Generative AI

Weijie Zhao

11/21/2023

HW 5: Model Inference

- This homework does NOT have sample code
- Write a program for handwritten digit dataset
- Output a file with predictions for each data instance
- You are free to use any model or data to train your model offline
- We will only do model inference during the test
- Two scripts are mandatory:
  - compile.sh
  - run.sh <test_dataset> <output_file>

HW 5: Model Inference

- The testing dataset is a variation of mnist.t (added gaussian noise)
- We will have 10 cases where we vary the level of noises
  - N(0,0), N(0,1), N(0,2), ... N(0,9)
  - N(0,0) corresponds to no noises, i.e., plain mnist.t
- Each test case contains 10k instances.
- A test case is considered correct if the test accuracy is no less than 50%

HW 5: Model Inference

- No 3rd party code is allowed.
- 10 test cases. Each case weights 1 pt.
- The compilation is considered failed if it does not finish in 5 minute.
- A test case is considered incorrect if it does not finish in 2 minutes.
- Correct GPU solutions will get 5 pts bonus.
- The summation of the execution time across 10 cases will be used to rank correct solutions.

- Due: 11/30/2022 1:00 pm EST

Testing Environment

- ssh yourusername@granger.cs.rit.edu
- Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz
- 48 threads in total (2 sockets, 12 cores per socket, 2 threads per core)
- 251 GB memory
- GPU: Tesla P4

- Testing limit:
  - 8 threads taskset -c
  - 2 GPU

Testing Environment

- ssh yourusername@granger.cs.rit.edu
- Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz
- 48 threads in total (2 sockets, 12 cores per socket, 2 threads per core)
- 251 GB memory
- GPU: Tesla P4

