Attention and Transformers

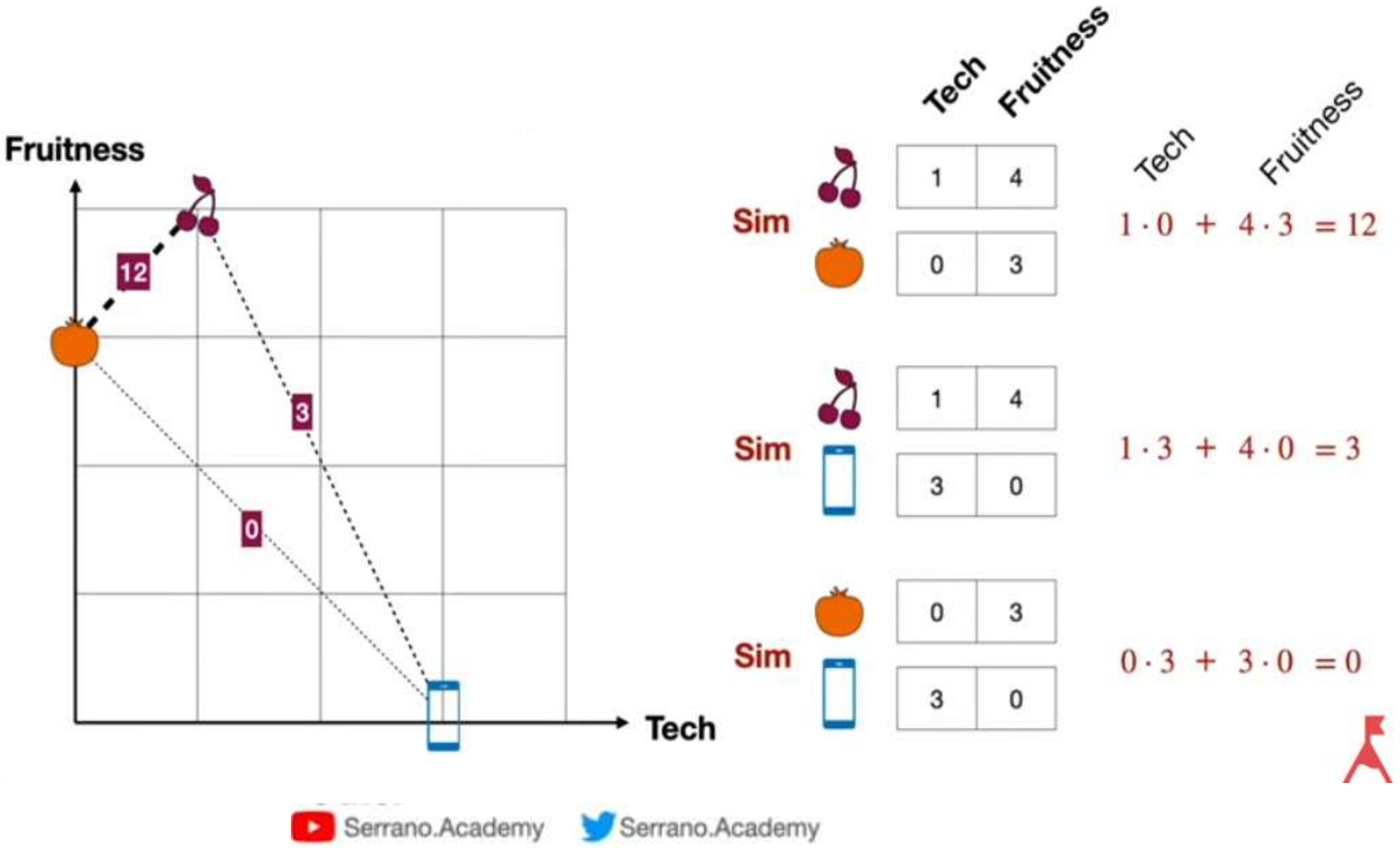
Saehwa Kim

Information and Communications Engineering
Hankuk University of Foreign Studies

Outline

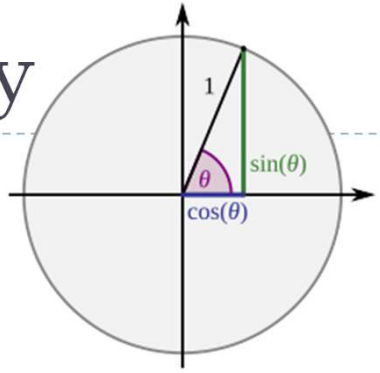
- Dot Product
- Attention Scores
- New Word Embedding with Attention Scores
- Attention in Transformers
- Transformer Architecture
- GPT and Transformers

