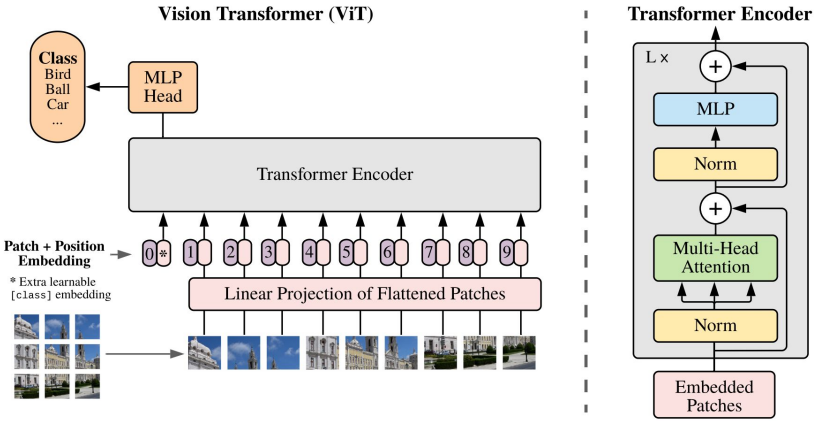
Vision Transformers and Self-supervised learning

Elena Izzo
elena.izzo@phd.unipd.it

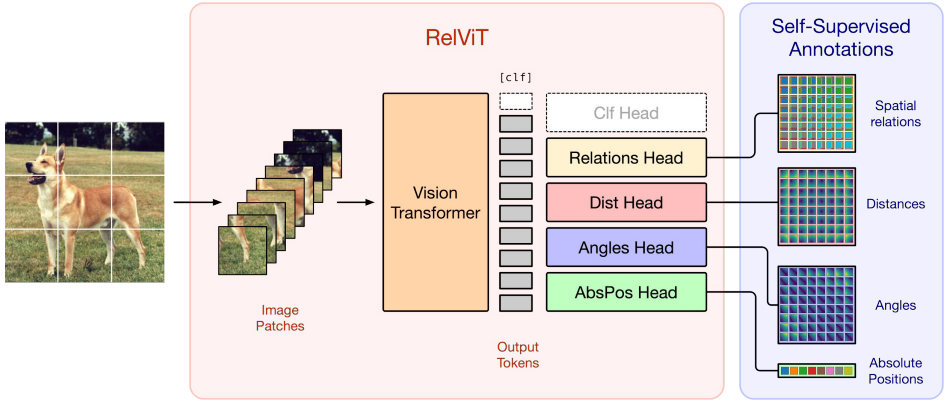
UNIVERSITÀ DEGLI STUDI DI PADOVA

Visual Intelligence Machine Perception Group

The ingredients



Self-Supervised learning (SSL)




Vision Transformers

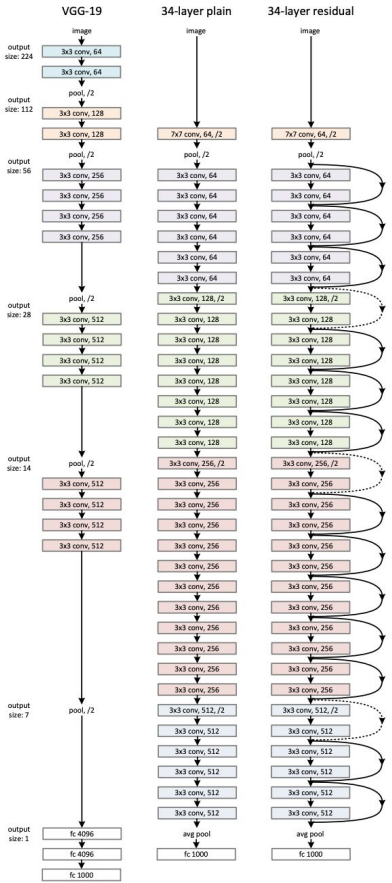
Inductive bias in CNNs

- **translation equivariance**:

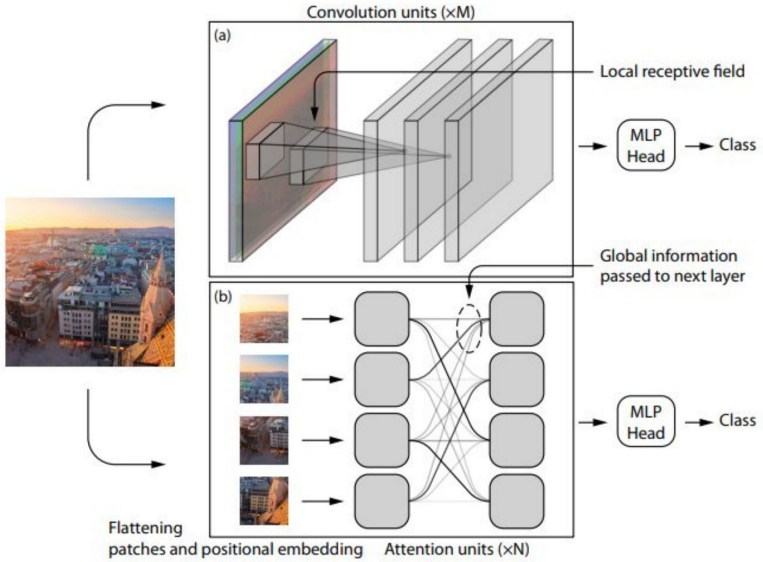
  CNNs are naturally equivariant to translation due to the weight sharing

