Quantum Self-Attention

Dominic Pasquali

March, 23 2022

2022 CERN openlab Technical Workshop

QUANTUM
TECHNOLOGY
INITIATIVE

Agenda

- Motivation for Quantum GANs & Quantum Self-Attention
- Classical Self-Attention
- Quantum Self-Attention
- Multi-Head Attention

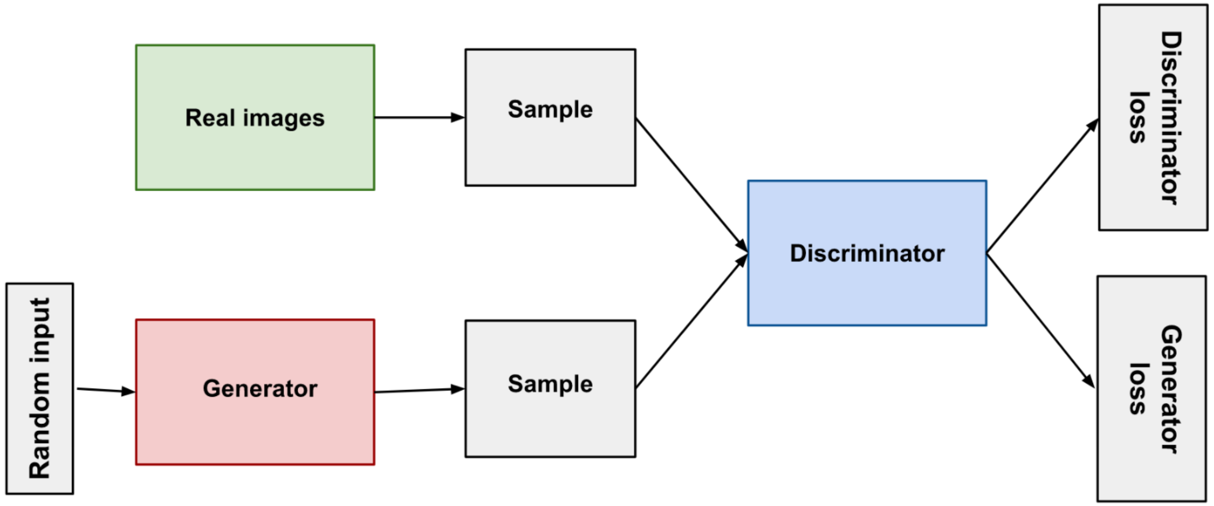
Motivation For Quantum GANs & Quantum Self-Attention

Why GANs:

- Simulation of particle transport through matter is fundamental for understanding the physics of High Energy Physics (HEP) experiments
- Most of LHC CPU budget (~ 1M CPU-years!!!) is dedicated to Monte Carlo simulation
- Faster approach: Replace Monte Carlo simulation with deep learning algorithms (e.g. GANs)

Why QGANs:

- compressed data representation in quantum states
- expect faster training with less number of parameters
- potential advantage of Quantum GAN[1]



Explore different prototypes of Quantum GANs to improve model

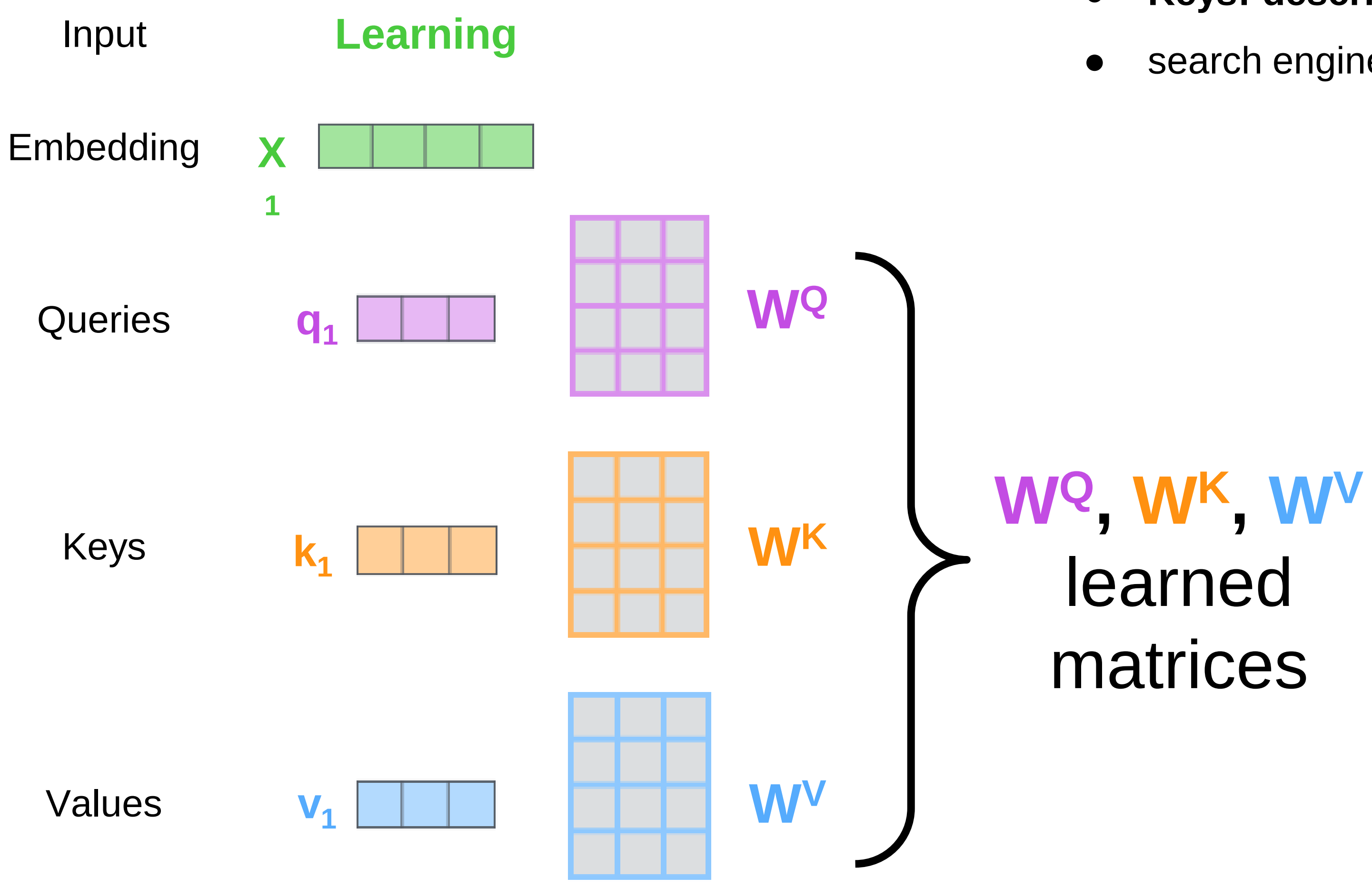
- **Quantum Self-Attention** in **Classical GANs** to boost performance in hybrid architecture

