Université Grenoble Alpes

LIG

Vision Transformer

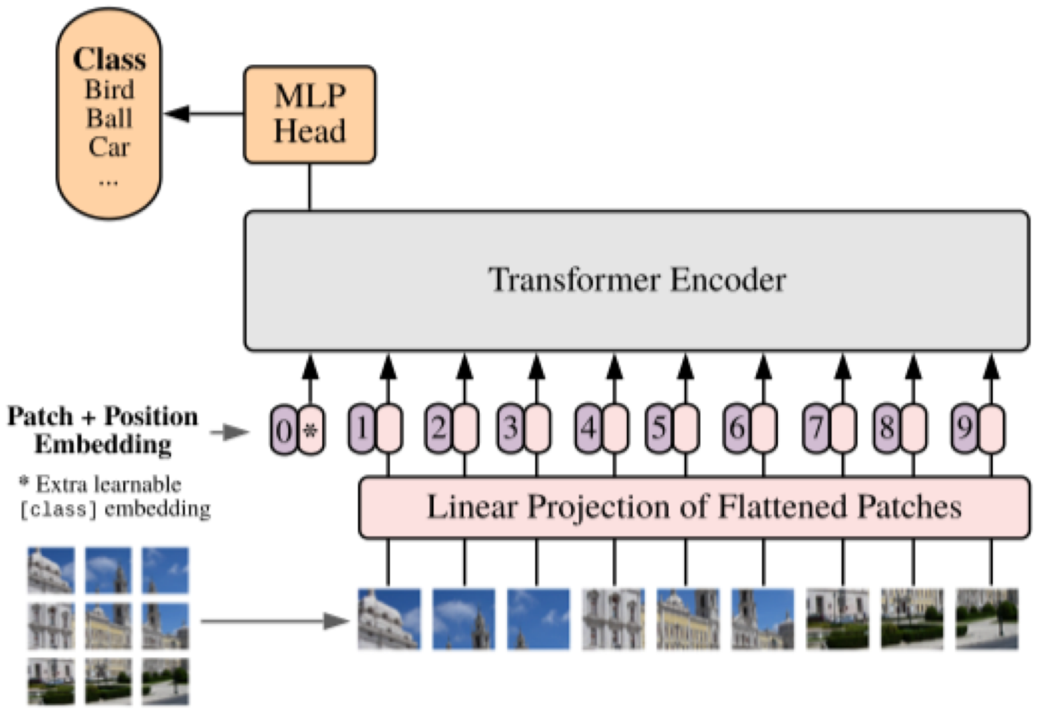
Yangtao Wang
James Crowley

Outline

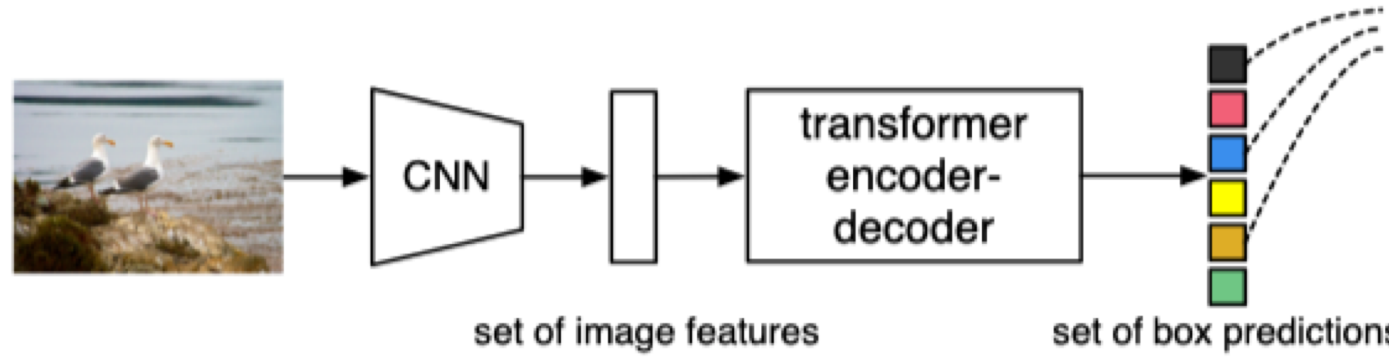
- Introduction
- Embeddings
  - Image embeddings
  - Positional embeddings
- Efficient attention mechanism
- Self-supervised method