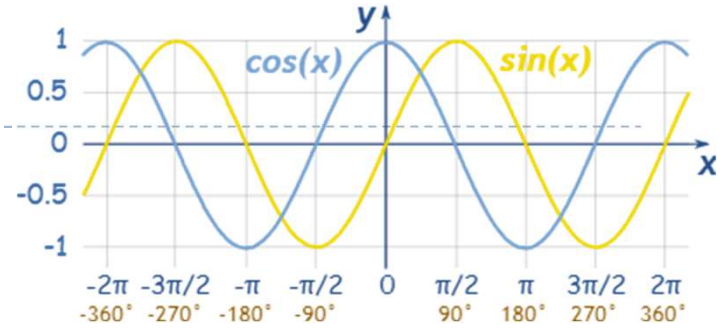
Dot Product

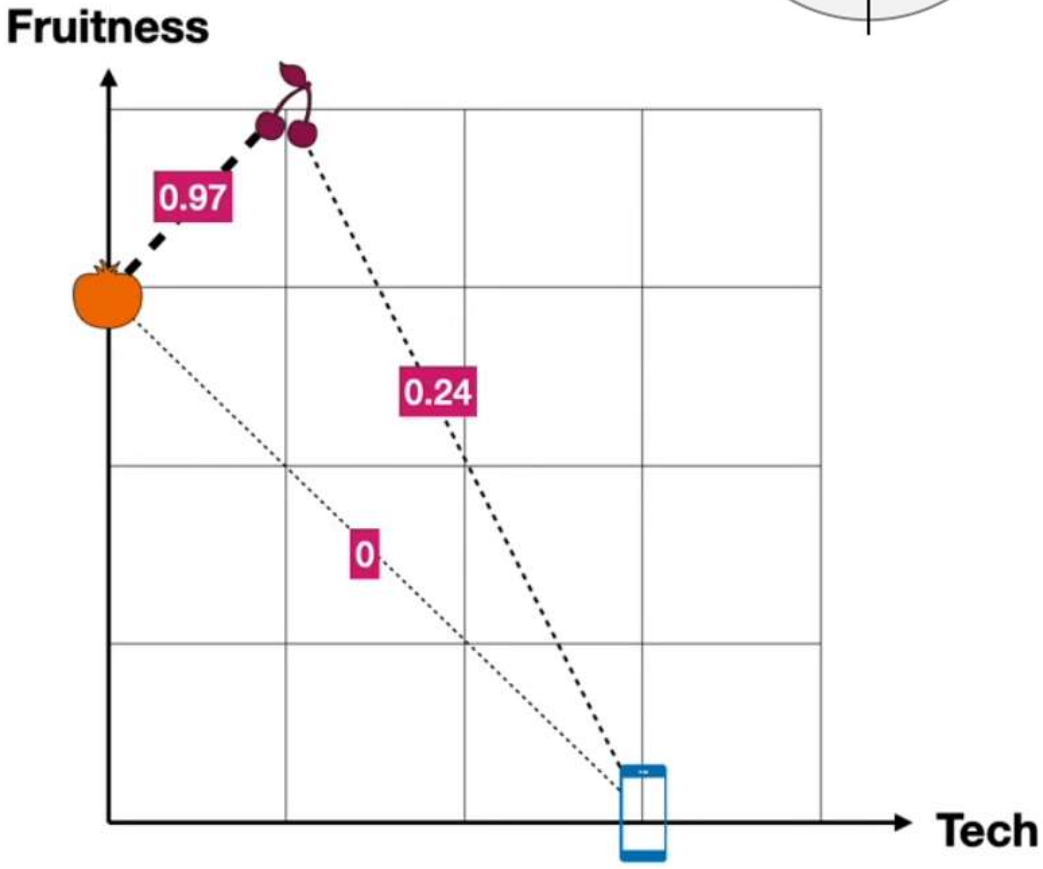


https://www.youtube.com/watch?v=UPtG_38Oq8o

Cosine Similarity







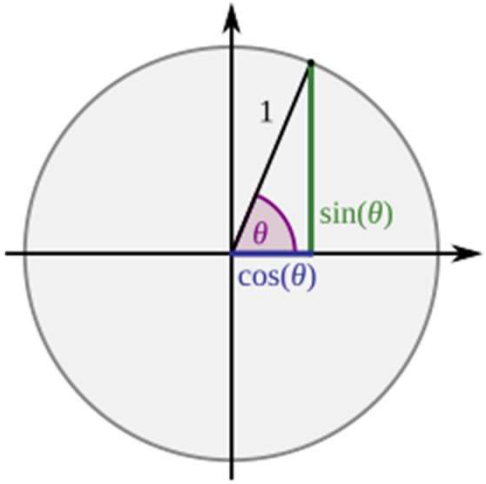
$cos(14°) = 0.97$

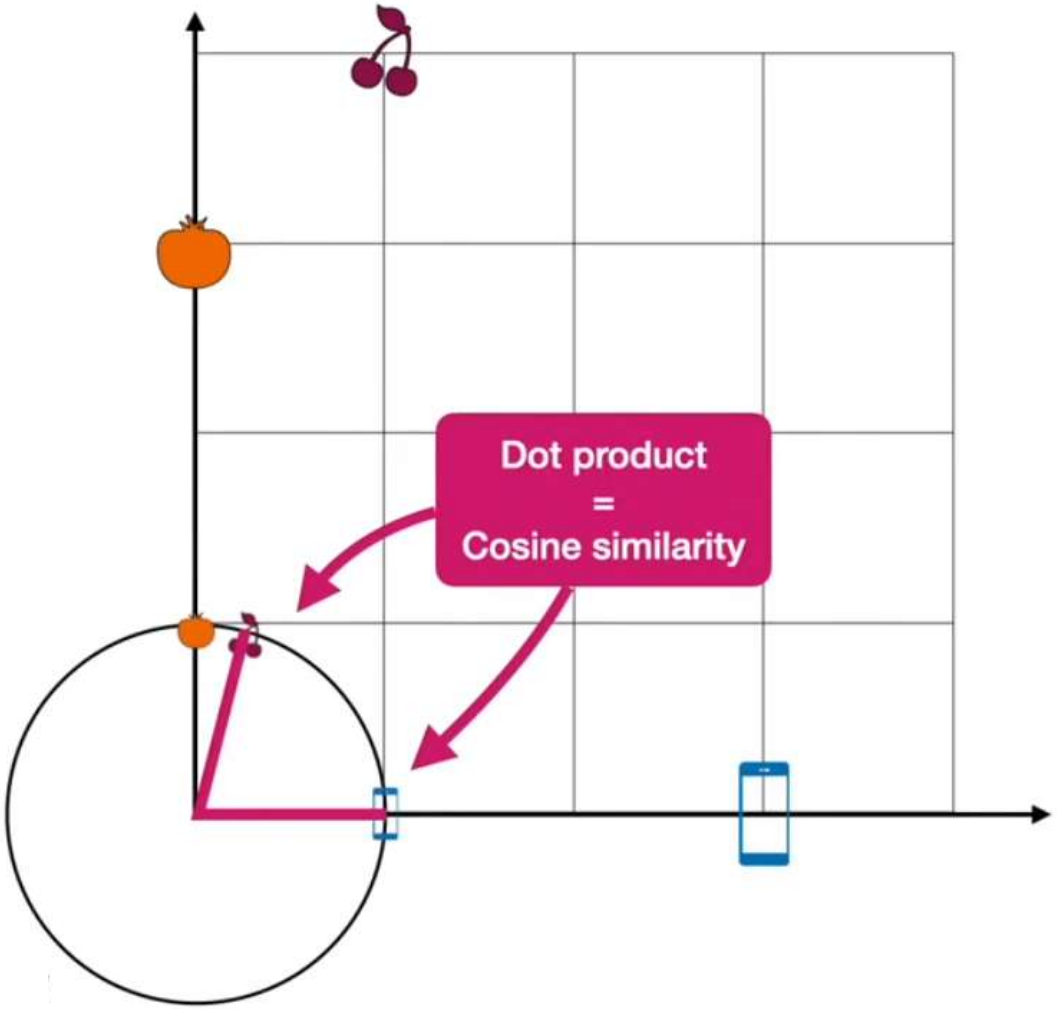
$cos(76°) = 0.24$

$cos(90°) = 0$

https://www.youtube.com/watch?v=UPtG_38Oq8o

Dot Product ≈ Cosine Similarity (1/2)



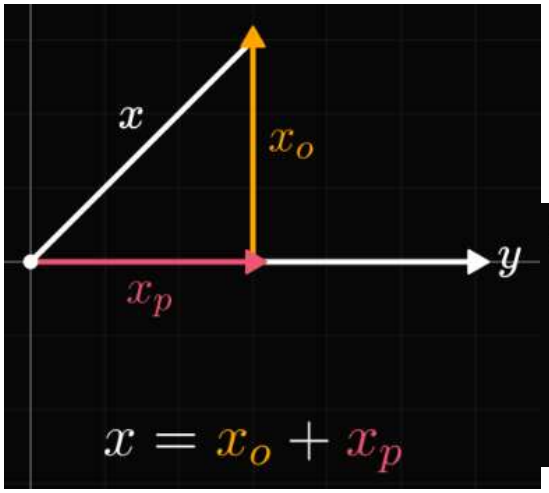


https://www.youtube.com/watch?v=UPtG_38Oq8o

Dot Product ≈ Cosine Similarity (2/2)

$$\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i$$

$$x = (x_1, \ldots, x_n)$$
$$y = (y_1, \ldots, y_n)$$



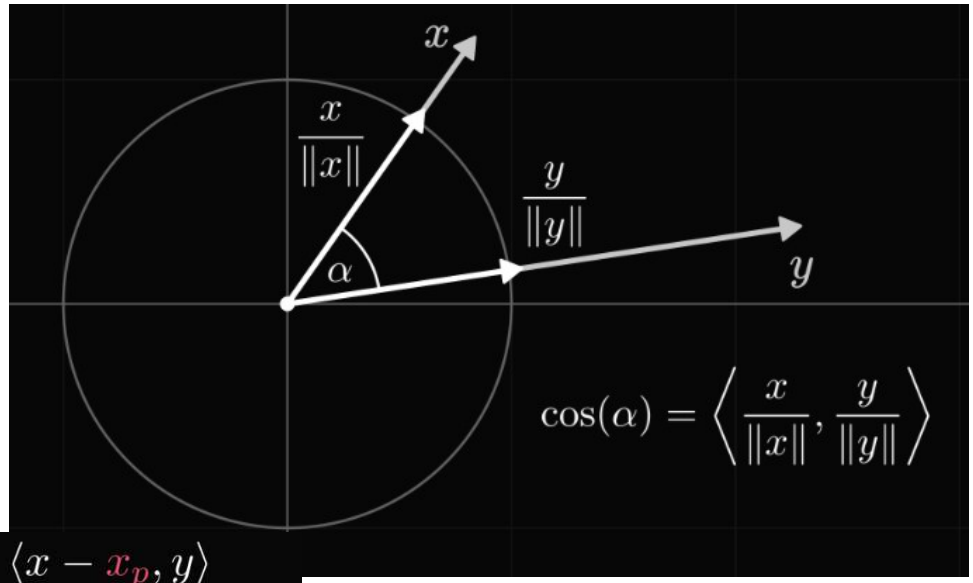
$$x = x_o + x_p$$

$$x_p = cy, \quad (c \in \mathbb{R})$$
$$\langle x_o, y \rangle = 0$$

$$\begin{aligned} \langle x_o, y \rangle &= \langle x - x_p, y \rangle \\ &= \langle x - cy, y \rangle \\ &= \langle x, y \rangle - c\langle y, y \rangle \\ &= 0 \end{aligned}$$

$$c = \frac{\langle x, y \rangle}{\langle y, y \rangle} = \frac{\langle x, y \rangle}{\|y\|^2}$$

$\langle y, y \rangle = \|y\|^2$
(magnitude of y)

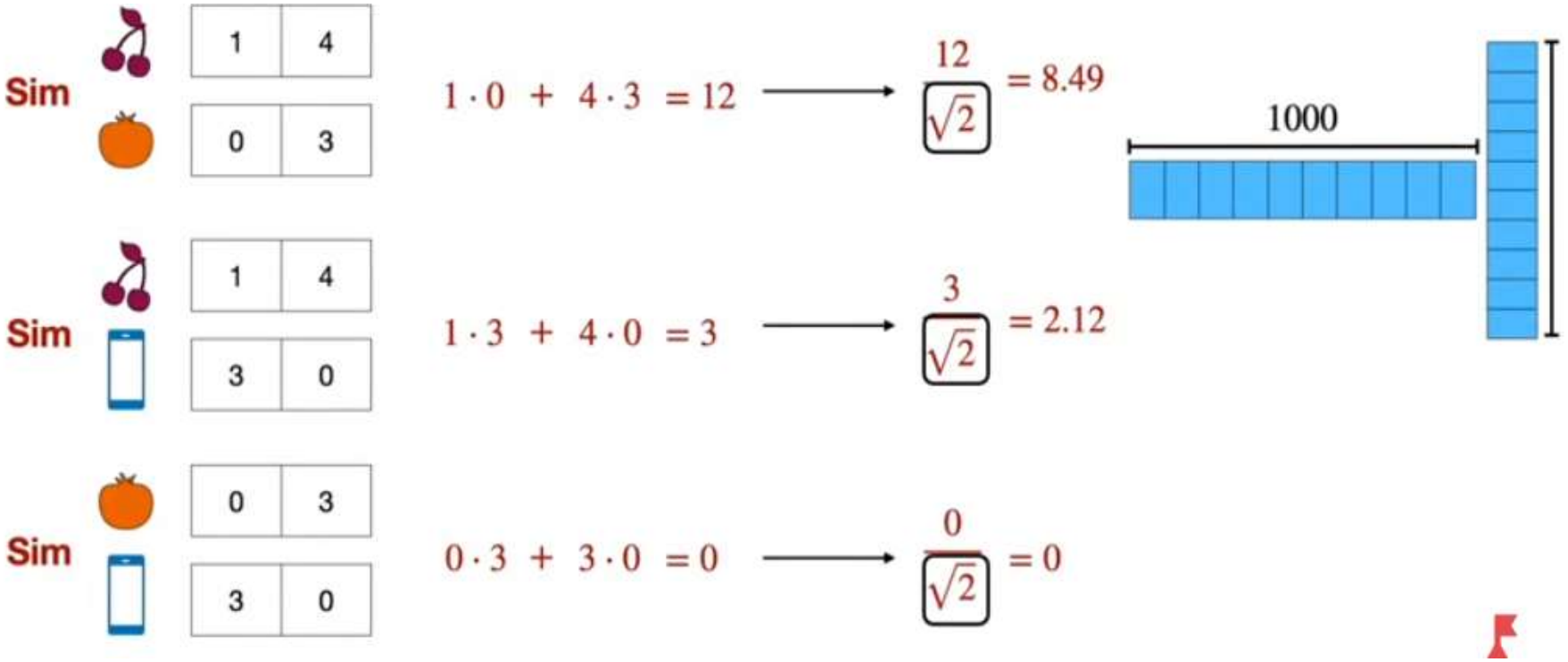


$$\cos(\alpha) = \left\langle \frac{x}{\|x\|}, \frac{y}{\|y\|} \right\rangle = \frac{\langle x, y \rangle}{\|x\|\|y\|}$$