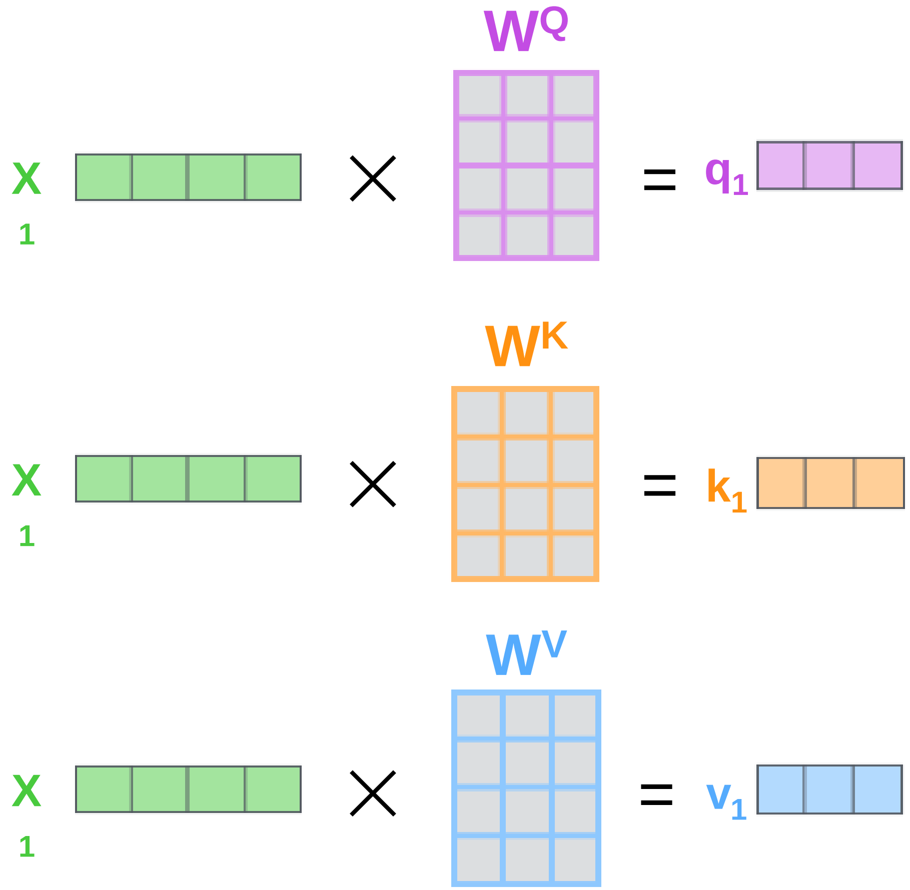
Classical Self-Attention

Query, Key, Value concept $\boxed{\rightarrow}$ analogous to retrieval systems

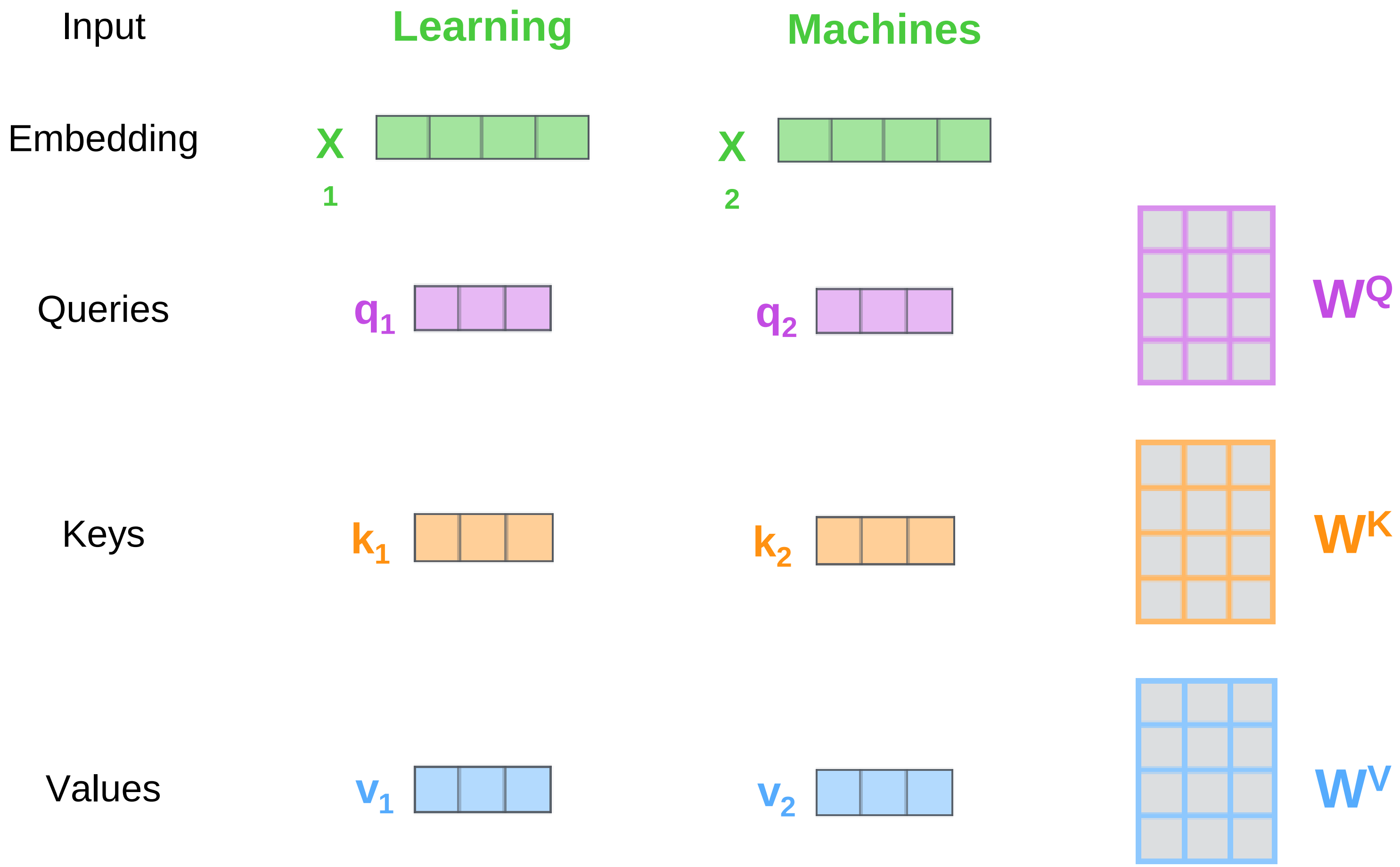
Example: When you searching for videos on YouTube's search engine

- search engine maps the **Query (text in the search bar)** against Keys