- Testing limit:
  - 8 threads taskset -c
  - 2 GPU

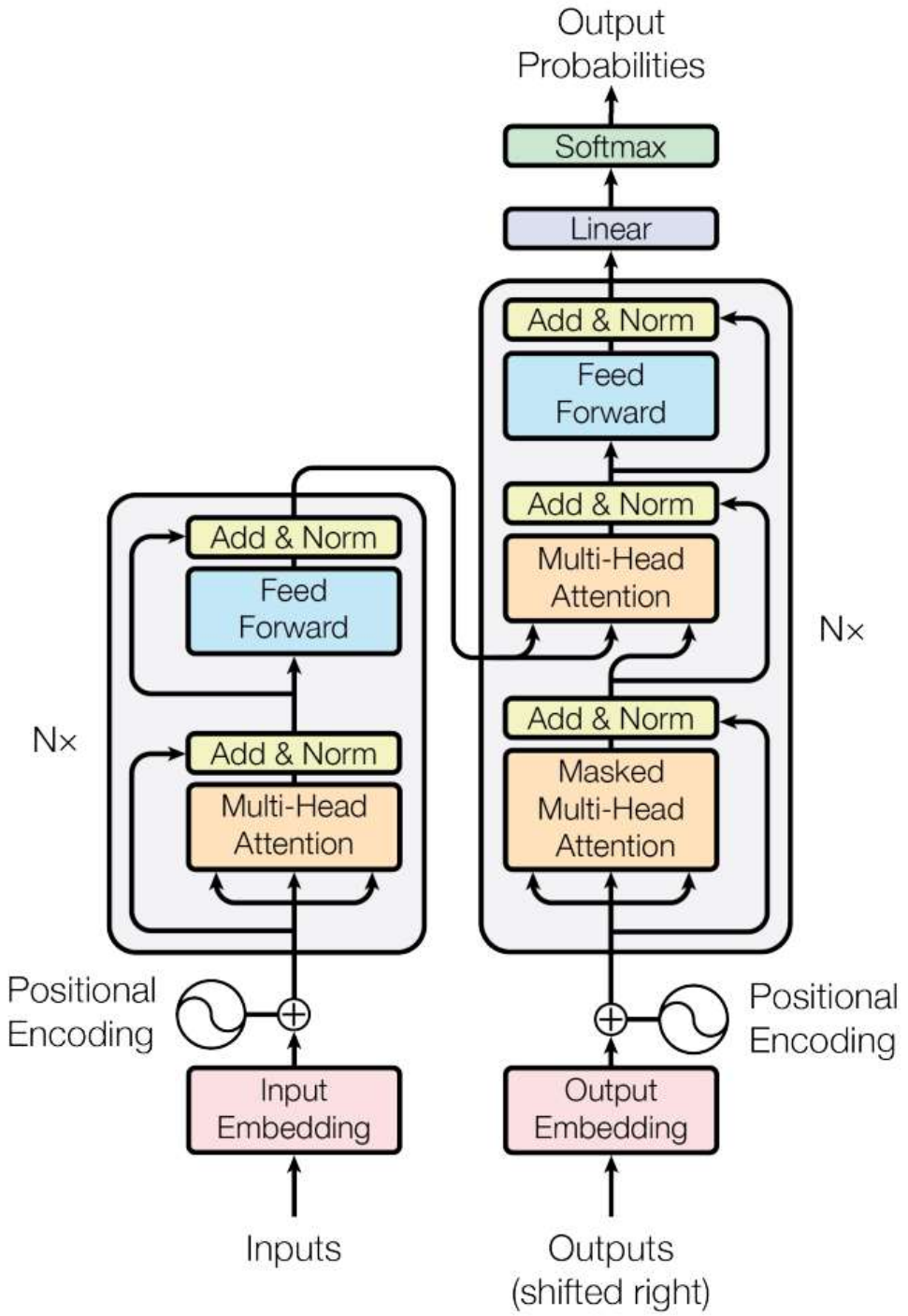
Transformers

- Attention is All You Need
- Attention
- Self-attention
- Masked self-attention
- Positional encoding

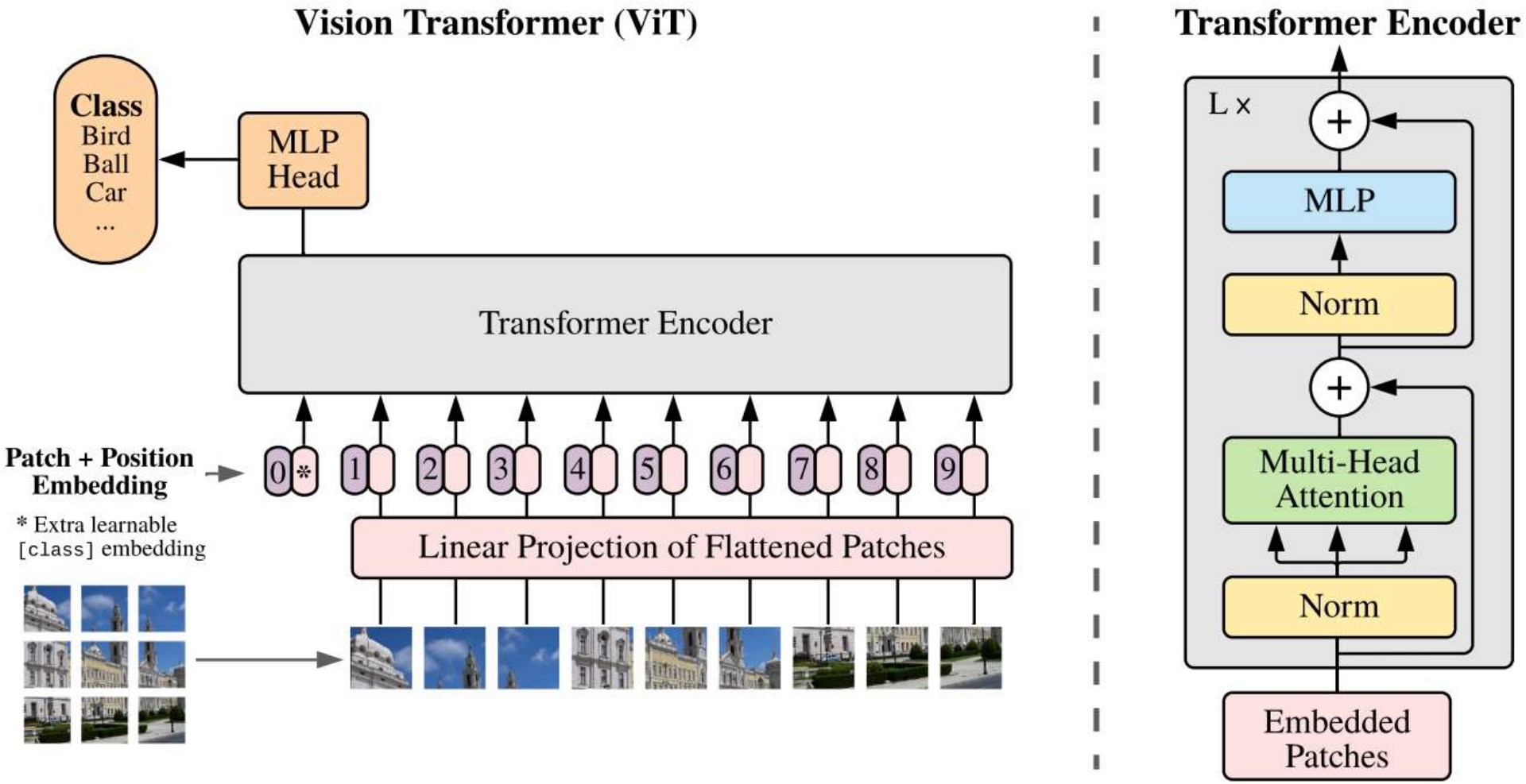
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