https://tivadardanka.com/blog/how-the-dot-product-measures-similarity

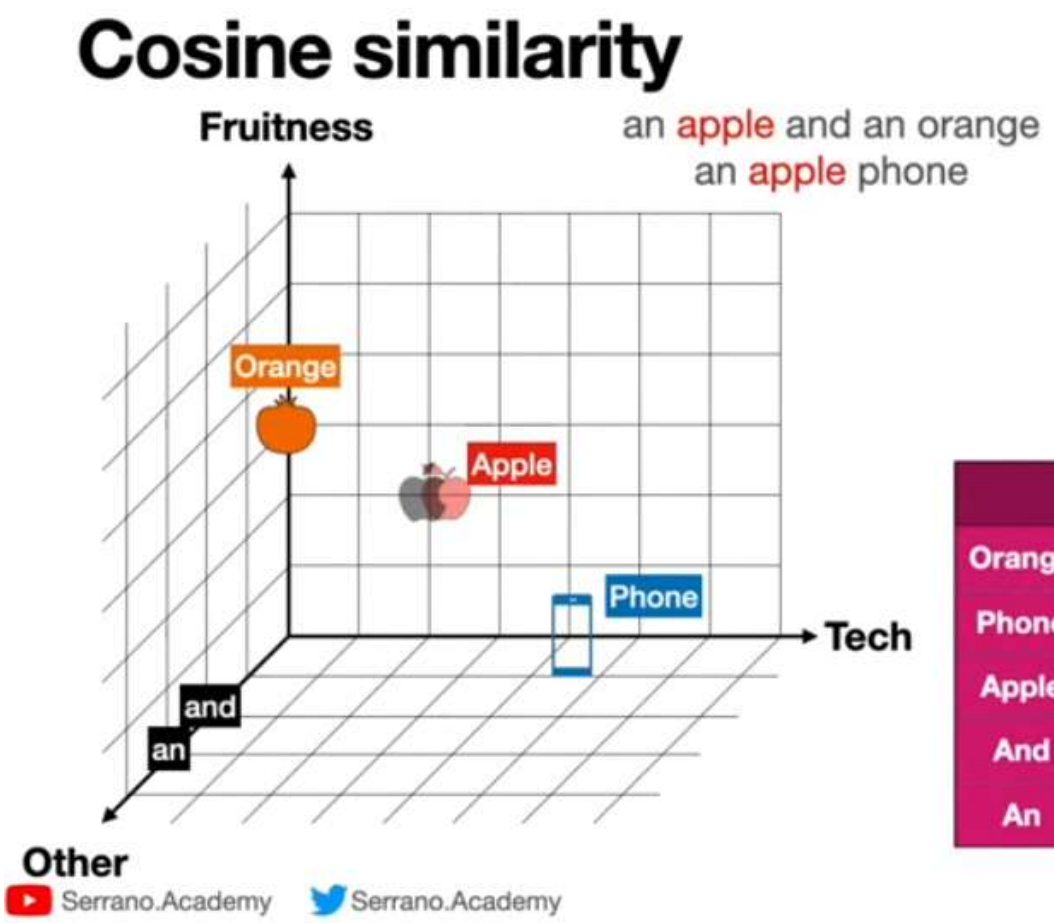
Scaled Dot Product

- Dot product divided by the square root of the length of the vector



https://www.youtube.com/watch?v=UPtG_38Oq8o

Attention Scores



	Tech	Fruitness	Other
Orange	0	3	0
Phone	4	0	0
Apple	2	2	0
And	0	0	2
An	0	0	3

	Orange	Phone	Apple	And	An
Orange	1	0	0.71	0	0
Phone	0	1	0.71	0	0
Apple	0.71	0.71	1	0	0
And	0	0	0	1	1
An	0	0	0	1	1

https://www.youtube.com/watch?v=UPtG_38Oq8o

New Word Embedding with Attention Scores (1/4)

an apple and an orange

	Orange	Apple	And	An
Orange	1	0.71	0	0
Apple	0.71	1	0	0
And	0	0	1	1
An	0	0	1	1

Orange → 1 Orange + 0.71 Apple

Apple → 0.71 Orange + 1 Apple

And → 1 And + 1 An

An → 1 An + 1 And

an apple phone

	Phone	Apple	An
Phone	1	0.71	0
Apple	0.71	1	0
An	0	0	1

Phone → 1 Phone + 0.71 Apple

Apple → 0.71 Phone + 1 Apple

An → 1 An

https://www.youtube.com/watch?v=UPtG_38Oq8o

New Word Embedding with Attention Scores (2/4)

- Scaling

an apple and an orange

	Orange	Apple	And	An
Orange	1	0.71	0	0
Apple	0.71	1	0	0
And	0	0	1	1
An	0	0	1	1

Orange → 0.57 Orange + 0.43 Apple

Apple → 0.43 Orange + 0.57 Apple

And → 0.5 And + 0.5 An

An → 0.5 An + 0.5 And

an apple phone

	Phone	Apple	An
Phone	1	0.71	0
Apple	0.71	1	0
An	0	0	1

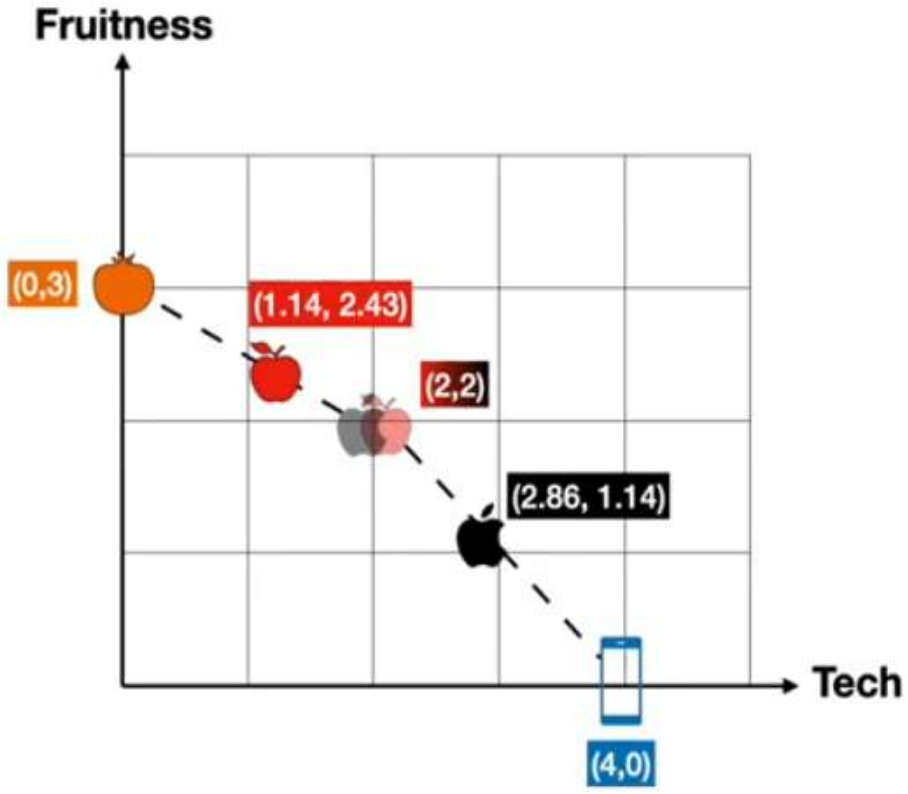
Phone → 0.57 Phone + 0.43 Apple

Apple → 0.43 Phone + 0.57 Apple

An → 1 An

https://www.youtube.com/watch?v=UPtG_38Oq8o

New Word Embedding with Attention Scores (3/4)