- **Keys: descriptors** (video title, description, etc.) of YouTube videos
- search engine returns the **best matched videos (Values)**





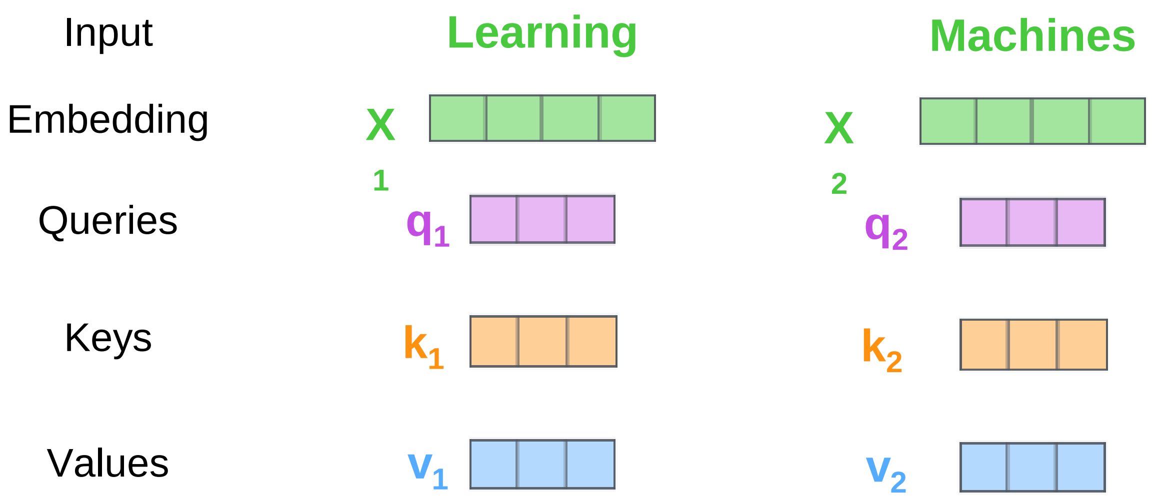
Classical Self-Attention



Adding more words adds more resulting vectors (while using the same learned matrices)

