HW 10.

Yuanteng Chen

3. Vision Transformer

(a) ① Vision transformer

→ Encoder-style transformer.

② Language Transformer.

→ Decoder-style transformer.

Because we want to produce an image representation using an encoder transformer and generate tokens from image representation using an decoder transformer

(b): (i) Each column creates a query while each row creates a key and a value.

(ii).

	<SOS>	a	mountain	range	<PAD>
<SOS>					X
a	X				X
mountain	X	X			X
range	X	X	X		X
<PAD>	X	X	X	X	X
<ENC1>					X
<ENC2>					X
<ENC3>					X

c C) : What is the Big-O runtime complexity of the attension operation after the modification? c Assume each window consists of K by K patches)

Solution:

$(\frac{H}{P})^2$ patches in total, but each patch only attends k^2 patches.

So. $O(\frac{H^2}{P^2} \cdot k^2 \cdot D)$

Corigin complexity : $(\frac{H^2}{P^2} \cdot \frac{H^2}{P^2} \cdot D)$