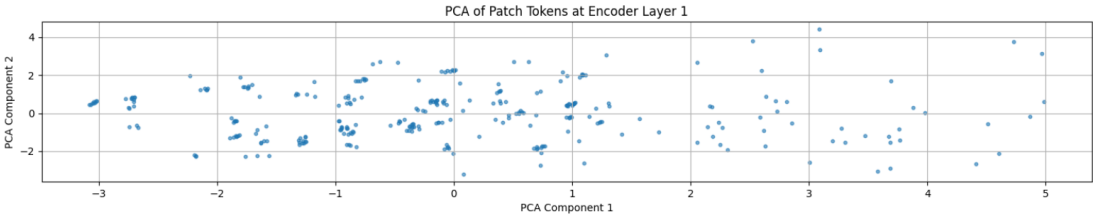
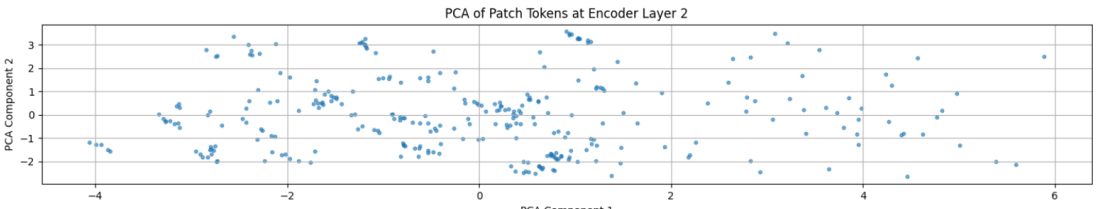
Update on Vision Transformer (01)

Sai Krishna Shanigarapu
Indian Institute of Technology Hyderabad

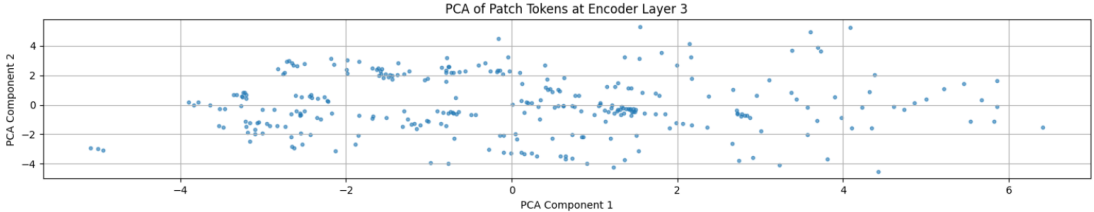
June 9, 2025

Intermediate patch embeddings between layers





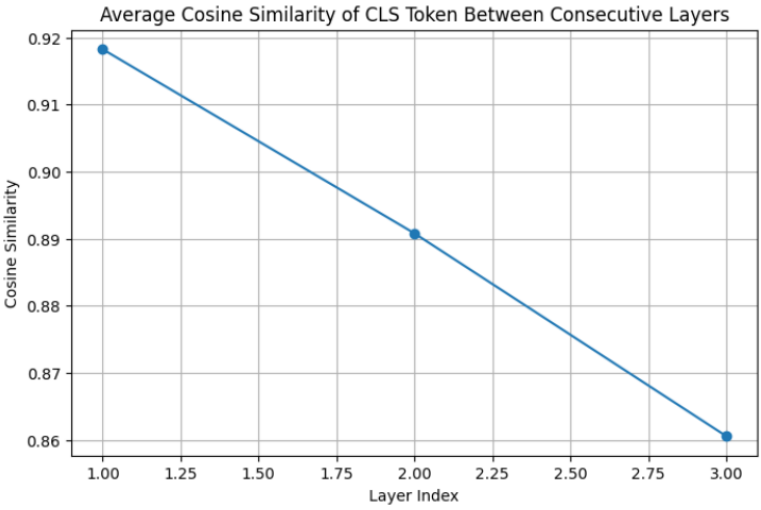
Cont...


PCA of Patch Tokens at Encoder Layer 3
PCA Component 2
PCA Component 1
PCA of Patch Tokens at Encoder Layer 4
PCA Component 2
PCA Component 1

Cont...

- As the transformer layers process the patch tokens, their internal representation changes. Which is nothing but the model is learning higher level features from a raw pixel input.
- The shift in the PCA space indicates that patches are evolving.
- shift along the x axis indicates there is largest variation in patch representation along x axis.
- This consistent shift in patch token PCA distribution implies that the Vision transformer is learning. (layer by layer)

Cosine Similarity of CLS token b/w consecutive layers



Cont...

- Cosine similarity is a measure of how similar two vectors are (in direction sense).
- From the graph we observe that the Cosine similarity decreases as the layers increase.
- at layer index 1, the cosine similarity is 0.92 (around 1) - thus there isn't much change in the CLS token. However as the layer index increases cosine similarity decreases (changing) indicating there are some changes made to CLS - CLS token is being updated accordingly.
- The more the number of layers, the better the result will be.