For now, let's only consider 2 words as our input: "Learning Machines"

Classical Self-Attention



Classical Self-Attention

Computing output for the word *Learning*

Input	Learning	Machines
Embedding	x_1	x_2
Queries	q_1	q_2
Keys	k_1	k_2
Values	v_1	v_2
Score	$q_1 \cdot k_1 = 112$	$q_1 \cdot k_2 = 96$

Classical Self-Attention

d_{key} = dimension of key vector
- Leads to more stable gradients
- Hyperparameter (!!!)
- Other values may be used

$$\downarrow$$

$\sqrt{d_{key}} = \sqrt{64} = 8$

(8 is the value used in *Attention Is All You Need* (2017))

Computing output for the word *Learning*

Input	Learning	Machines
Embedding	x_1	x_2
Queries	q_1	q_2
Keys	k_1	k_2
Values	v_1	v_2
Score	$q_1 \bullet k_1 = 112$	$q_1 \bullet k_2 = 96$
Score ÷ $\sqrt{d_{key}}$	$112 / \sqrt{d_{key}} = 14$	$96 / \sqrt{d_{key}} = 12$

Classical Self-Attention

d_{key} = dimension of key vector
- Leads to more stable gradients
- Hyperparameter (!!!)
- Other values may be used

$\downarrow$

$\sqrt{d_{key}} = \sqrt{64} = 8$

(8 is the value used in the original self-attention paper)

Computing output for the word *Learning*

Input	Learning	Machines
Embedding	x_1	x_2
Queries	q_1	q_2
Keys	k_1	k_2
Values	v_1	v_2
Score	$q_1 \cdot k_1 = 112$	$q_1 \cdot k_2 = 96$
Score $\div$ $\sqrt{d_{key}}$	$112 / \sqrt{d_{key}} = 14$	$96 / \sqrt{d_{key}} = 12$
Softmax	$s_1 = 0.88$	$s_2 = 0.12$

Classical Self-Attention

d_{key} = dimension of key vector
- Leads to more stable gradients
- Hyperparameter (!!!)
- Other values may be used

$\boxed{\downarrow}$

$\sqrt{d_{key}} = \sqrt{64} = 8$

(8 is the value used in the original self-attention paper)

Computing output for the word *Learning*

Input	Learning	Machines
Embedding	X_1	X_2
Queries	q_1	q_2
Keys	k_1	k_2
Values	v_1	v_2
Score	$q_1 \cdot k_1 = 112$	$q_1 \cdot k_2 = 96$
Score $\div \sqrt{d_{key}}$	$112 / \sqrt{d_{key}} = 14$	$96 / \sqrt{d_{key}} = 12$
Softmax	$s_1 = 0.88$	$s_2 = 0.12$
Softmax • Value	$m_1 = s_1 \cdot v_1 =$	$m_2 = s_2 \cdot v_2 =$

Classical Self-Attention

d_{key} = dimension of key vector
- Leads to more stable gradients
- Hyperparameter (!!!)
- Other values may be used

$\downarrow$

$\sqrt{d_{key}} = \sqrt{64} = 8$

(8 is the value used in the original self-attention paper)

Computing output for the word *Learning*

Input	Learning	Machines
Embedding	X_1	X_2
Queries	q_1	q_2
Keys	k_1	k_2
Values	v_1	v_2
Score	$q_1 \bullet k_1 = 112$	$q_1 \bullet k_2 = 96$
Score $\div \sqrt{d_{key}}$	$112 / \sqrt{d_{key}} = 14$	$96 / \sqrt{d_{key}} = 12$
Softmax	$s_1 = 0.88$	$s_2 = 0.12$
Softmax • Value	$m_1 = s_1 \bullet v_1 =$	$m_2 = s_2 \bullet v_2 =$
Sum	$z_1 = m_1 + m_2 =$	

QUANTUM TECHNOLOGY INITIATIVE

Classical Self-Attention

d_{key} = dimension of key vector
- Leads to more stable gradients
- Hyperparameter (!!!)
- Other values may be used

$\downarrow$

$\sqrt{d_{key}} = \sqrt{64} = 8$

(8 is the value used in the original self-attention paper)

This is only for the word *Learning*!