- **locality:**

  pixels are strictly related to their neighbors

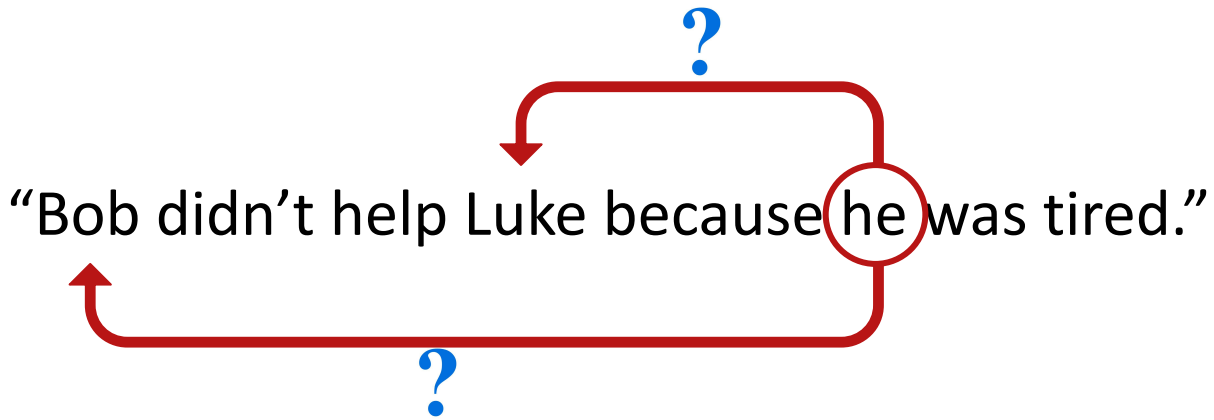


From Convolution to Attention

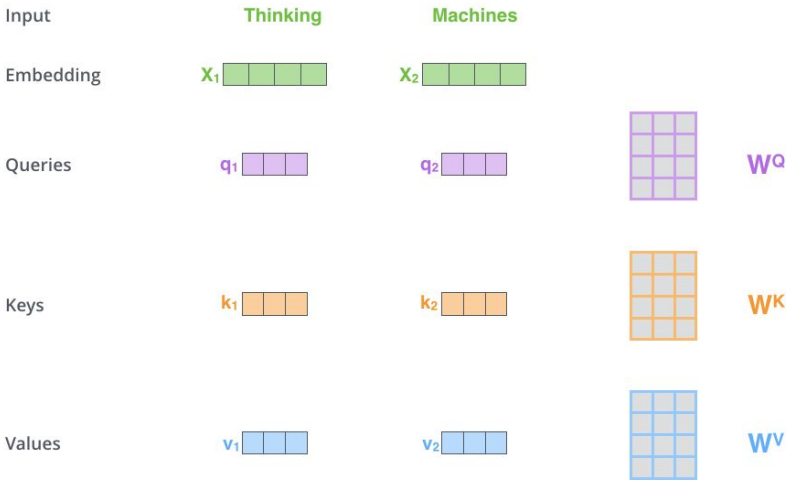


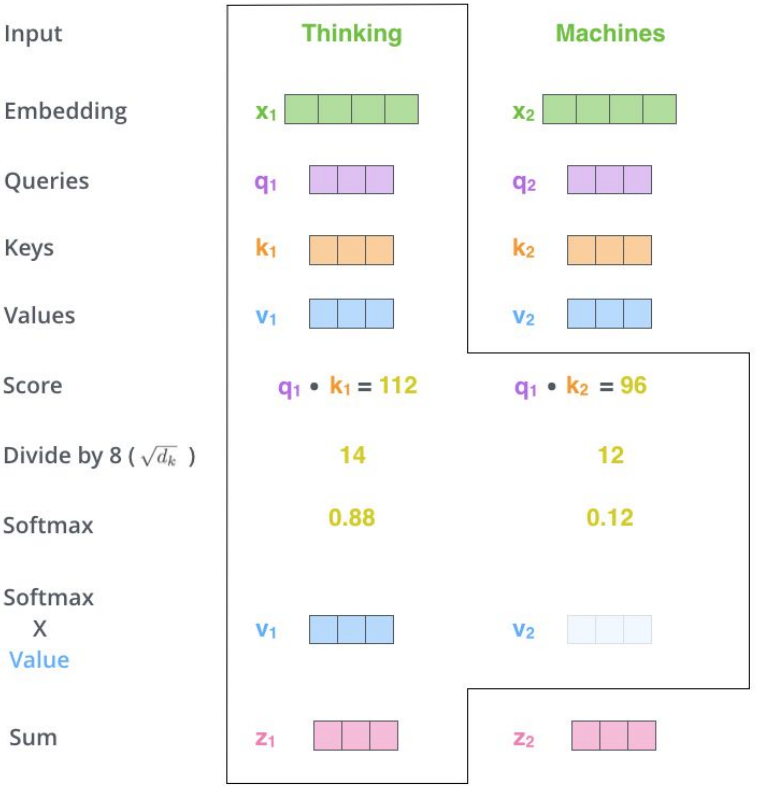
- An alternative to convolutions is the attention mechanism
- It lacks translation equivariance and locality
- The global attention has obtained great success in Natural Language Processing
- Recently, it has been applied in computer vision

The Attention

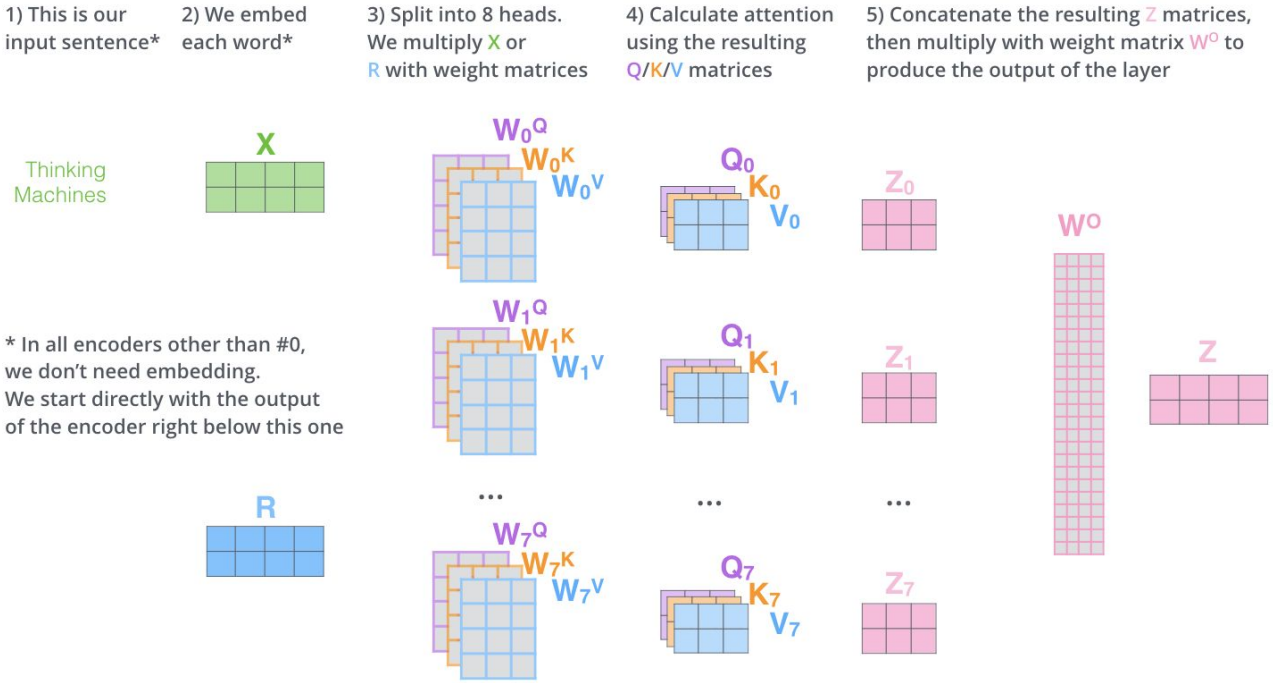


The Attention

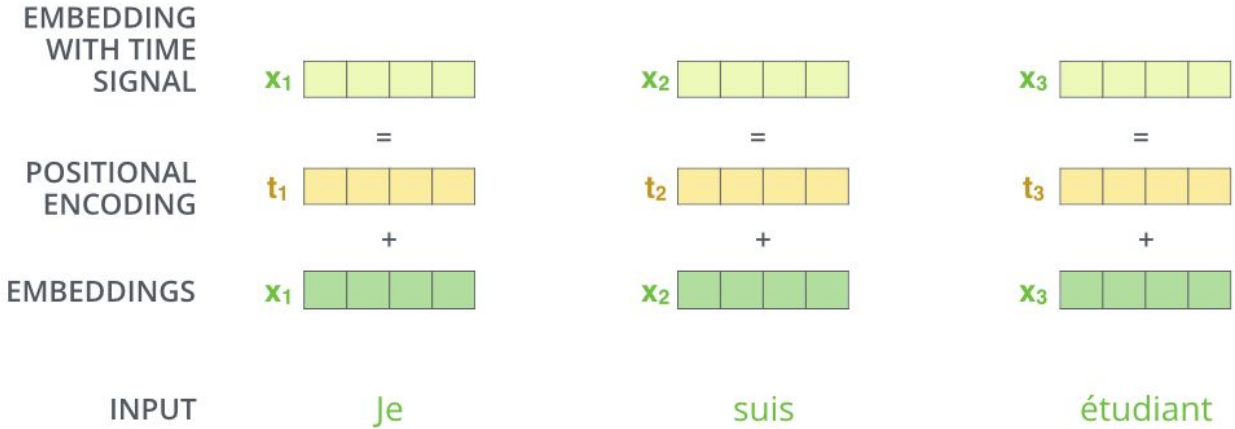




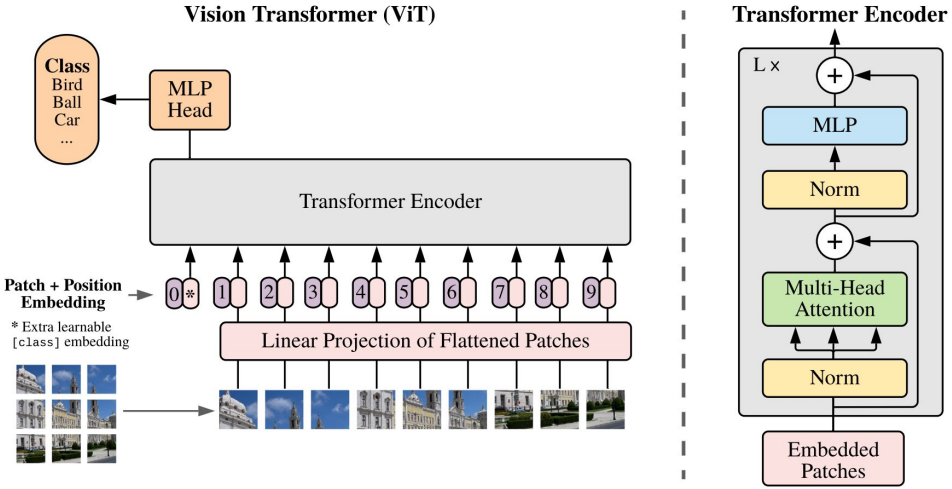
The Multi-Head Attention



Positional Encoding



Vision Transformers (ViTs)