Transformer Architecture

Vision Transformer (ViT)



DETR



Only encoder

Classification: ViT[1], DeiT, PVT, MsViT, Swin-T
Object detection: PVT
Segmentation: DINO

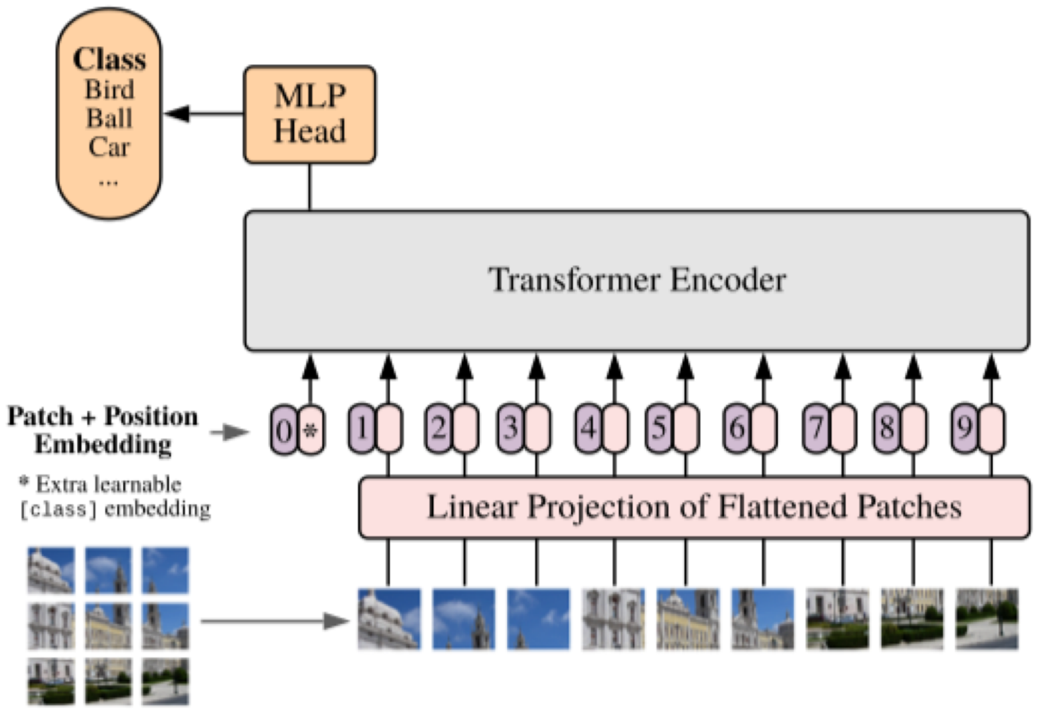
Encoder-decoder

Object detection: DETR[2], deformable DETR
Tracking: TransCenter, …

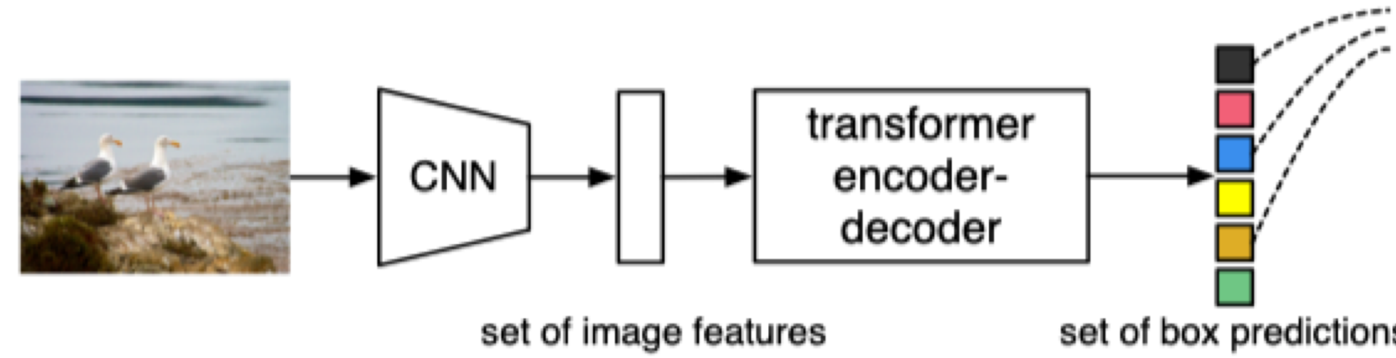
[1]Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Tran
[2]Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020, August). End-to-end object detection with transformers. In *European Conference on Computer Vision* (p