Input	**Learning**	**Machines**
Embedding	x_1	x_2
Queries	q_1	q_2
Keys	k_1	k_2
Values	v_1	v_2
Score	$q_1 \cdot k_1 = 112$	$q_1 \cdot k_2 = 96$
Score $\div \sqrt{d_{key}}$	$112 / \sqrt{d_{key}} = 14$	$96 / \sqrt{d_{key}} = 12$
Softmax	$s_1 = 0.88$	$s_2 = 0.12$
Softmax • Value	$m_1 = s_1 \cdot v_1 =$	$m_2 = s_2 \cdot v_2 =$
Sum	$z_1 = m_1 + m_2 =$	

Classical Self-Attention

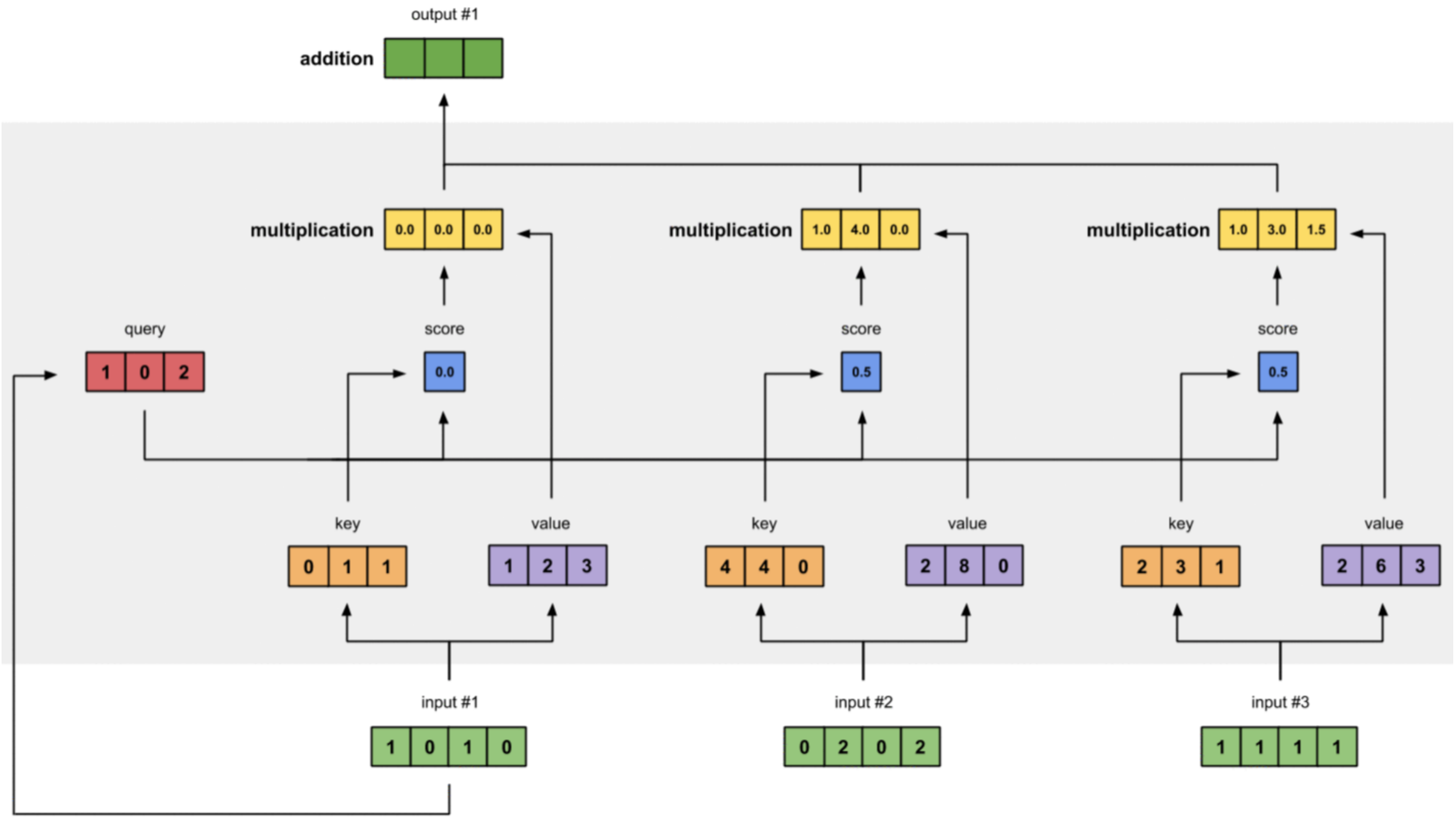
Input	Learning	Machines
Embedding	x_1	x_2
Queries	q_1	q_2
Keys	k_1	k_2
Values	v_1	v_2
Score	$\boxed{q_2 \cdot k_1 = 16}$	$\boxed{q_2 \cdot k_2 = 24}$
Score ÷ $\sqrt{d_{key}}$	$16 / \sqrt{d_{key}} = 2$	$24 / \sqrt{d_{key}} = 3$
Softmax	$s_1 = 0.27$	$s_2 = 0.73$
Softmax • Value	$m_1 = s_1 \cdot v_1 =$	$m_2 = s_2 \cdot v_2 =$
Sum		$\boxed{z_2 = m_1 + m_2 =}$