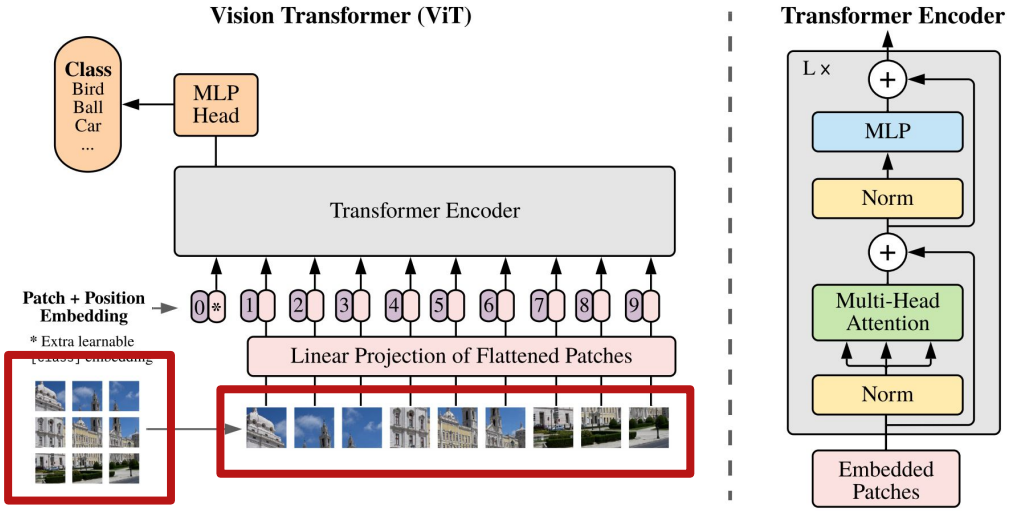


ViT is a **recent model** in the area of computer vision

It works considering patches as parts of the image such as words are parts of a sentence

Components

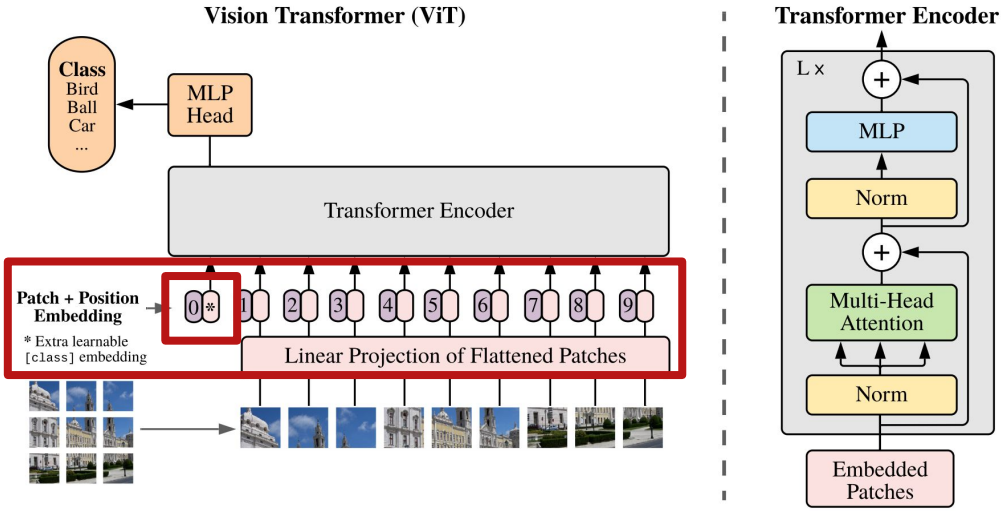
- The image is divided into patches of 16x16 pixels
- The set of image patches is the input of the model
- Each patch is flattened and linearly projected



Components

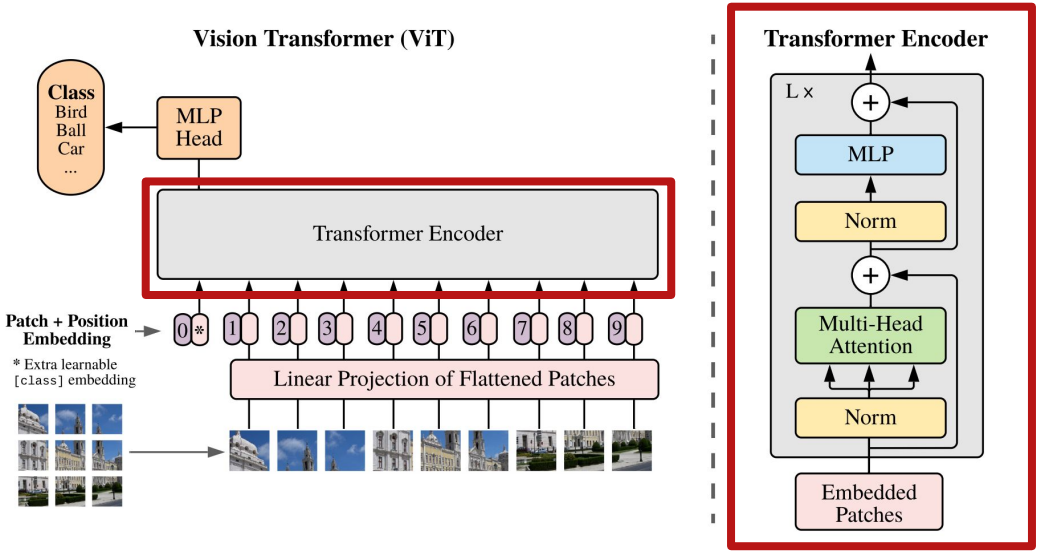
- A random vector is inserted at first position of the set of linear projected patches: this is called clf token as is used for classification at the end of computation

- Each linear projection is combined (added) with a positional encoded embedding



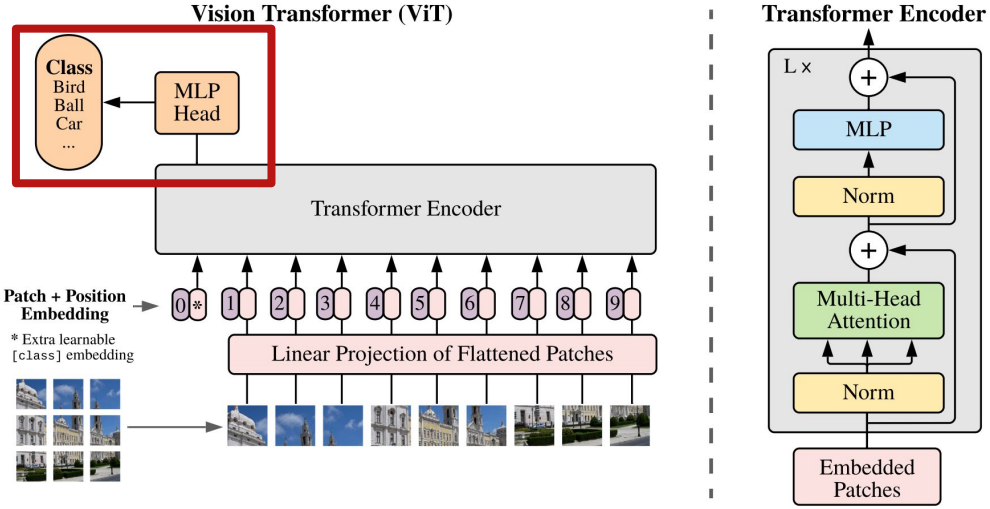
Components

- The projected patches, combined with the encoded positions, are passed to the transformer encoder
- The transformer encoder is the same as the original transformer architecture