Image Embeddings

Vision Transformer (ViT)



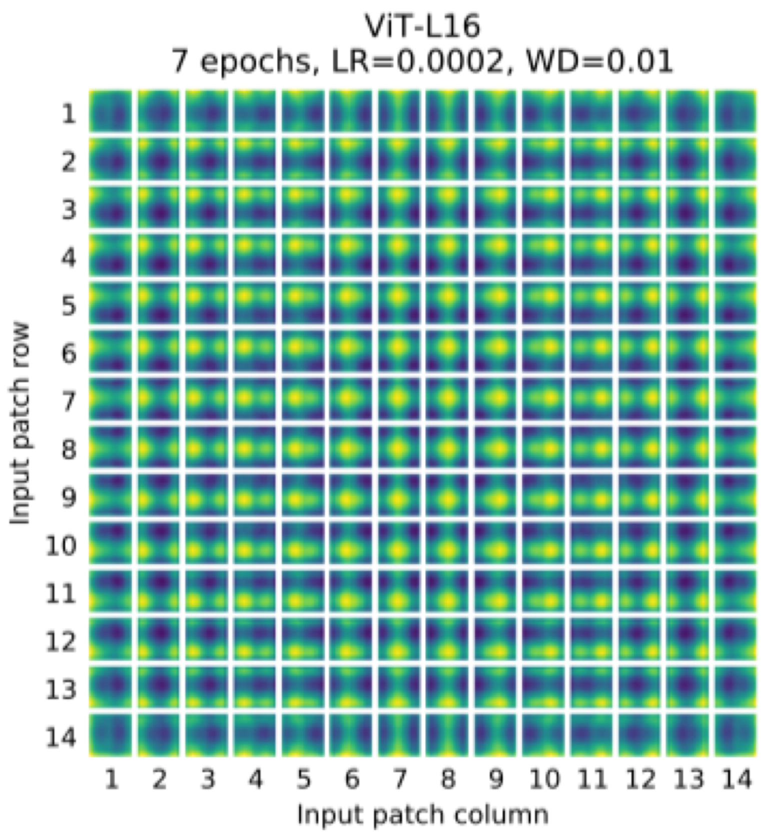
DETR



Method 1: Splitting raw image into patches of 4x4, 8x8, 16x16, 32x32.
Method 2: Using CNN feature map.

[1]Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: T
[2]Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020, August). End-to-end object detection with transformers. In *European Conference on Computer Vision*

Positional Embeddings



1) **Learnable absolute 1D positional embeddings (ViT, DeiT[3])**
   Encode the inputs as a sequence of patches in the raster order.

2) **Learnable 2D positional embeddings (MsViT[4])**
   Encode the inputs as a grid of patches in two dimensions.

3) **Relative positional Embeddings (SwinT[5])**
   Encode the relative distance between patches.

$$\text{Attention}(Q, K, V) = \text{SoftMax}(QK^T/\sqrt{d} + B)V,$$

[3] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2020). Training data-efficient image transformers & distillation through attention. arXiv prepri
[4] Zhang, P., Dai, X., Yang, J., Xiao, B., Yuan, L., Zhang, L., and Gao, J. (2021). Multiscale vision longformer: A new vision transformer for high-resolution image encoding. ar
[5] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv

Positional Embeddings

Different adding strategies:

1) Add positional embeddings to the inputs before feeding the inputs to the Transformer encoder. (e.g. ViT, DeiT)
2) Learn and add positional embeddings to the inputs at the beginning of each layer (e.g. PVT[6])
3) Add a learned positional embeddings to the inputs at the beginning of each layer (shared between layers).

Pos. Emb.	Default/Stem	Every Layer	Every Layer-Shared
No Pos. Emb.	0.61382	N/A	N/A
1-D Pos. Emb.	0.64206	0.63964	0.64292
2-D Pos. Emb.	0.64001	0.64046	0.64022
Rel. Pos. Emb.	0.64032	N/A	N/A

Table 8: Results of the ablation study on positional embeddings with ViT-B/16 model evaluated on ImageNet 5-shot linear.

