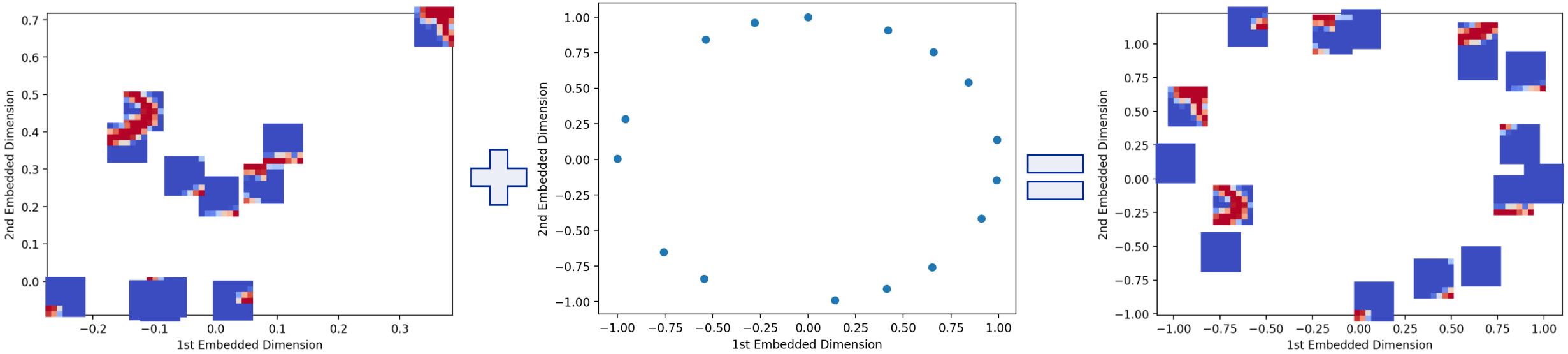
VISION TRANSFORMER PREPROCESSING POSITIONAL ENCODING

- Can also be interpreted as position on the unit circle in D_{emb} dimensions
- E.g. with $D_{\text{emb}} = 2$: $\begin{cases} \sin\left(\dfrac{i}{10\,000^{j/D_{\text{emb}}}}\right) & \text{for even } j \\ \cos\left(\dfrac{i}{10\,000^{(j-1)/D_{\text{emb}}}}\right) & \text{for odd } j \end{cases}$



VISION TRANSFORMER PREPROCESSING POSITIONAL ENCODING

- Can also be interpreted as position on the unit circle in D_{emb} dimensions



SUMMARY

- Unlike CNNs, Transformers do not posses an inherent positional bias. Hence, they are better suited to assess long range information in images but need more data for the training and more compute recourses
- To use the NLP Transformer-Architecture several preprocessing steps need to be performed:
  - Patching: dividing the image into quadratic crops
  - Linear Embedding: Projects all image patches into a lower dimensional space as vectors referred to as tokens
  - Class Token: Adds a token to the image tokens that will be used later to extract class information from the Transformer
  - Positional Encoding: Adds a unique pattern ontop the input tokens to encode their respective position in the image.