Components

- The first output of the sequence, related to the clf input token, is passed to a MLP Head
- The MLP Head is a fully connected (even a single linear) layer that makes the final prediction

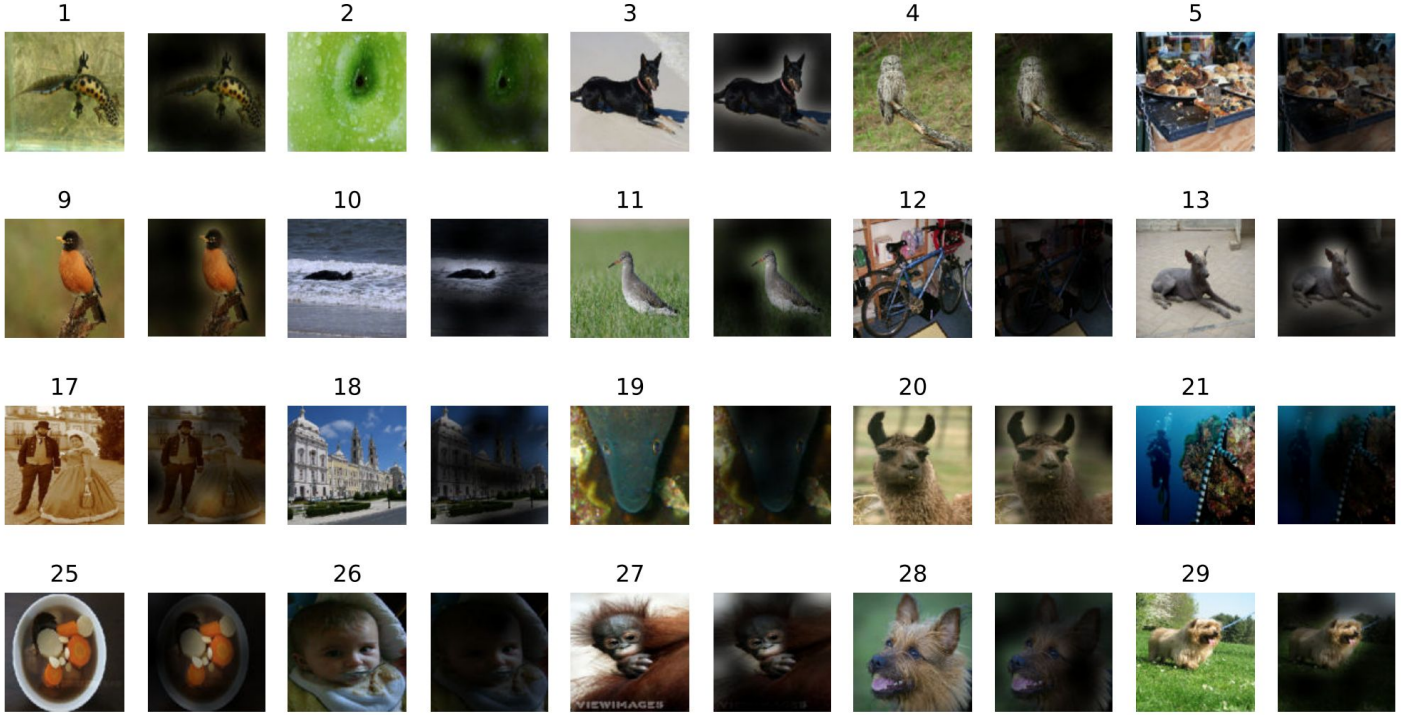


Model Variants

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Table 1: Details of Vision Transformer model variants.

ViT Results

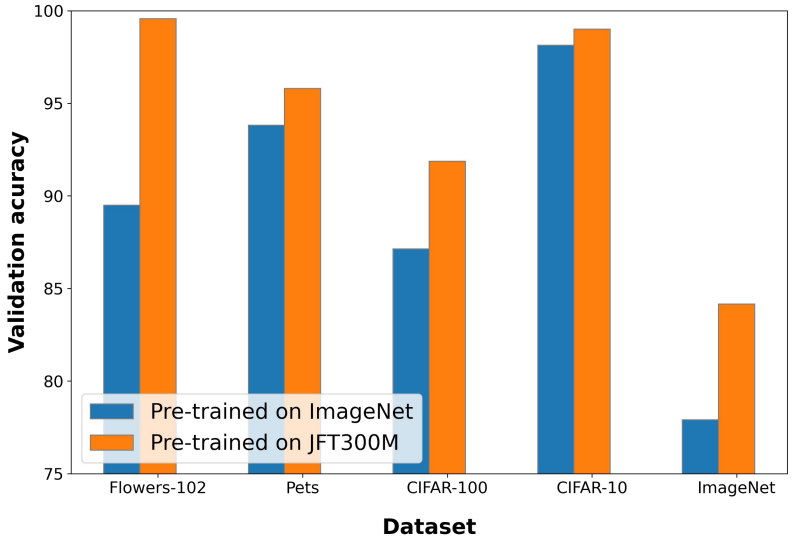


ViT Results

	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21K (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	–
CIFAR-100	**94.55** ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	–
Oxford-IIIT Pets	**97.56** ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02	**99.74** ± 0.00	99.61 ± 0.02	99.63 ± 0.03	–
VTAB (19 tasks)	**77.63** ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

Table 2: Comparison with state of the art on popular image classification datasets benchmarks. Vision Transformer models pre-trained on the JFT300M dataset often match or outperform ResNet-based baselines while taking substantially less computational resources to pre-train. *Slightly improved 88.5% result reported in Touvron et al. (2020).

Training set size Problem



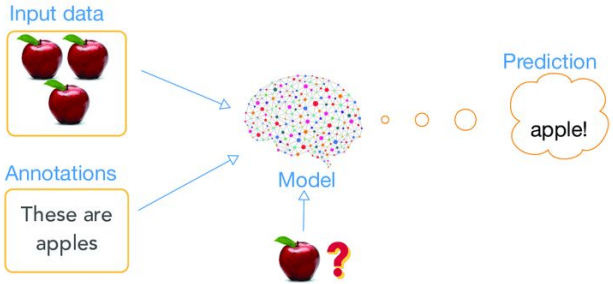
- ➔ Vision Transformer requires a lot of data in order to shine
- ➔ Performance strictly related to the training set size used during pre-training
- ➔ If trained from scratch, Vision Transformer is less accurate on small dataset
- ➔ Many efforts for making ViT more effective

Self-supervised learning

Self-supervised learning

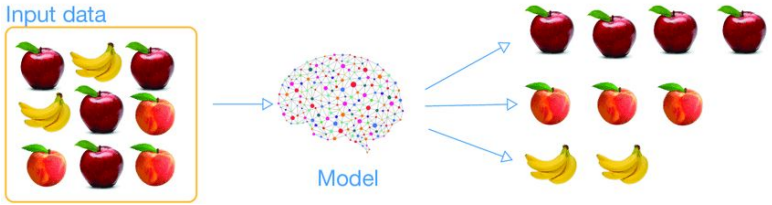
Supervised learning

- "traditional" learning paradigm
- input-output pairs
- labels provided by a "supervisor"
- labeling bias



Unsupervised learning

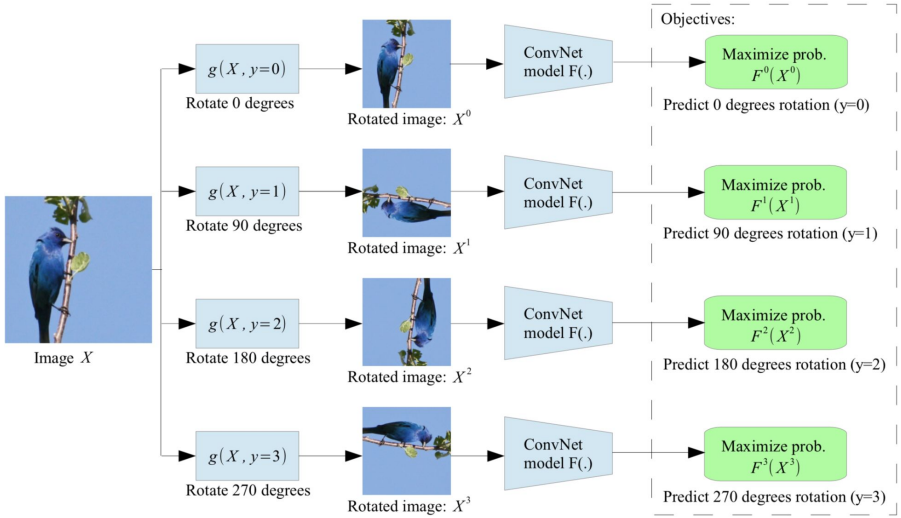
- unlabeled data
- based on patterns and regularities in data
- no human annotator