Doing this for the word *Machines* is just as easy

Creating a score with the query from *Machines*

Summing over each *m* from the new values

Classical Self-Attention: Another Look



Quantum Self-Attention

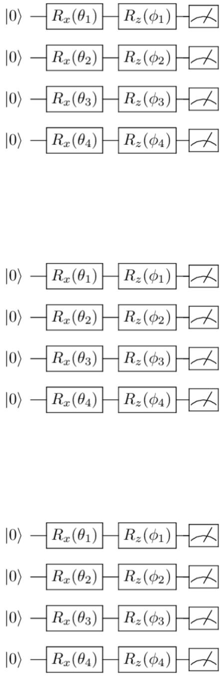
Input | Learning

Embedding x_1

Queries q_1

Keys k_1

Values v_1



W^Q

W^K

W^V

W^Q, W^K, W^V

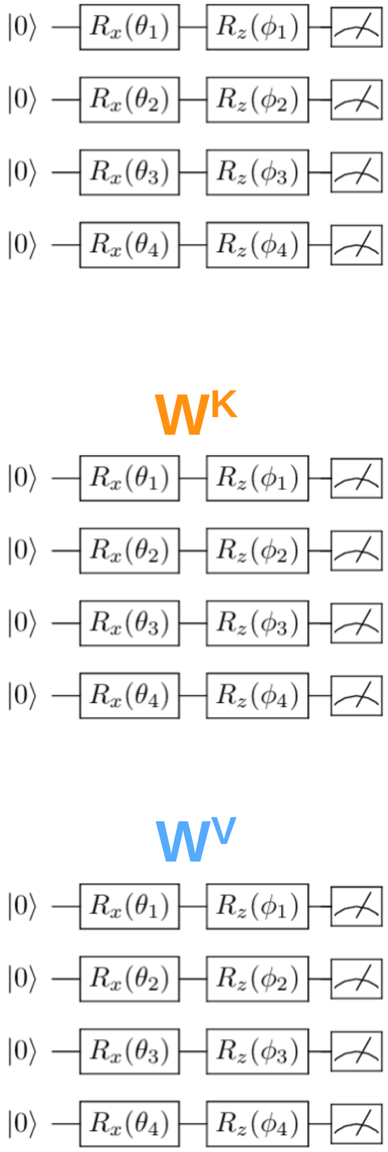
~~learned matrices~~

quantum circuits!