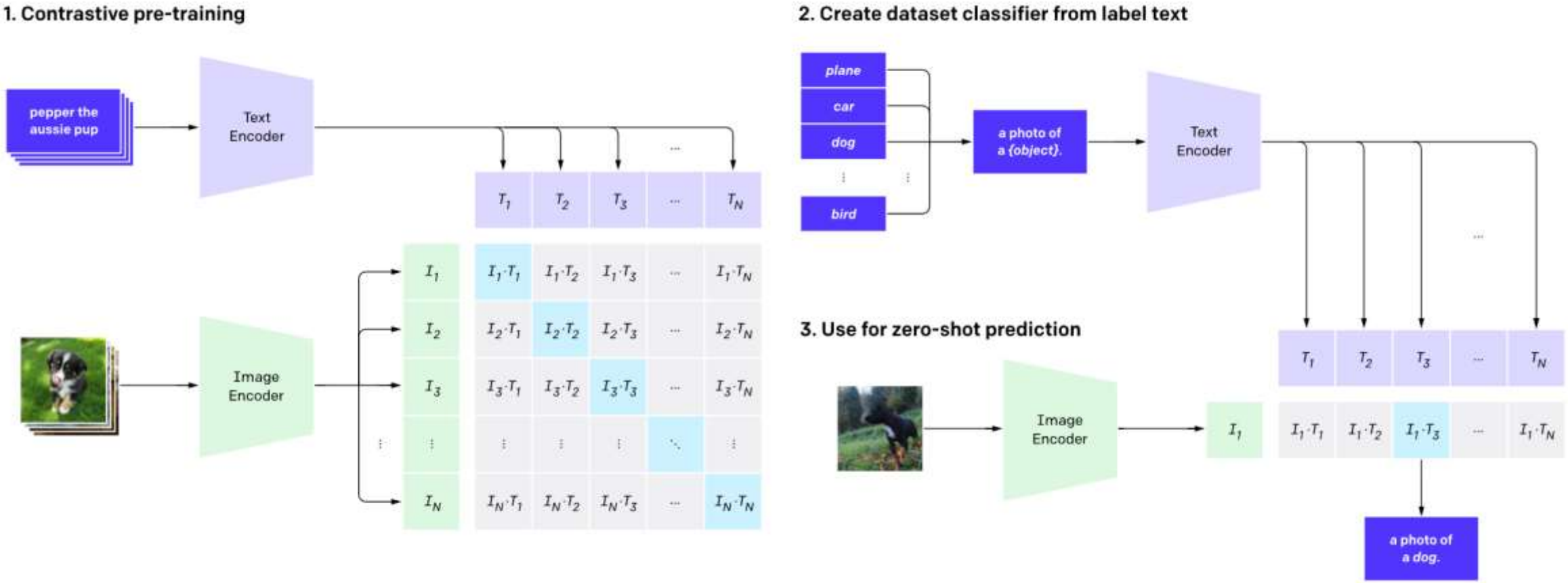


Vision Transformer



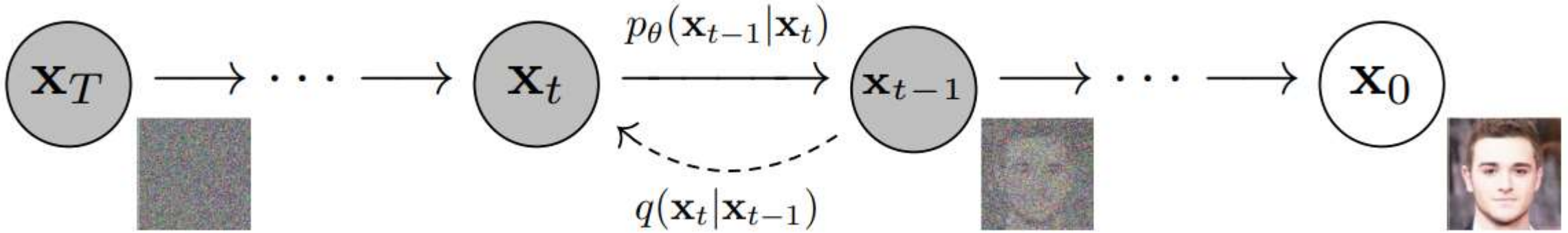
https://arxiv.org/pdf/2010.11929v2.pdf

CLIP