Self-supervised learning

- labeled data
- "implicit" labels
- no labeling bias
- solid background knowledge

Self-supervised learning - examples



- The prediction of the orientation is an example of SSL
- Given an input image, a random rotation is applied on it
- The model has to predict the rotation applied on the input image

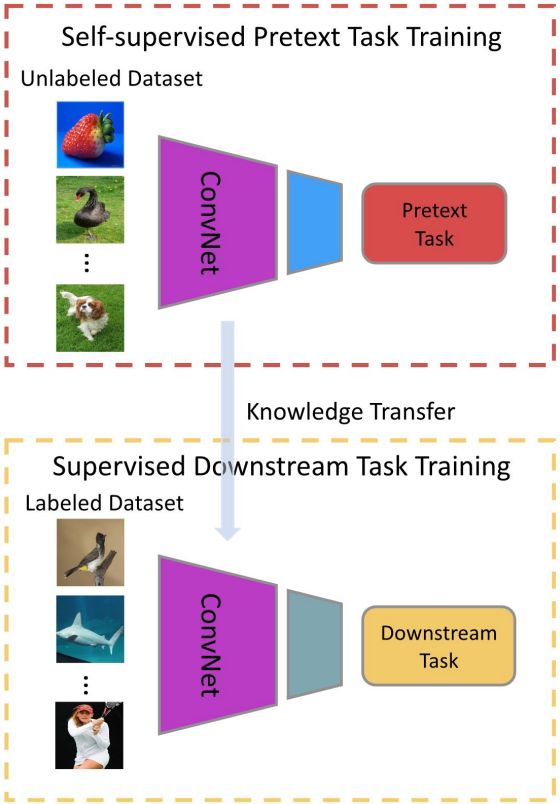
Self-supervised learning - examples

- Image colorization can be used as a SSL strategy
- Given an RGB image, a grayscale filter is applied
- The grayscale image is passed to the model
- The model has to predict the original RGB image

Self-supervised learning - examples

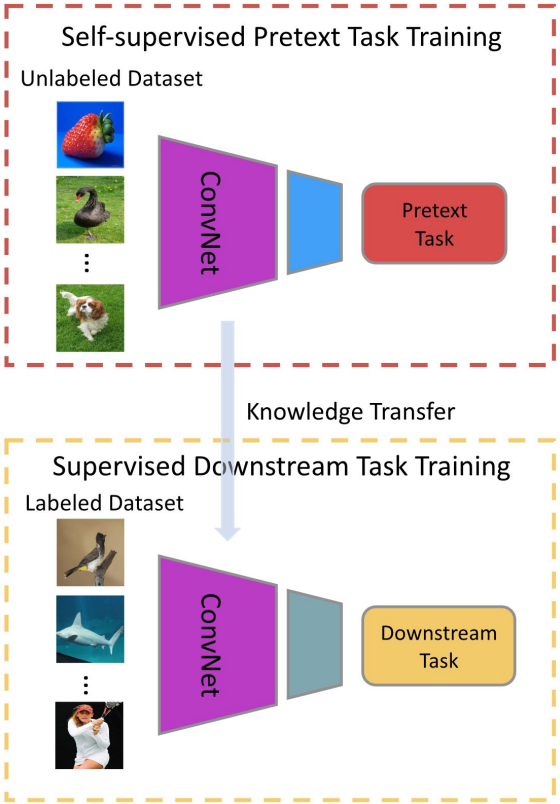
- The jigsaw puzzle as a SSL task
- An input image is divided into patches
- A random permutation is applied to the patches
- The model has to predict the original positions of each patch

The pipeline of the Self-supervised learning



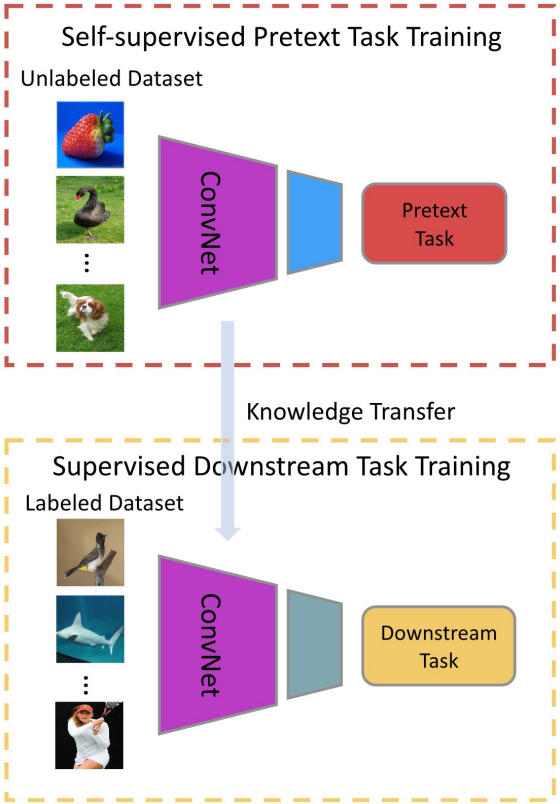
1. Self-supervised training phase: the model has to solve a pre-defined pretext task
2. The learned parameters are used as a pre-trained model
3. Transfer learning to a downstream computer vision tasks by fine-tuning

Why Self-supervised learning?



- The pretext task is based on some attributes of data
- Solving the SSL task, the model can learn both low-level features (corners, edges, textures, …), and high-level features (objects, scenes, object parts, …)
- The learned visual features can be helpful for other downstream tasks

When Self-supervised learning?



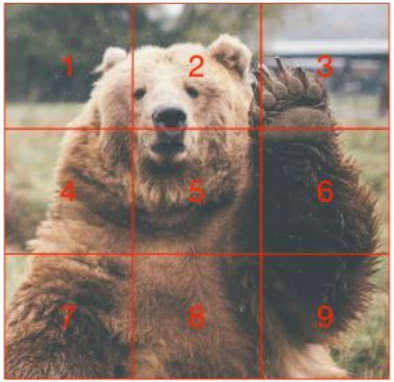
- **Expensive data annotations**: to reduce the collection and annotation of large-scale datasets
- **Small data available**: to improve performance and overcome over-fitting
- **One model, many tasks**: to create a solid background

SSL in Vision Transformer


UNIVERSITÀ
DEGLI STUDI
DI PADOVA

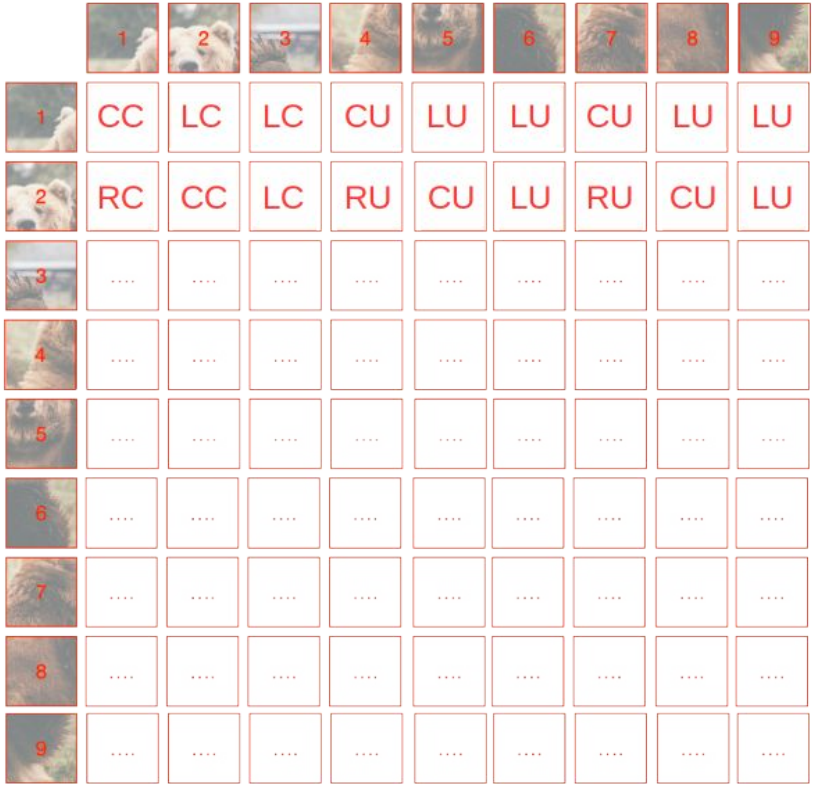
RelViT

Relational Vision Transformer (RelViT)