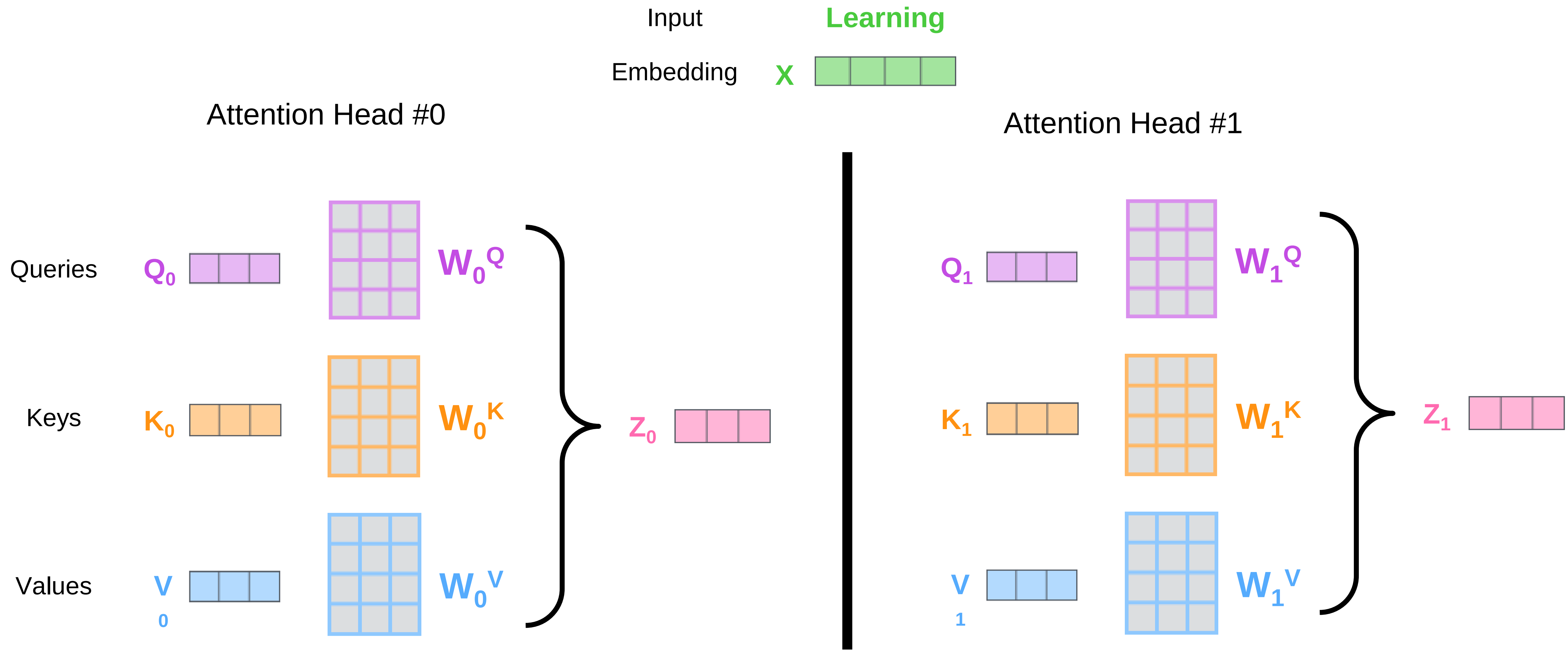
$x_1 \times W^Q = q_1$

$x_1 \times W^K = k_1$

$x_1 \times W^V = v_1$

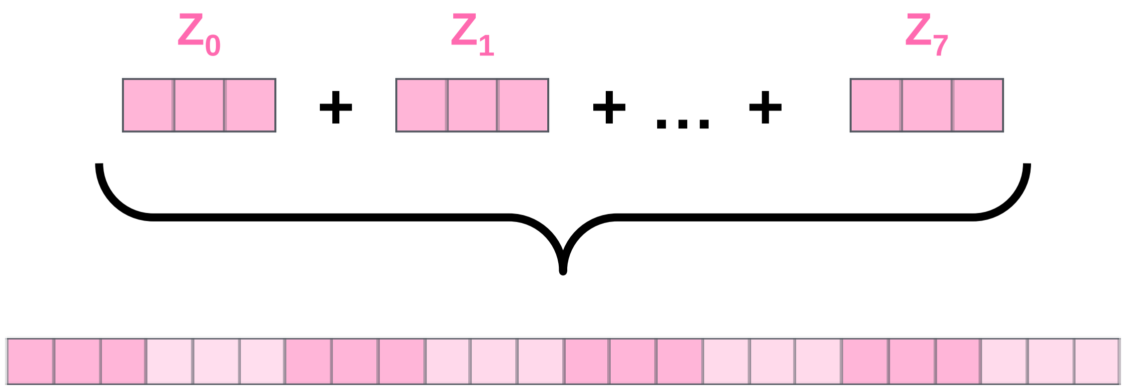


Multi-Head Attention

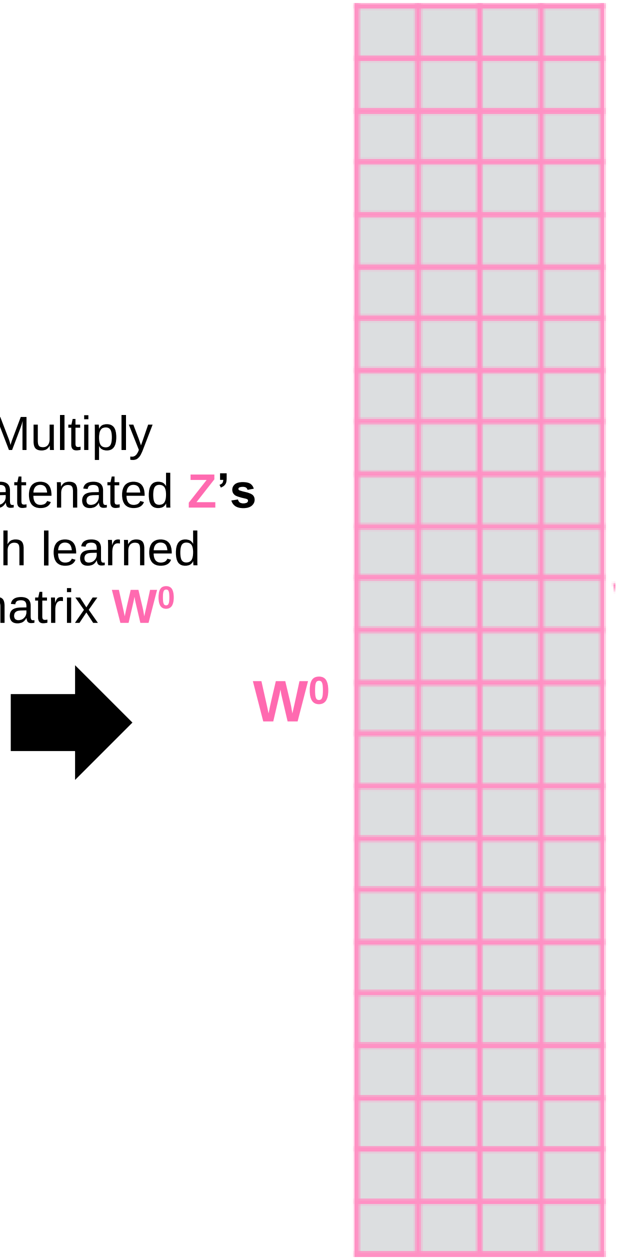


Multi-Head Attention

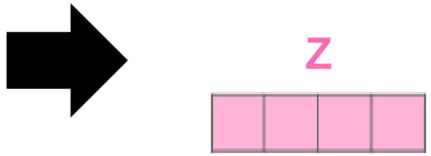
Concatenate Z from each attention head:



Multiply concatenated Z’s with learned matrix W^0



Resulting matrix Z contains information from all attention heads

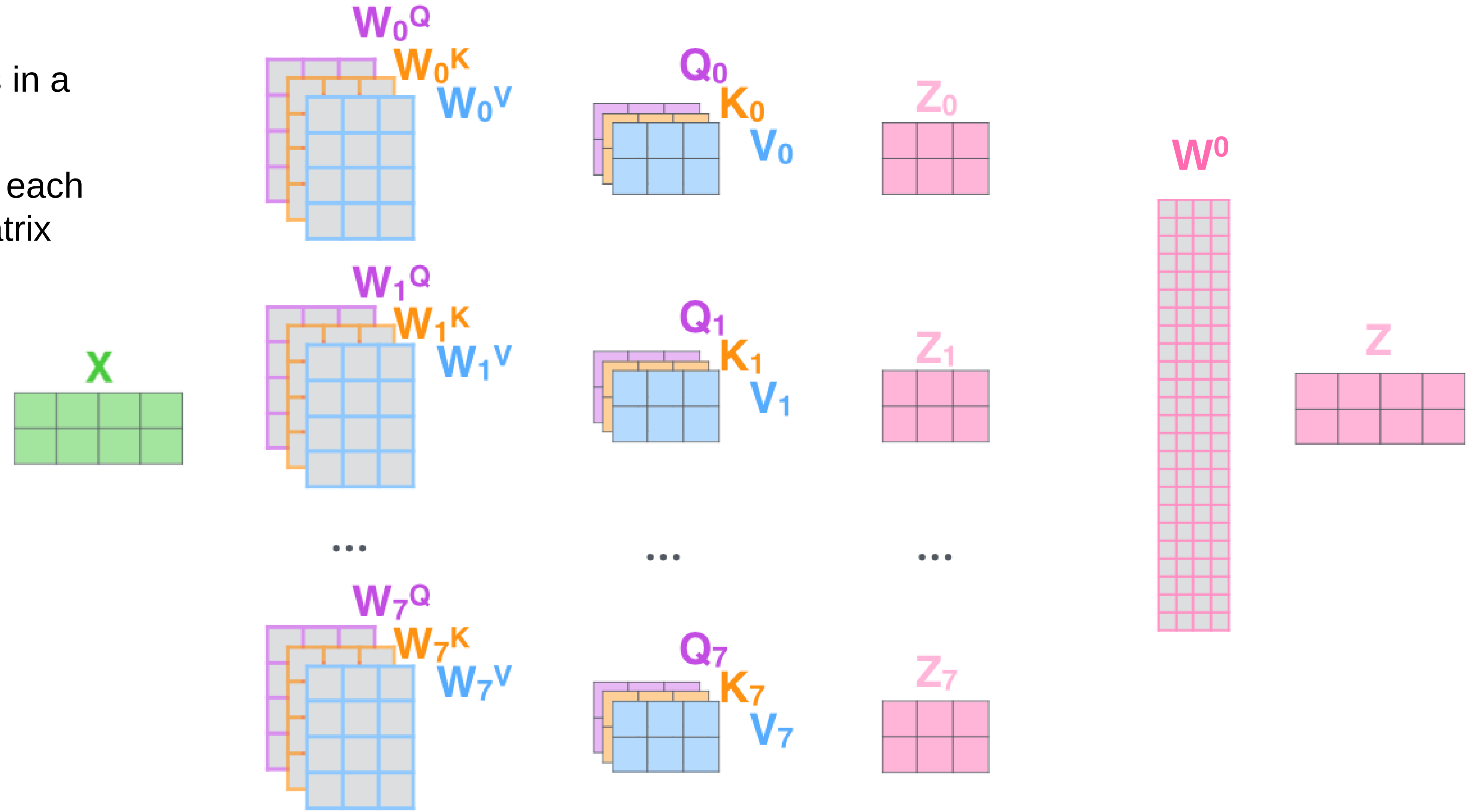


Efficient Multi-Head Attention

Stacking words results in a larger matrix

Allows for representing each input as a (larger) matrix

Learning Machines

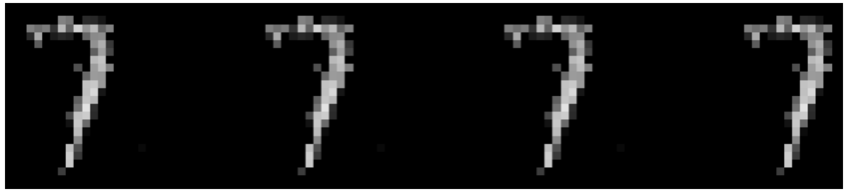


CERN Quantum Technology Initiative

Accelerating Quantum Technology Research and Applications

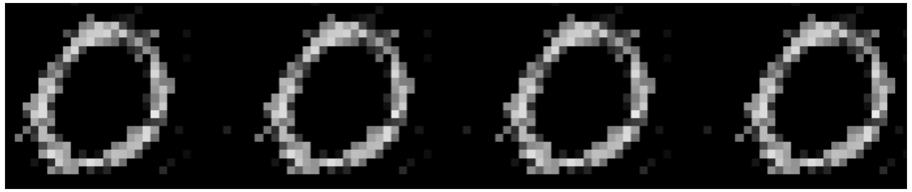
Thank You!

In < 50 epochs, we get (preliminary) results with MNIST dataset!



Dr. Sofia Vallecorsa

Dr. Michele Grossi



Computational Resources:

CERN openlab

Questions?