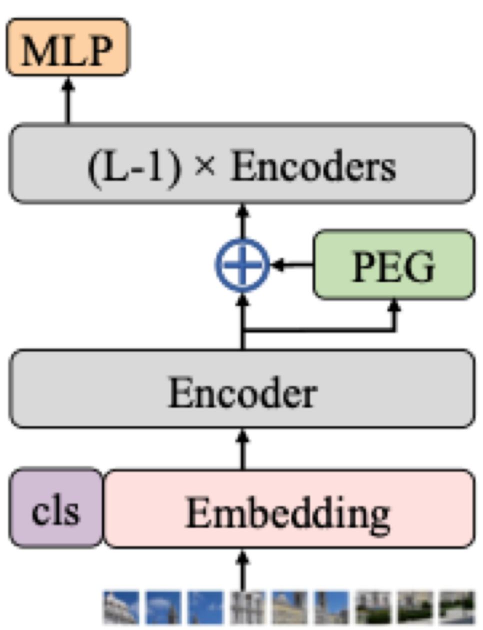
[6] Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., and Shao, L. (2021). Pyramid vision transformer: A versatile backbone for dense prediction withou

Problem for fixed size embeddings

Problem: If we want to fine-tune the model on higher resolution images, the pre-trained position embeddings may no longer be meaningful.

Method 1: Performing a 2D interpolation(bicubic) of the pre-trained position embeddings, according to their location in the original image, while keeping the patch size the same.

Method 2: Conditional Positional encoding (CPE[7])



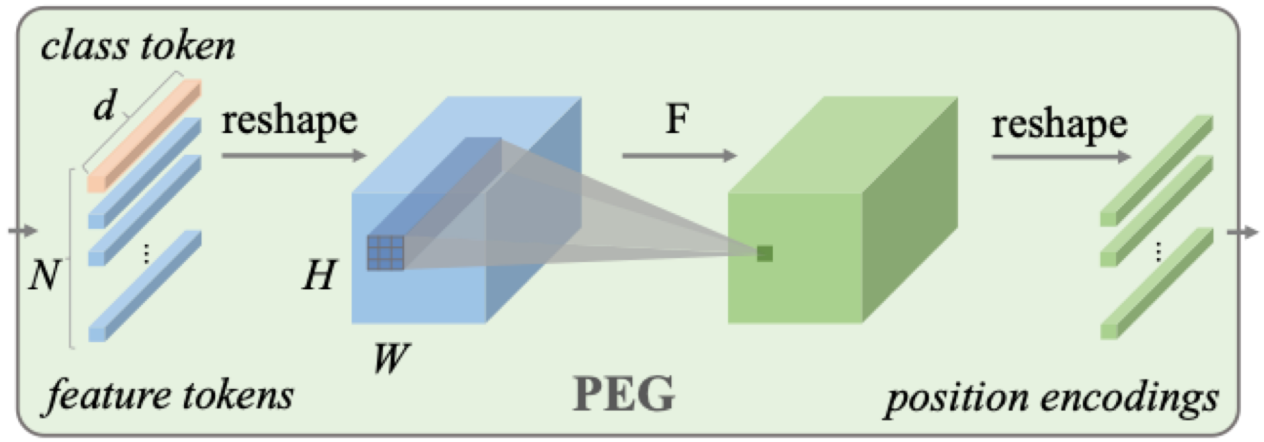
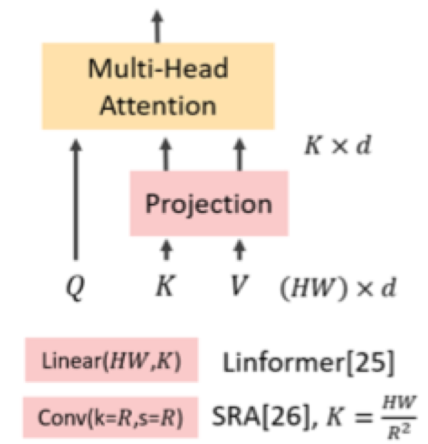


Figure 2. Schematic illustration of Positional Encoding Generator (PEG). Note d is the embedding size, N is the number of tokens. The function $\mathcal{F}$ can be depth-wise, separable convolution or other complicated blocks.