an apple and an orange

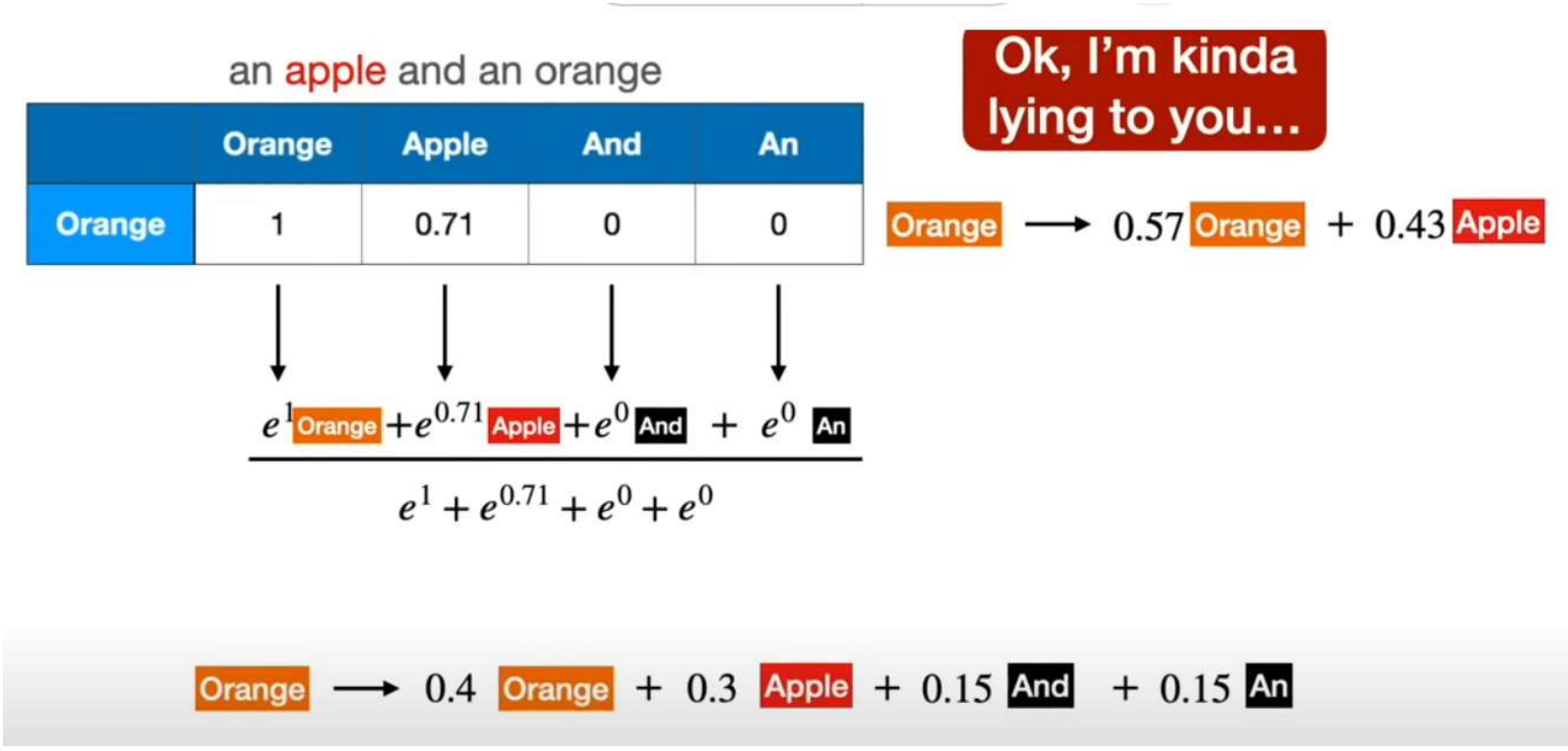
Apple → 0.43 Orange + 0.57 Apple

an apple phone

Apple → 0.43 Phone + 0.57 Apple

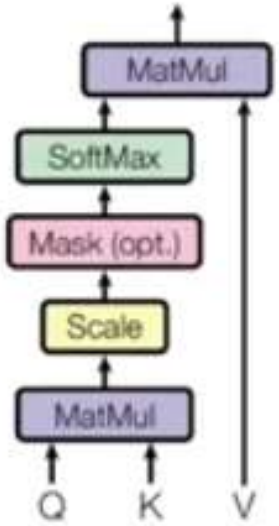
New Word Embedding with Attention Scores (4/4)

- Softmax normalization

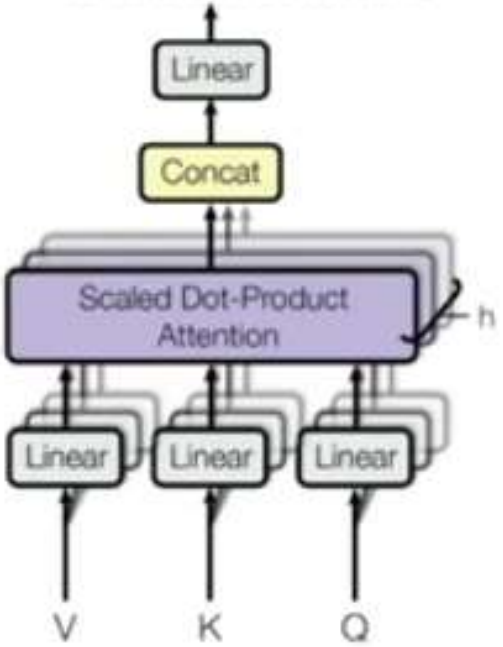


Attention in Transformers

Scaled Dot-Product Attention



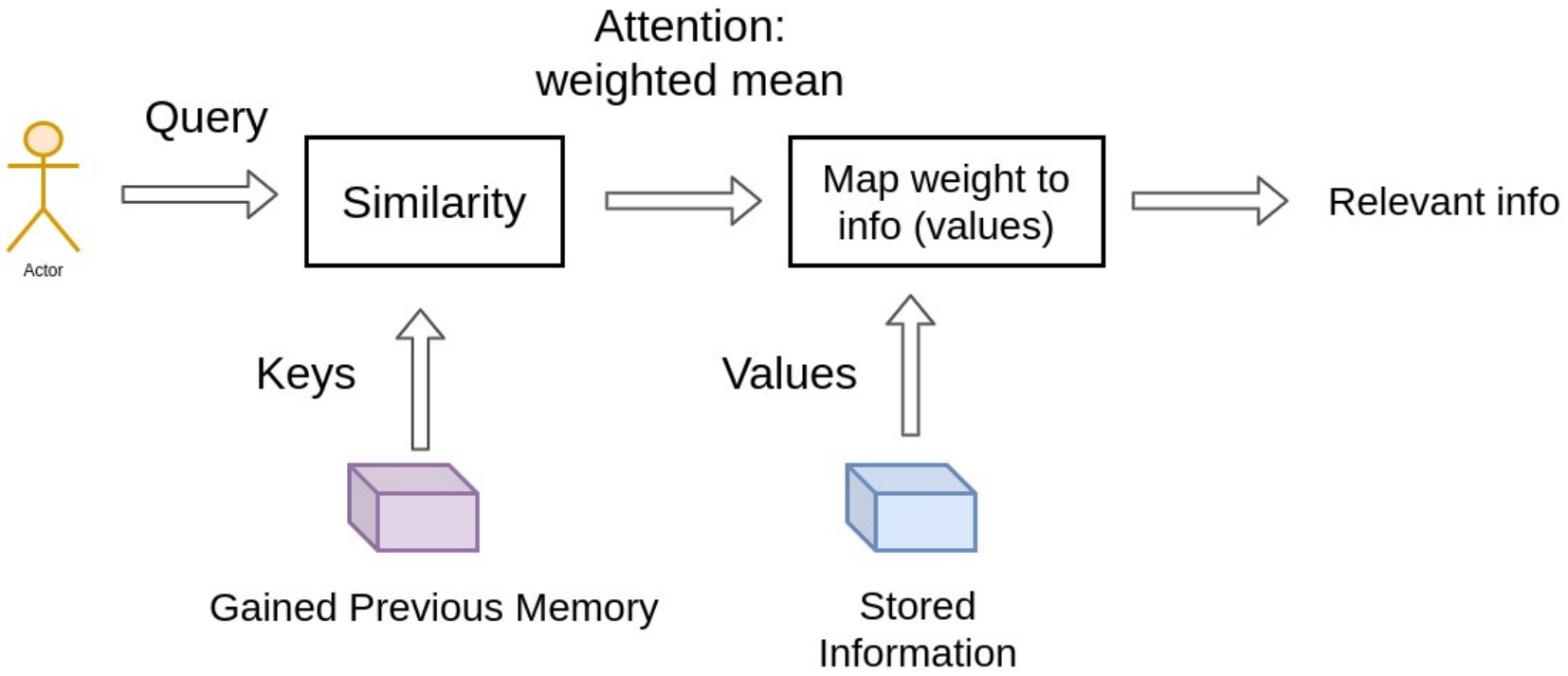
Multi-Head Attention



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

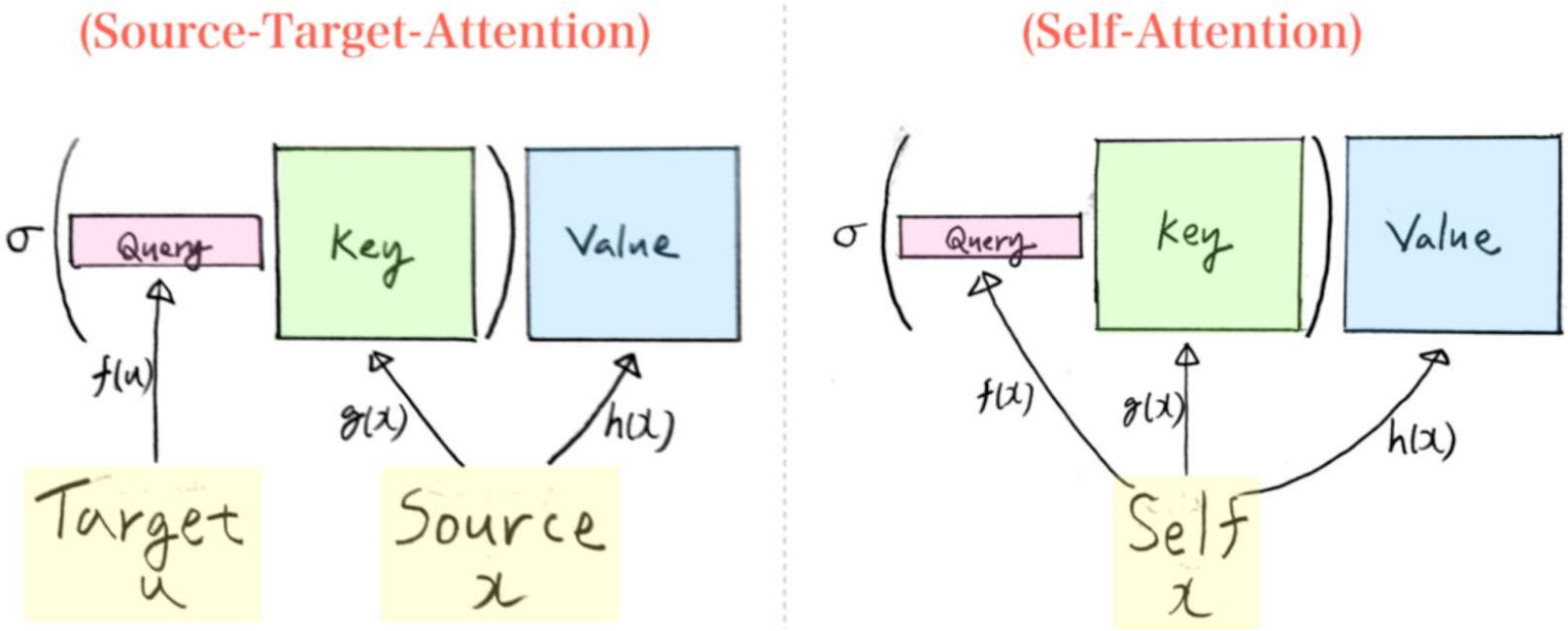
$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_\text{i}^Q, KW_\text{i}^K, VW_\text{i}^V)$$