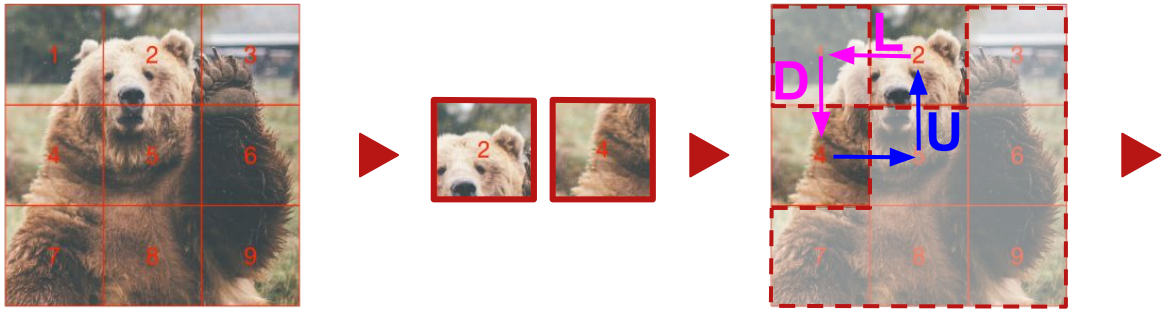


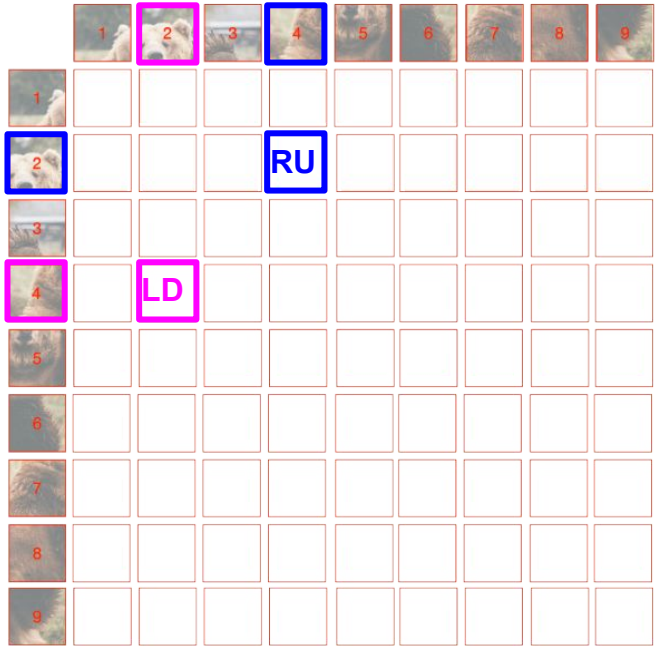
Spatial Classes

0 → left-up (LU)
1 → center-up (CU)
2 → right-up (RU)
3 → left-center (LC)
4 → center-center (CC)
5 → right-center (RC)
6 → left-down (LD)
7 → center-down (CD)
8 → right-down (RD)



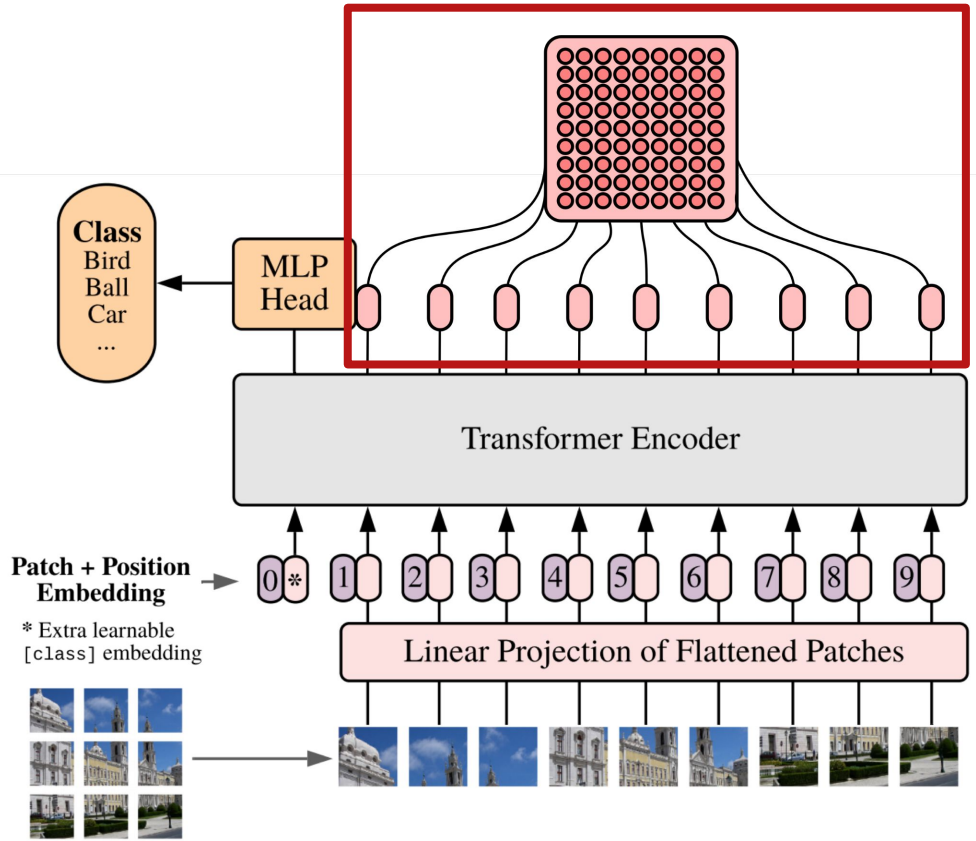
Relational Vision Transformer (RelViT)



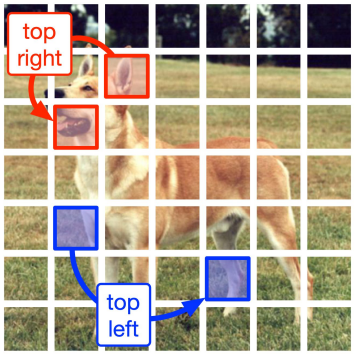


Network Architecture

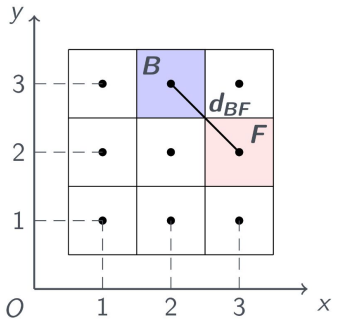
Predict spatial relations from the output of the Transformer Encoder



The self-supervised tasks

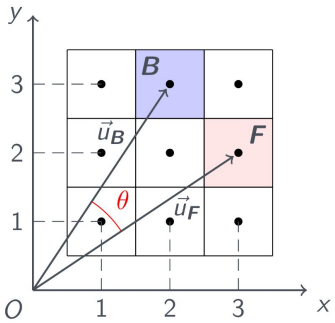


Spatial Relations:
to recognize the relation class of couples of image patches



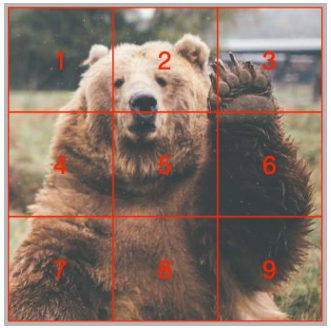
$$\vec{p}_B = (x_B; y_B)$$
$$\vec{p}_F = (x_F; y_F)$$
$$d_{BF} = \|\vec{p}_B - \vec{p}_F\|$$

Distances:
to learn euclidean distances among patches locations in the original image.