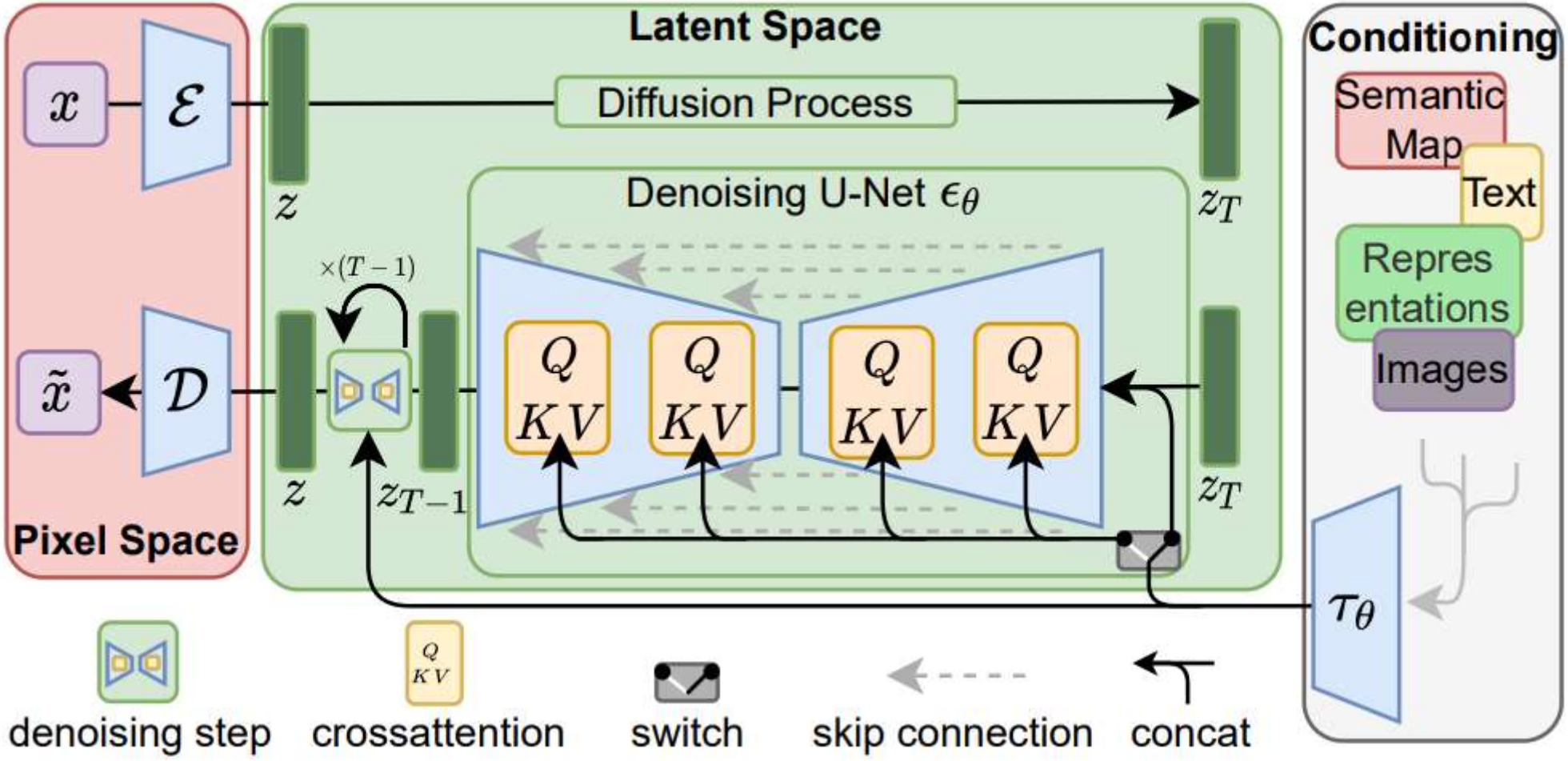
https://openai.com/research/clip

Diffusion Models



https://arxiv.org/pdf/2006.11239.pdf

Stable Diffusion



https://arxiv.org/pdf/2112.10752.pdf