Notion of Query, Key, and Values



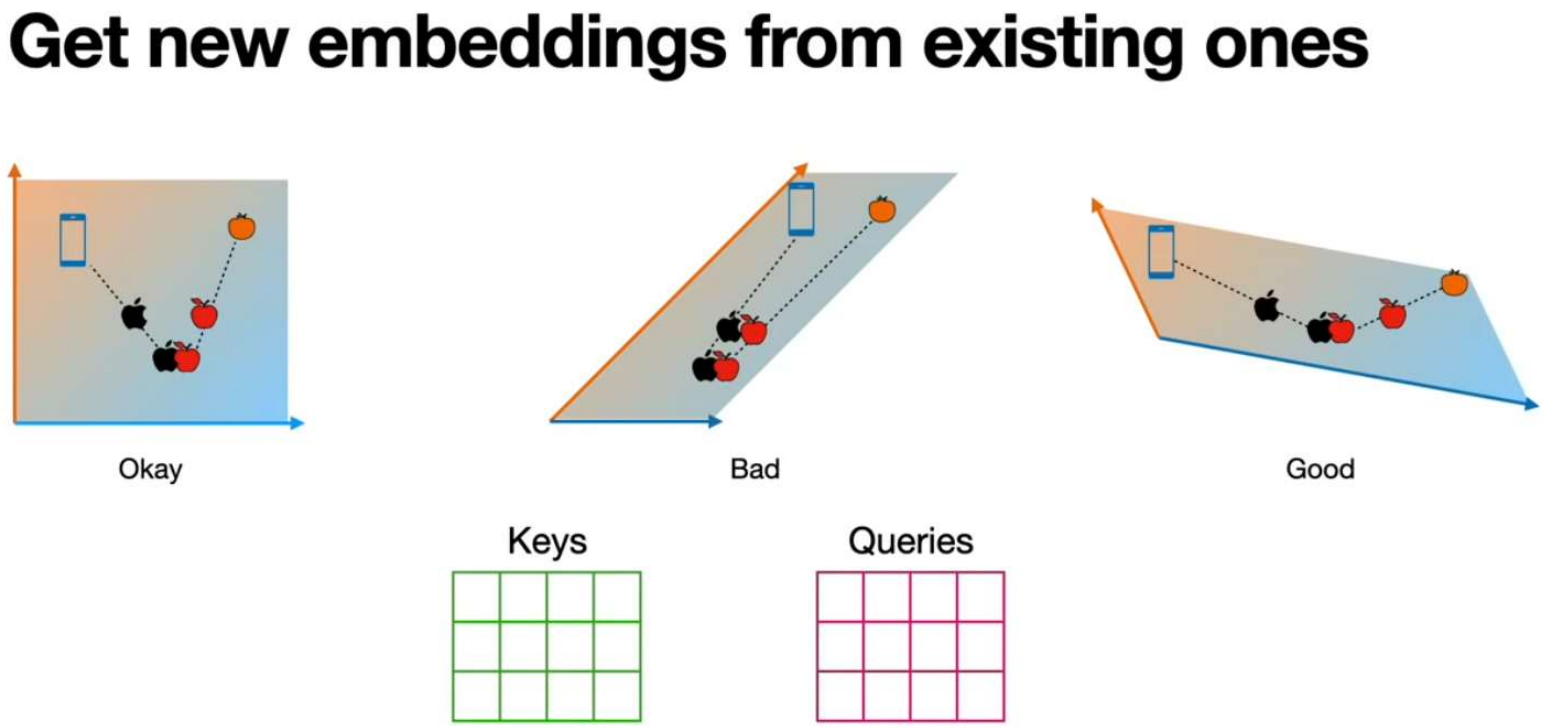
https://blog.sailor.plus/deep-learning/transformer-family/

Source-Target Attention vs. Self-Attention



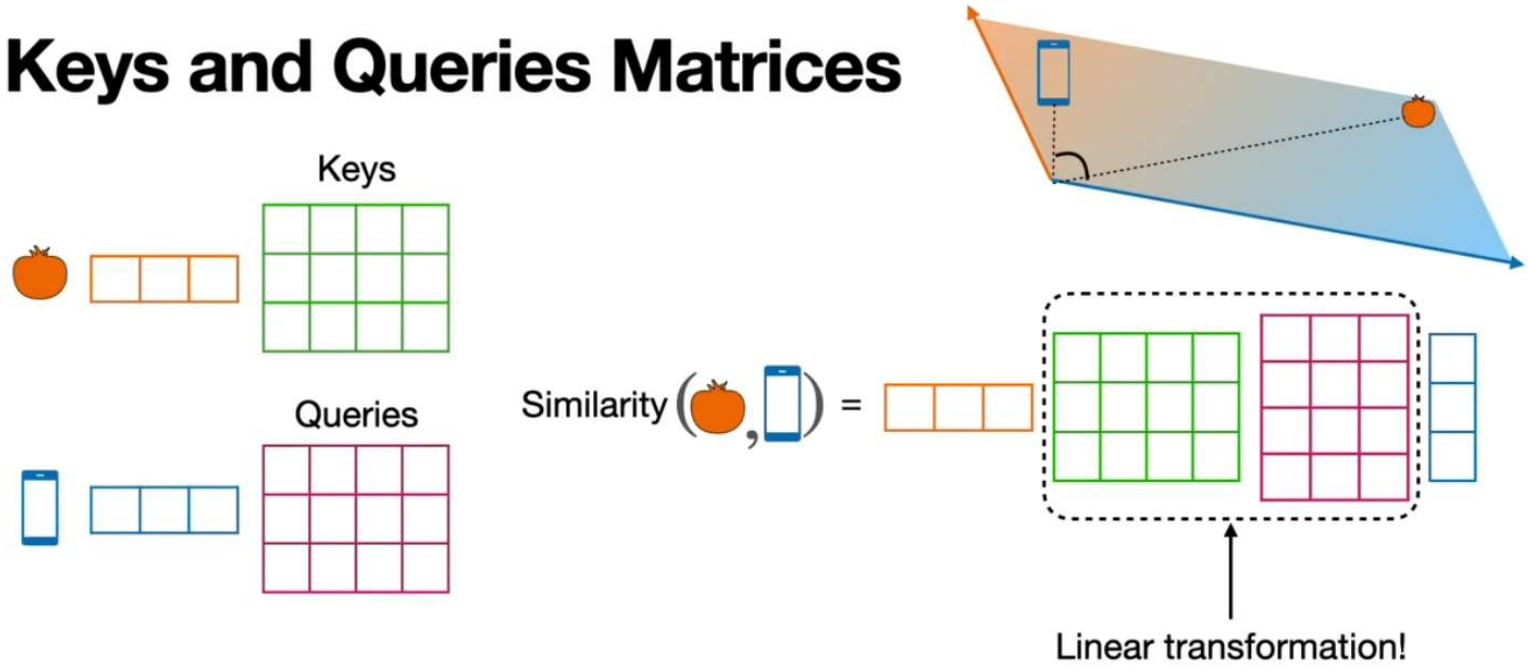
https://pretteyandnerdy.wordpress.com/2019/04/26/attention-is-all-you-need/

Weights for Keys and Queries



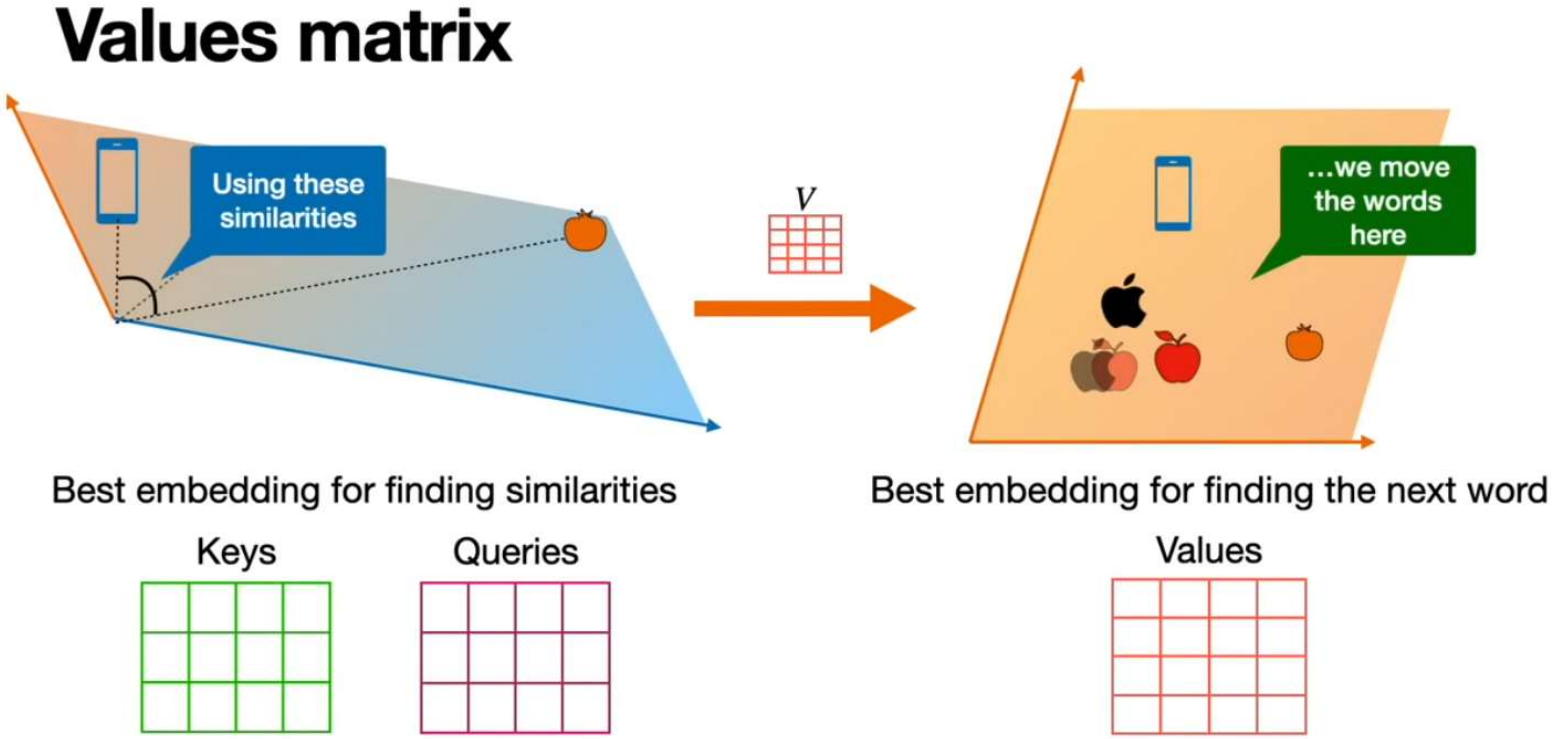
https://www.youtube.com/watch?v=UPtG_38Oq8o