[7] Chu, X., Zhang, B., Tian, Z., Wei, X., & Xia, H. (2021). Do We Really Need Explicit Position Encodings for Vision Transformers?. *arXiv preprint arXiv:2102.10882*.

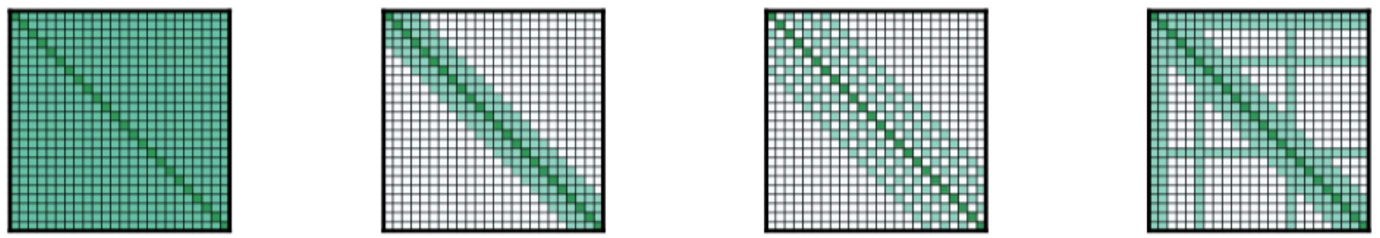
Efficient attention mechanism

1) Low rank based method: PVT[6]



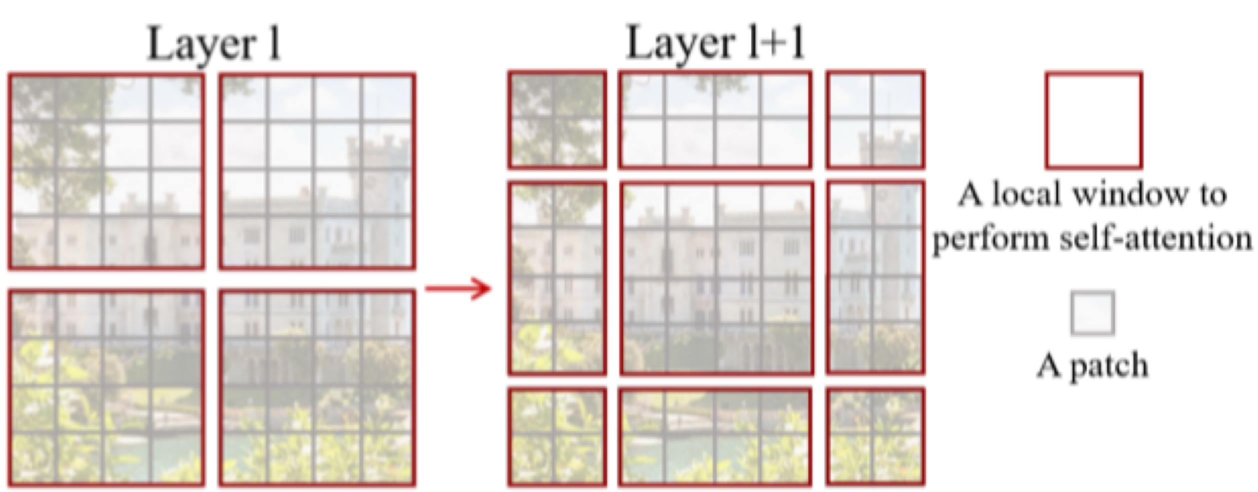
Low-rank based efficient Attention

2) Sparse attention mechanism:
Axial transformer, Longformer[8]



(a) Full n^2 attention (b) Sliding window attention (c) Dilated sliding window (d) Global+sliding window

3) Shifted window attention:
Swin Transformer[5]



[5] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2
[6] Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., and Shao, L. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without c
[8] Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150.

