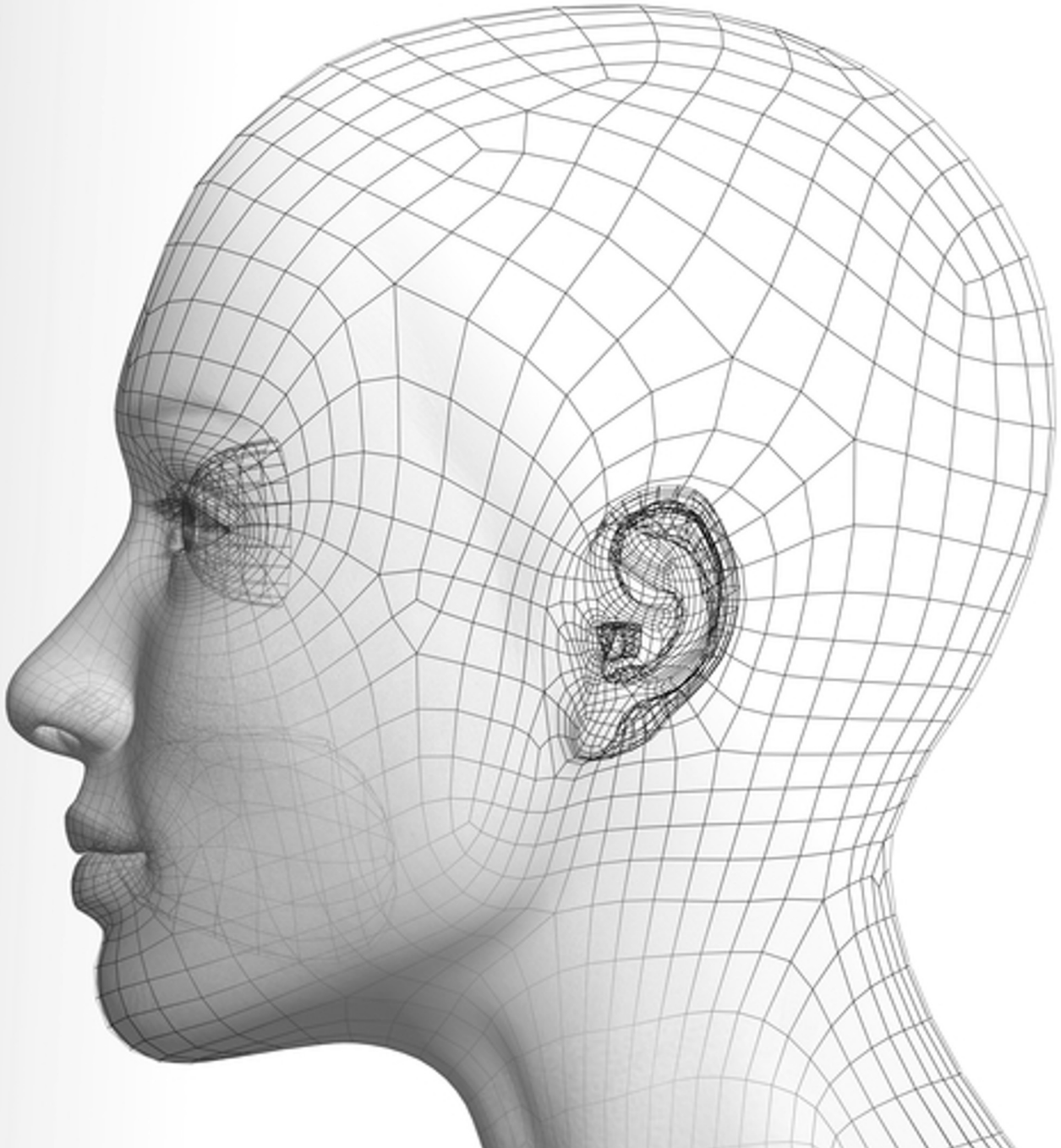
Tutorial 9
Attention

Xingang Pan
潘新钢

https://xingangpan.github.io/

https://twitter.com/XingangP



Question 1

The tutorial provides a simple walkthrough of the Vision Transformer. We hope you will be able to understand how it works by looking at the actual data flow during inference.

t9q1.ipynb

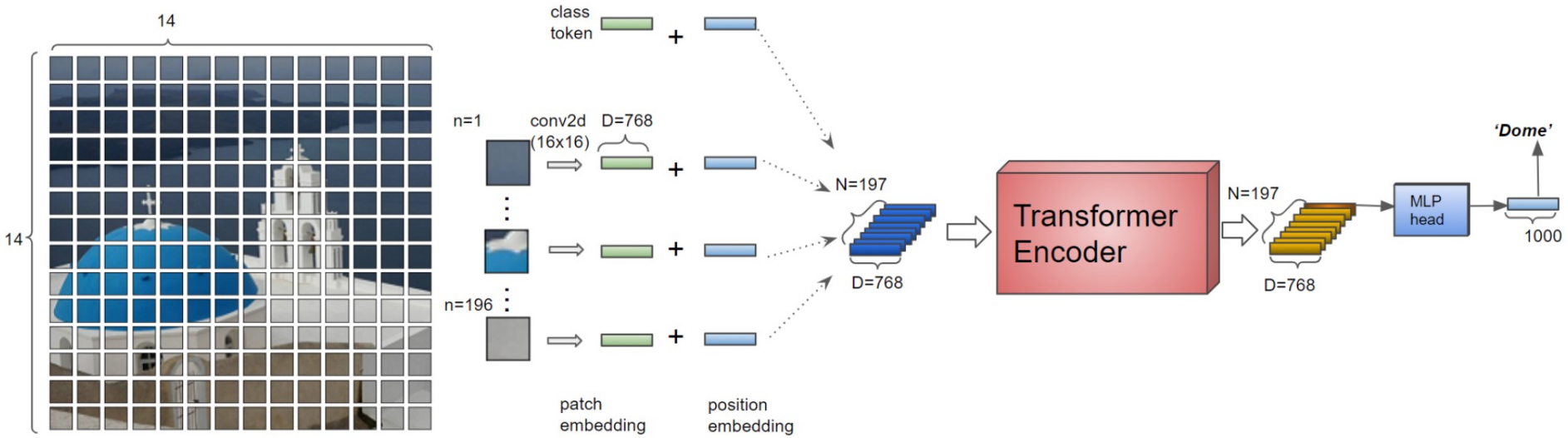


Figure 1. Vision Transformer inference pipeline.

1. Split Image into Patches
   The input image is split into 14 x 14 vectors with dimension of 768 by Conv2d (k=16x16) with stride=(16, 16).
2. Add Position Embeddings
   Learnable position embedding vectors are added to the patch embedding vectors and fed to the transformer encoder.
3. Transformer Encoder
   The embedding vectors are encoded by the transformer encoder. The dimension of input and output vectors are the same. Details of the encoder are depicted in Fig. 2.
4. MLP (Classification) Head
   The 0th output from the encoder is fed to the MLP head for classification to output the final classification results.

Question 1

t9q1.ipynb