Getting Similarity of Queries and Keys



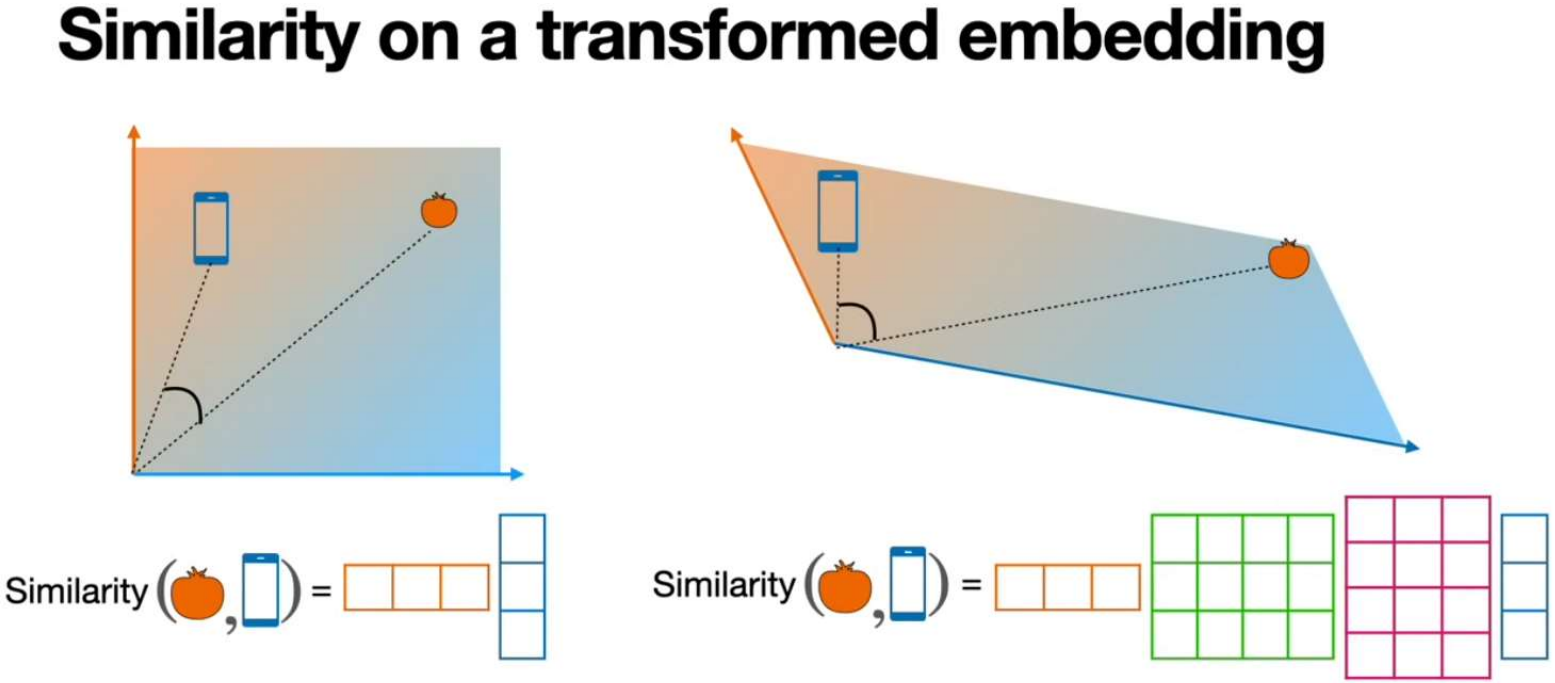
https://www.youtube.com/watch?v=UPtG_38Oq8o

Applying Value Matrix



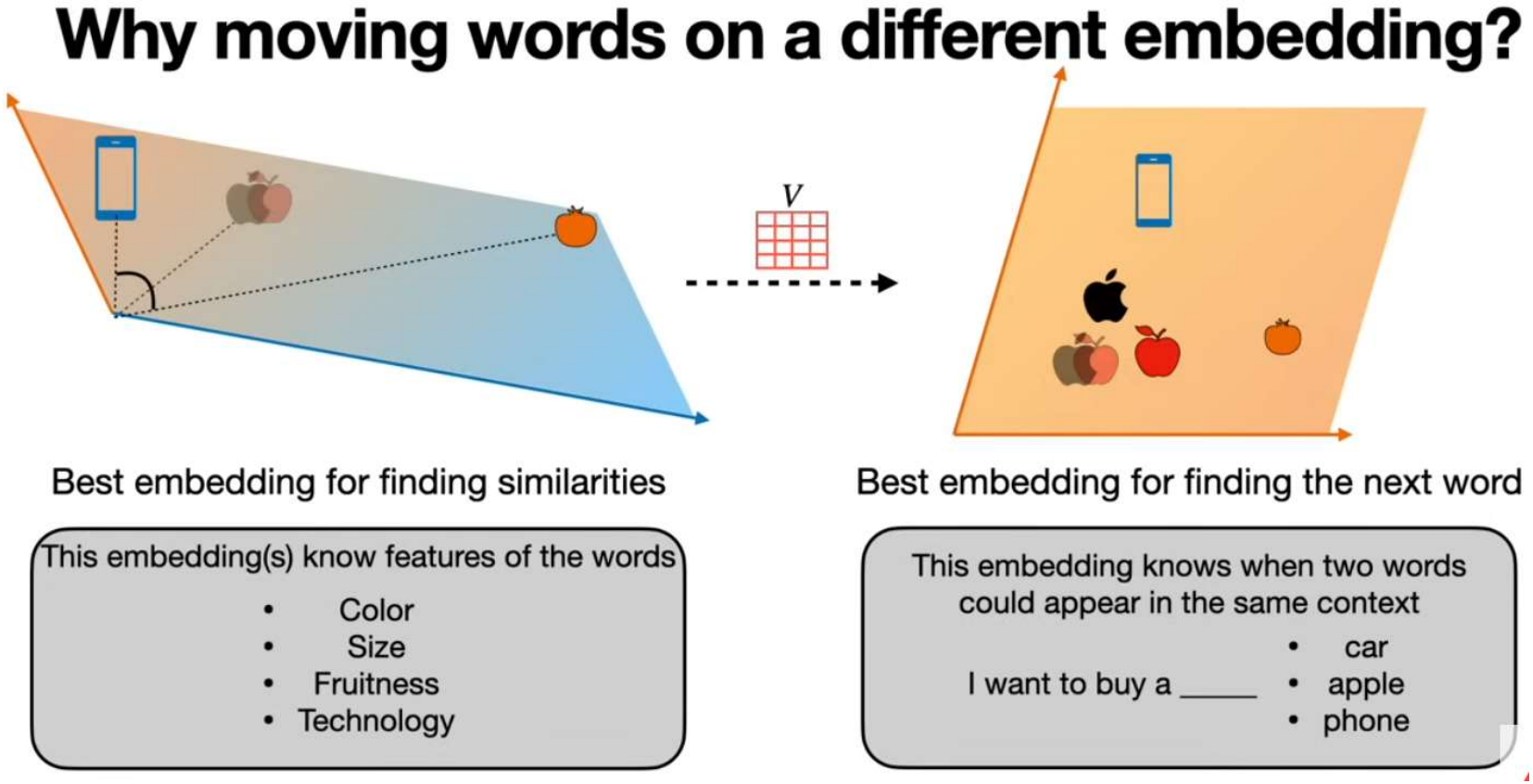
https://www.youtube.com/watch?v=UPtG_38Oq8o

Effects of Linear Transformations



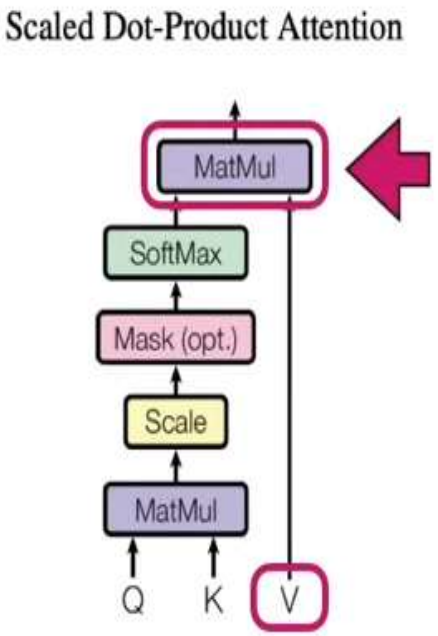
https://www.youtube.com/watch?v=UPtG_38Oq8o