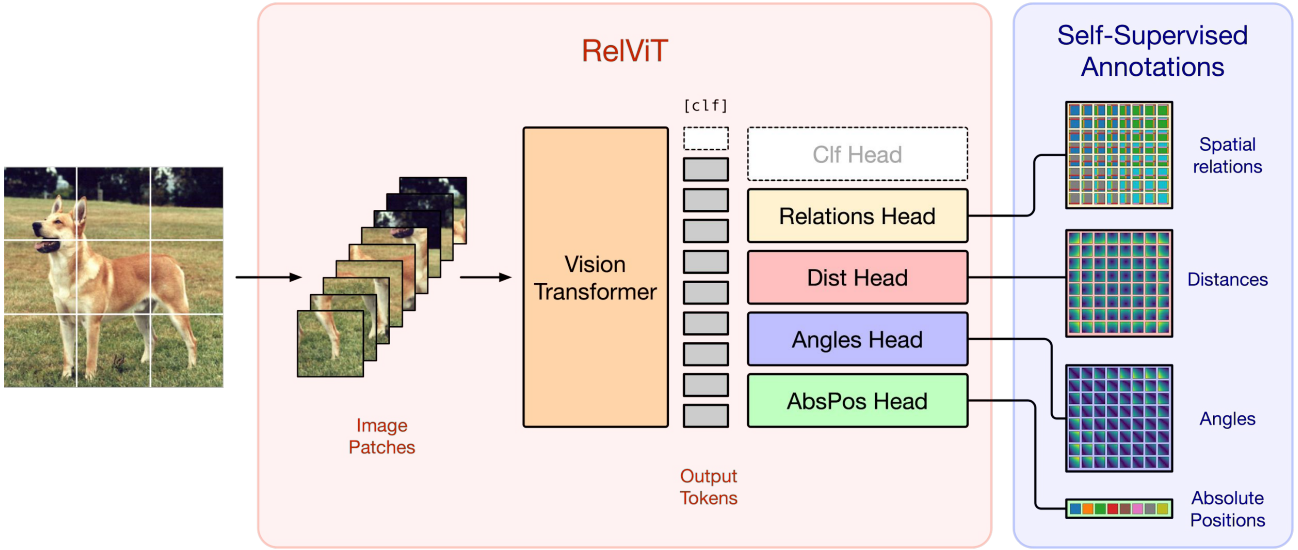
$$\theta = \arccos\left(\frac{\vec{u}_B \cdot \vec{v}_F}{\|\vec{u}_B\|\|\vec{v}_F\| + \epsilon}\right)$$

Angles:
to learn angles among couples of input patch locations.



Abs positions:
to recognize the 2D location of the input patch

Relational Vision Transformer (RelViT)



- ➔ all output tokens are optimized
- ➔ supervision and self-supervision share the same backbone
- ➔ multiple-tasks modality
- ➔ intuitive and simple implementation

Experimental Results

Results on various **small datasets.**

Self-supervised upstream from scratch trained for 100 epochs, plus **supervised fine-tuning** trained for 100 epochs.

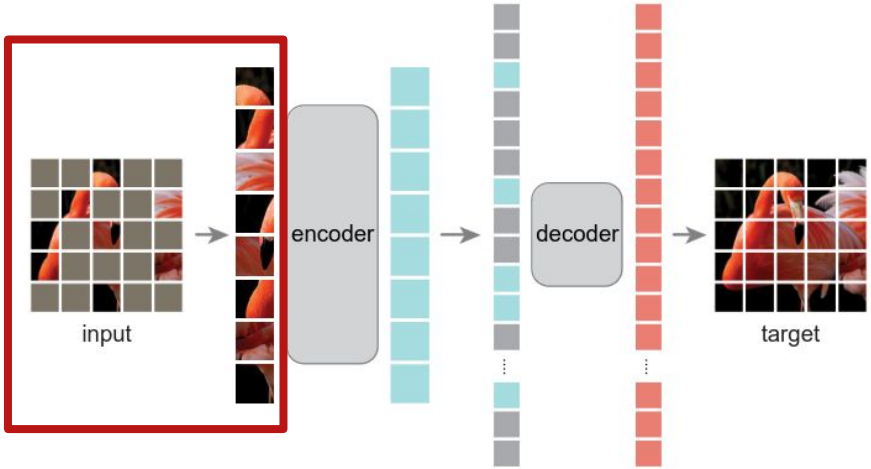
	Backbone	Supervised	RelViT	Improv.
CIFAR-10	ViT-S/4	86.09 ±0.46	**90.23** ±0.09	**+4.14** ↑
SVHN	ViT-S/4	96.01 ±0.07	**97.14** ±0.03	**+1.13** ↑
CIFAR-100	ViT-S/4	59.19 ±0.84	**64.99** ±0.46	**+5.85** ↑
Flower-102	ViT-S/32	42.08 ±0.29	**45.78** ±0.75	**+3.70** ↑
TinyImagenet	ViT-S/8	43.19 ±0.78	**51.98** ±0.20	**+8.79** ↑
Imagenet100	ViT-S/32	58.04 ±0.91	**66.46** ±0.45	**+8.42** ↑

MAE

Masked Autoencoders (MAE)

During pre-training:

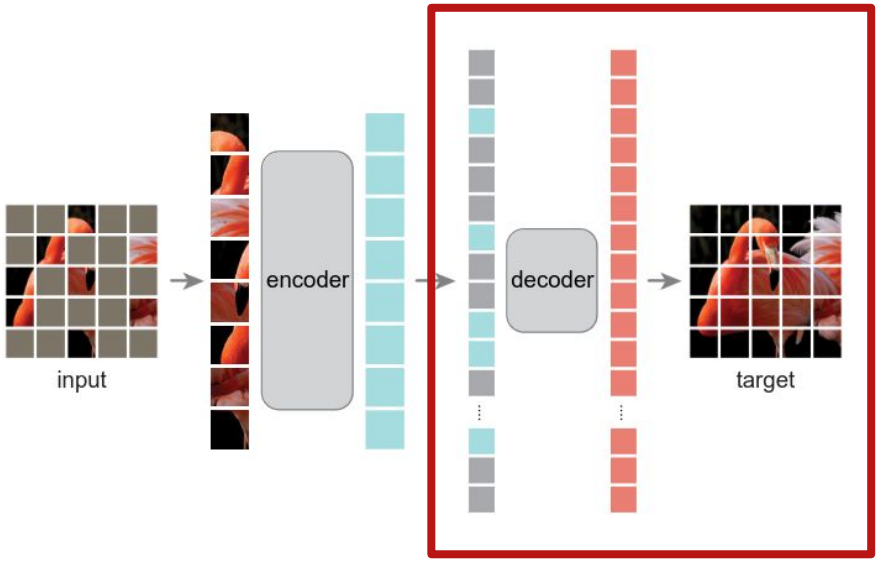
- The image is divided into patches as in the standard Vision Transformer
- A pre-defined number of patches is randomly masked
- Only the visible tokens are passed to the encoder



Masked Autoencoders (MAE)

During pre-training:

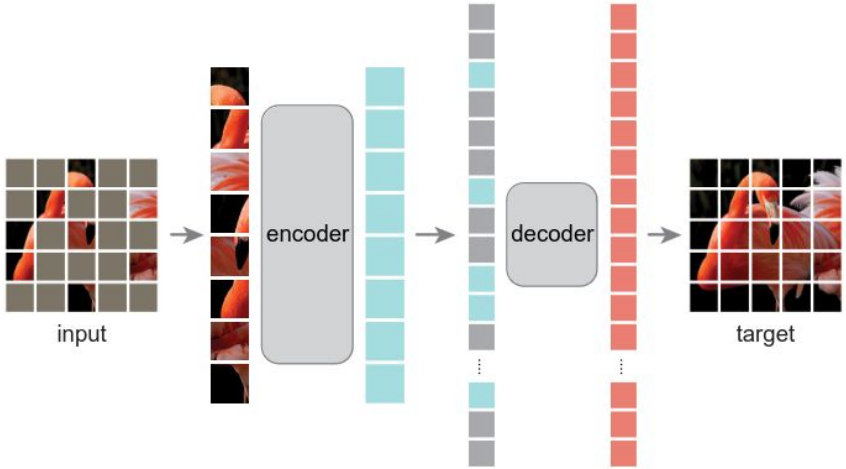
- After the encoder, the mask tokens are introduced
- The full set of tokens are processed by a small decoder
- The decoder reconstructs the original image in pixels