Self-supervised learning

1) BERT like strategy: SiT[9]

Task 1: Image Reconstruction(Inspired by MLM)
Random drop, random replacement, colour distortion, blurring, grey-scale

Task 2: Rotation Prediction
Rotate image by 0°,90°,180°,270°, classify the rotation by rotation token

Task 3: Contrastive learning (Inspired by next sentence prediction)
Given half of the negative samples from other image and adopt cosine similarity by using contrastive

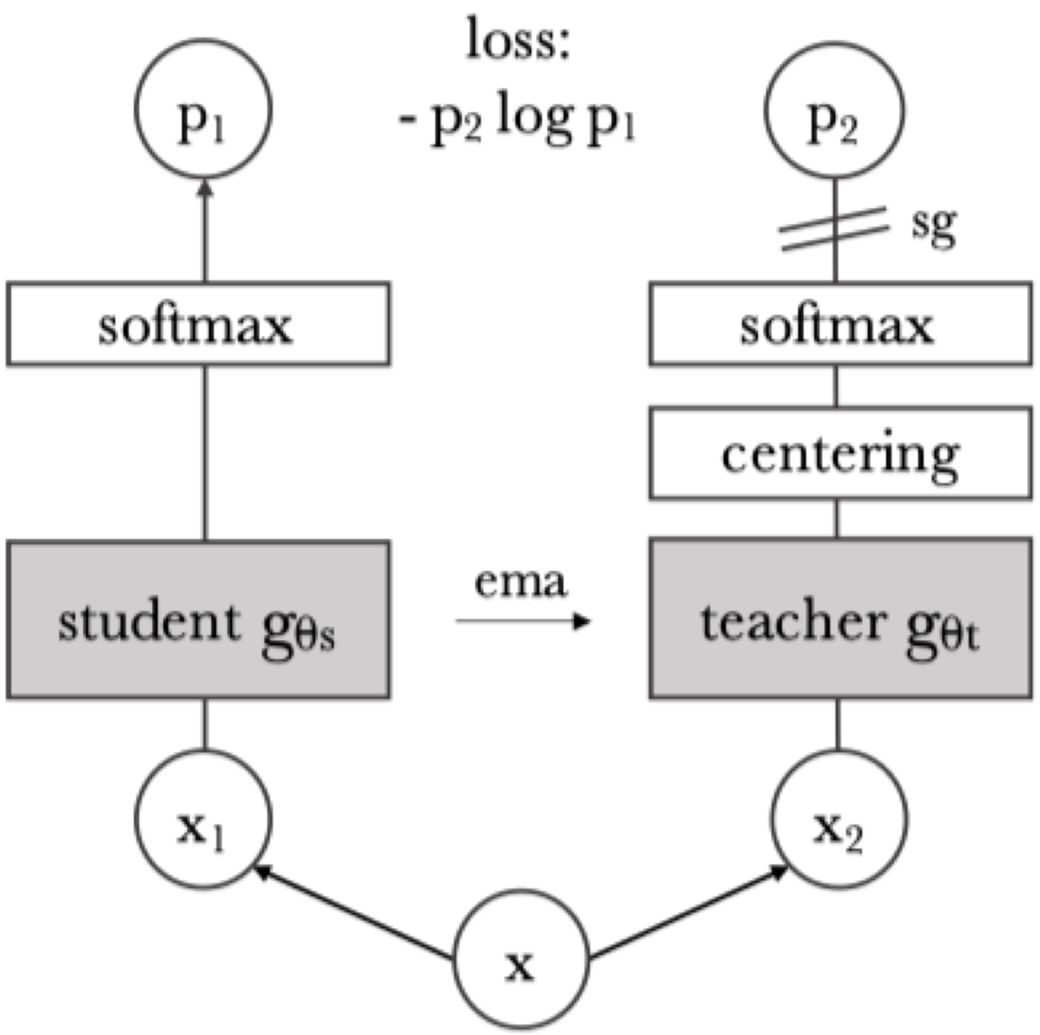
Original Image | Random Drop | Random Replace | Colour Distortion | Blurring | Grey-scale

[9] Atito, S., Awais, M., & Kittler, J. (2021). SiT: Self-supervised vIsion Transformer. *arXiv preprint arXiv:2104.03602*.

Self-supervised learning

1) Student-teacher based strategy: DINO[10]
Dynamic teacher network updated
by previous student weights with momentum encoder

Figure: Self-attention from a Vision Transformer with 8x8 patches trained with no supervision



[10] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging Properties in Self-Supervised Vision Transformers. *arXiv preprint arXiv:2104.14294*.

Transformer survey papers

1) Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2021). Transformers in Vision: A Survey. *arXiv preprint arXiv:2101.01169*.

2) Tay, Y., Dehghani, M., Bahri, D., and Metzler, D. (2020b). Efficient transformers: A survey. arXiv preprint arXiv:2009.06732.

Thank you