Effects of Value Matrix

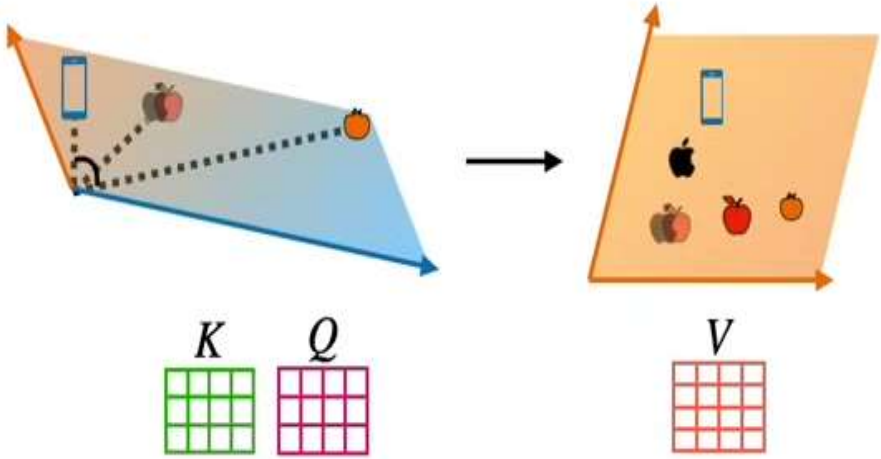


https://www.youtube.com/watch?v=UPtG_38Oq8o

Scaled Dot-Product Attention

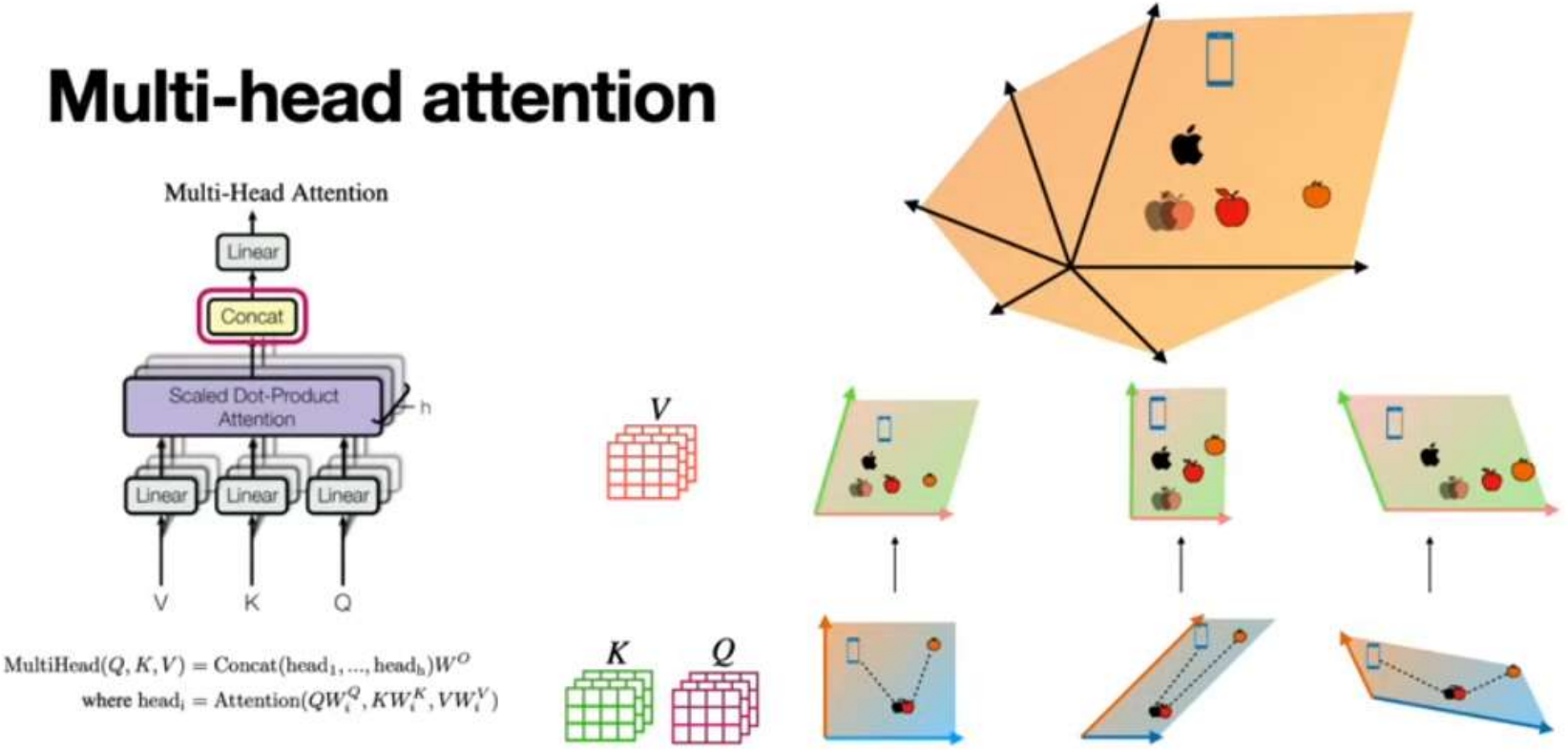


$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$



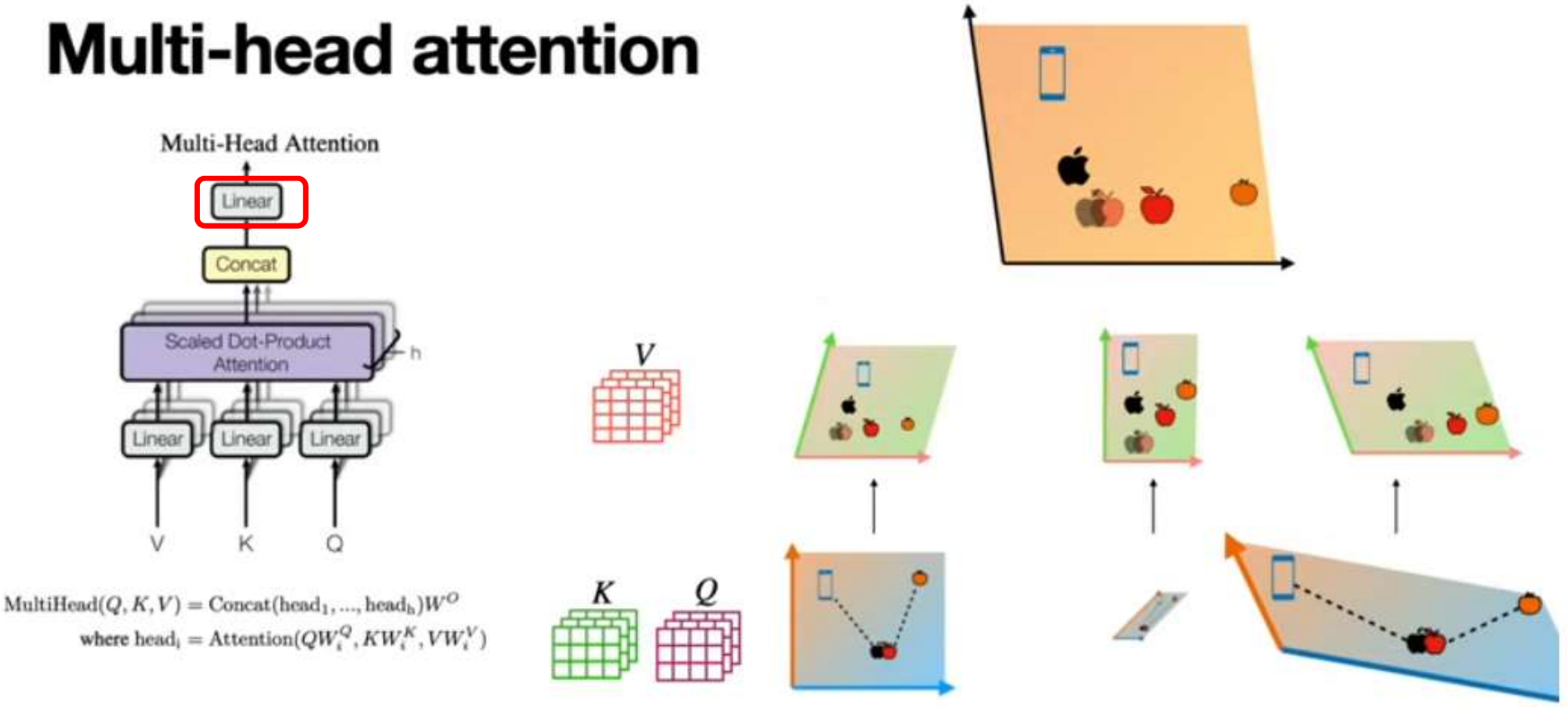
https://www.youtube.com/watch?v=UPtG_38Oq8o

Multi-head Attention (1/3)



https://www.youtube.com/watch?v=UPtG_38Oq8o

Multi-head Attention (2/3)



Multi-head Attention (3/3)

Value matrix

an apple and an orange

	Orange	Apple	And	An
Orange	0.4	0.3	0.15	0.15
Apple	0.3	0.4	0.15	0.15
And	0.15	0.15	0.5	0.5
An	0.15	0.15	0.5	0.5

Value matrix

=

	Orange	Apple	And	An
Orange	v_{11}	v_{12}	v_{13}	v_{14}
Apple	v_{21}	v_{22}	v_{23}	v_{24}
And	v_{31}	v_{32}	v_{33}	v_{34}
An	v_{41}	v_{42}	v_{43}	v_{44}

apple $\longrightarrow$ $0.3 \cdot$ orange $+0.4 \cdot$ apple $+0.15 \cdot$ and $+0.15 \cdot$ an

apple $\longrightarrow$ $v_{21} \cdot$ orange $+v_{22} \cdot$ apple $+v_{23} \cdot$ and $+v_{24} \cdot$ an

https://www.youtube.com/watch?v=UPtG_38Oq8o

Transformer Architecture (1/2)

