

Attention Layer with SiTo

Partition the Image into patches of size $s \times s$

1	2	3	4
5	6	7	8

1	2	3	4	5	6	7	8
---	---	---	---	---	---	---	---

(a) Base Token Selection

1	2	3	4	5	6	7	8
---	---	---	---	---	---	---	---

(b) Pruned Token Selection

①	2	③	4	⑤	6	7	8
---	---	---	---	---	---	---	---

Self-Attention

①	2	③	4	⑤	6	7	8
---	---	---	---	---	---	---	---

(c) Pruned Token Recovery

4	2	⑤	4	⑤	6	7	8
---	---	---	---	---	---	---	---

Cross-Attention

FFN

- ① Compute the Cosine Similarity $CosSim$ between N tokens.

1	1.0	0.6	0.4	0.9	0.6	0.7	0.3	0.8
2	0.6	1.0	0.5	0.5	0.7	0.3	0.5	0.6
3	0.4	0.5	1.0	0.7	0.9	0.8	0.3	0.8
4	0.9	0.5	0.7	1.0	0.7	0.6	0.3	0.8
5	0.6	0.7	0.9	0.7	1.0	0.8	0.2	0.6
6	0.7	0.3	0.8	0.6	0.8	1.0	0.1	0.3
7	0.3	0.5	0.3	0.3	0.2	0.1	1.0	0.4
8	0.8	0.6	0.8	0.8	0.6	0.3	0.4	1.0
1	2	3	4	5	6	7	8	

Sum

- ② $SimScore = Sum(CosSim, dim = 1)$

5.3	4.7	5.4	5.5	5.5	4.6	3.1	5.3
-----	-----	-----	-----	-----	-----	-----	-----

- ③ Add the *Guassain Noise*

0.2	0.1	0.3	0.4	0.2	0.1	0.3	0.2
-----	-----	-----	-----	-----	-----	-----	-----

- ④ $Noise\ SimScore = SimScore + Noise$

5.5	4.8	5.7	5.9	5.7	4.7	3.4	5.5
1	2	3	4	5	6	7	8

- ⑤ Find the **maximum value** in each patch.

5.5	4.8	5.7	5.9
5.7	4.7	3.4	5.5

- ⑥ Get the **Base Tokens** from each patch (○).

1	2	3	4
⑤	6	7	8

- ① Get the **Similarity** between the **Base Tokens** and others.

	1	2	3	6	7	8
4	0.9	0.5	0.7	0.6	0.3	0.8
5	0.6	0.7	0.9	0.8	0.2	0.6

Max

- ② Fine the maximum **value** and corresponding index.

1	2	3	6	7	8
0.9	0.7	0.9	0.8	0.3	0.8
4	5	5	5	4	4

- ③ Select the Top K Values.

1	3
4	5

- ④ Get the **Pruned Tokens** (○).

①	2	③	4
⑤	6	7	8

- ① Get the output of the attention layer.

①	2	③	4	⑤	6	7	8
①	2	③	4	⑤	6	7	8

Copy

- ② Recover the pruned tokens from base tokens.

①	2	③	4	⑤	6	7	8
①	2	③	4	⑤	6	7	8

- ③ Get the final outputs.

4	2	⑤	4
⑤	6	7	8