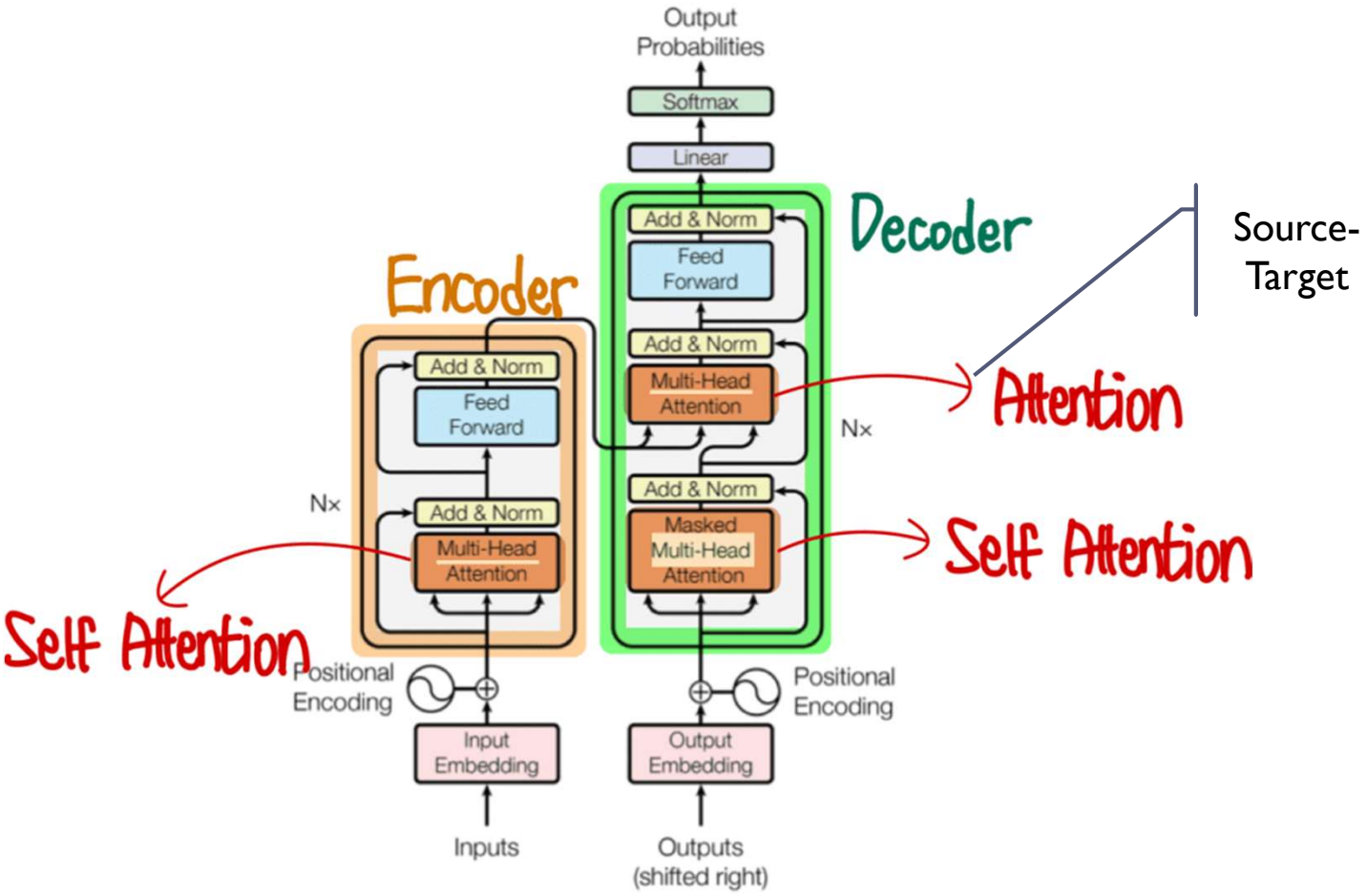
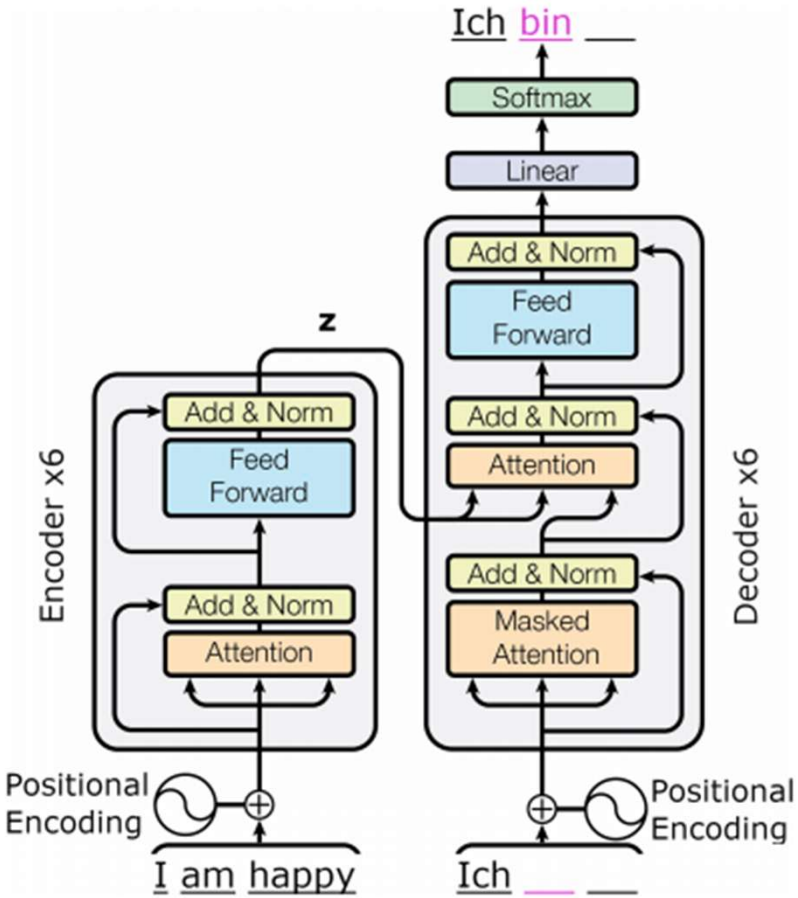


Figure 1: The Transformer - model architecture.

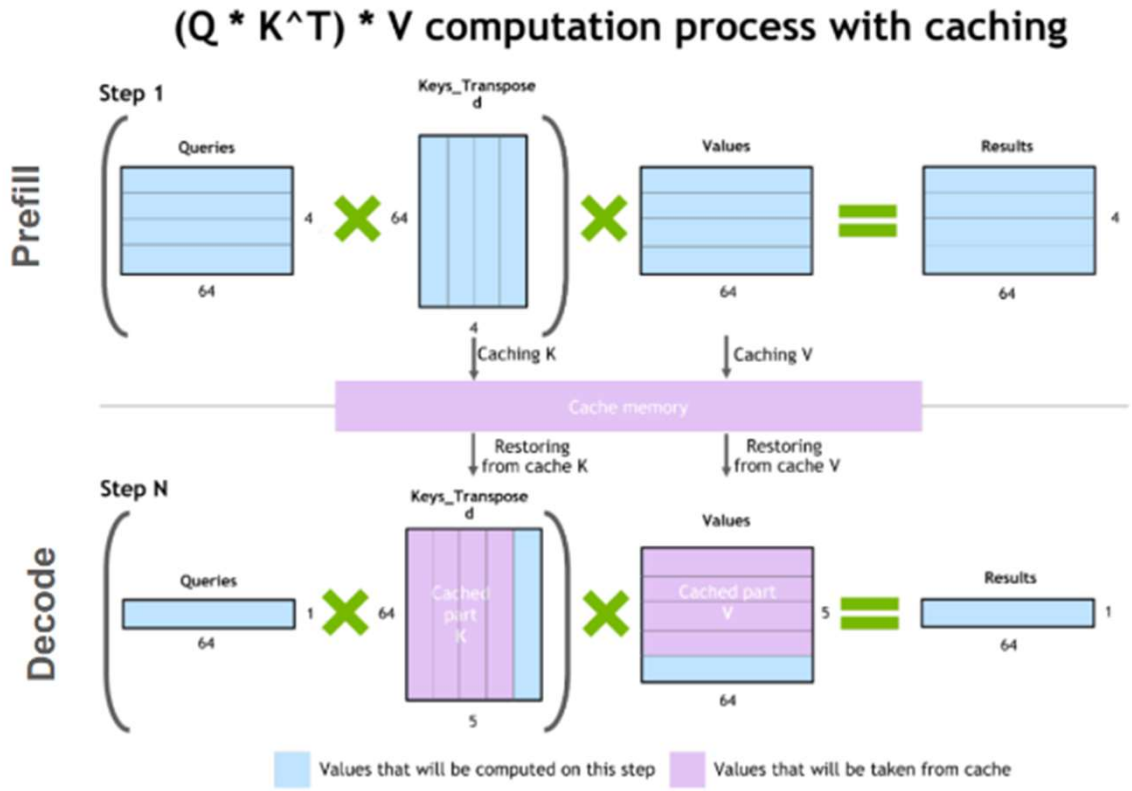
https://ffighting.net/deep-learning-basic/%EB%94%A5%EB%9F%AC%EB%8B%9D-%ED%95%B5%EC%8B%AC-%EA%B0%9C%EB%85%90/attention-and-self-attention-in-deep-learning/

Transformer Architecture (2/2)



Chat-GPT and Transformers

- Chat-GPT adopted decoder-only transformer architecture
- Key Value caching in inference



https://developer.nvidia.com/blog/mastering-llm-techniques-inference-optimization/