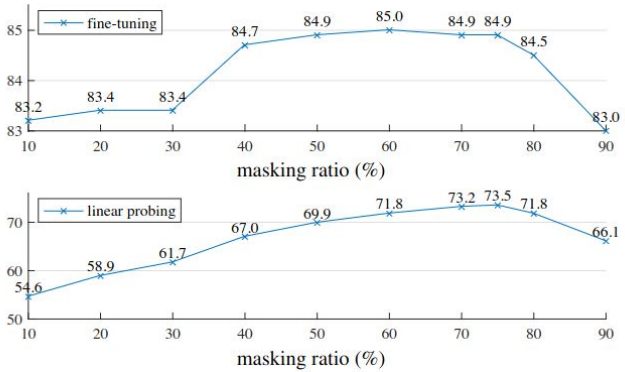
Masked Autoencoders (MAE)

After pre-training:

- The decoder is discarded
- The encoder is applied to unmasked image for solving downstream task



Masked Autoencoders (MAE)



- A high masking ratio works well for both fine-tuning and linear probing experiments

- Masking 75% of the patches, MAE is very effective on ViT-L

scratch, original [16]	scratch, our impl.	baseline MAE
76.5	82.5	84.9

Masked Autoencoders (MAE)

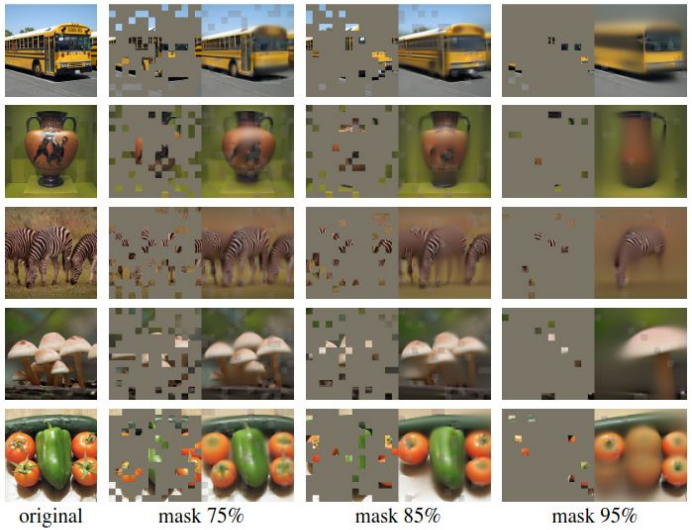


Figure 4. Reconstructions of ImageNet *validation* images using an MAE pre-trained with a masking ratio of 75% but applied on inputs with higher masking ratios. The predictions differ plausibly from the original images, showing that the method can generalize.

I-JEPA

Image-based Joint-Embedding Predictive Architecture (I-JEPA)

- A context block, a portion of the original image, is sampled from the image
- The context block is passed to the context encoder, that is a standard ViT

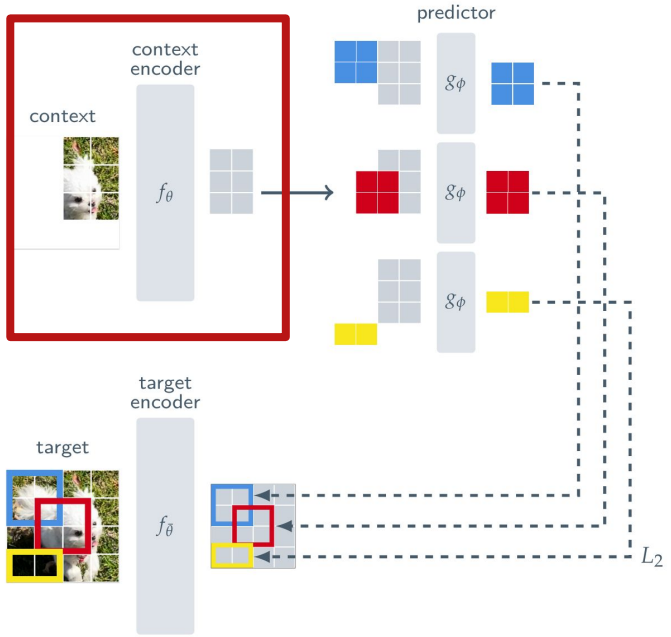


Image-based Joint-Embedding Predictive Architecture (I-JEPA)

In parallel:

- The original image is processed by a target encoder
- M target blocks are defined which are portion of the image
- The target encoder predicts the representations of a target block at a specific position

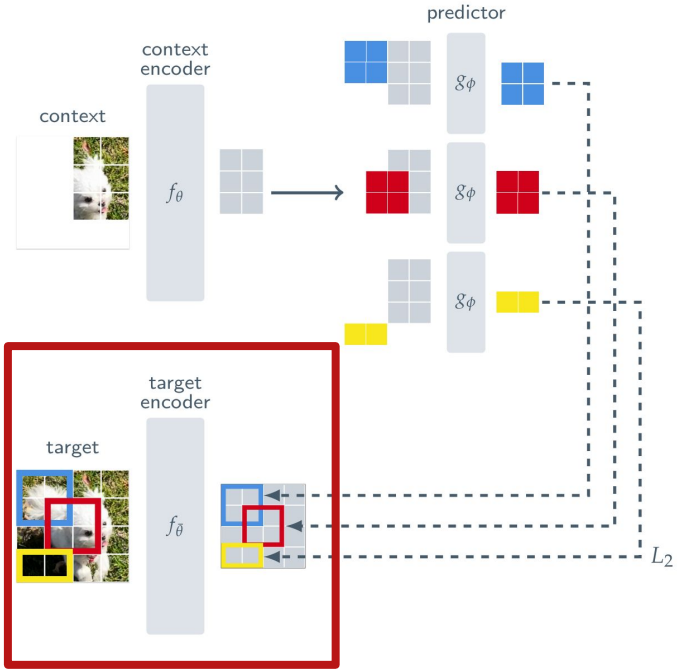


Image-based Joint-Embedding Predictive Architecture (I-JEPA)

- Exploiting the PE, the output of the context encoder is passed to M predictors, one for each target blocks
- The predictions of each predictor is compared with the representations got from the target encoder

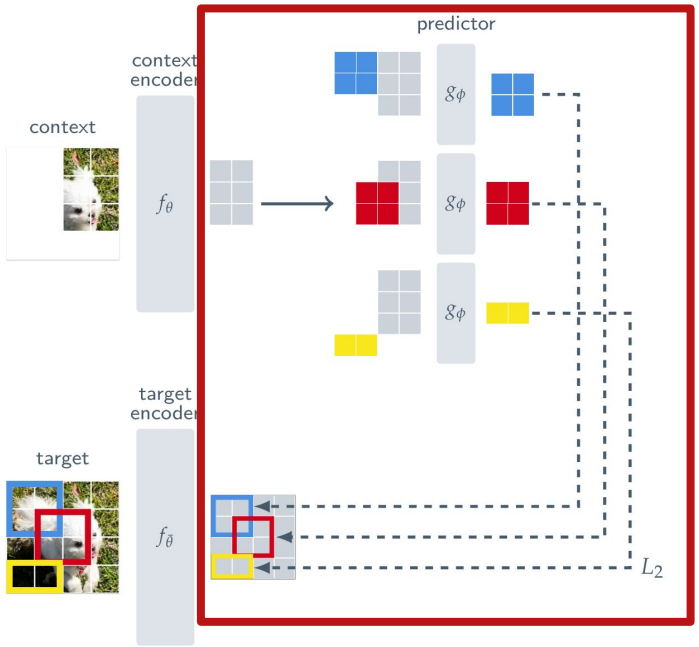
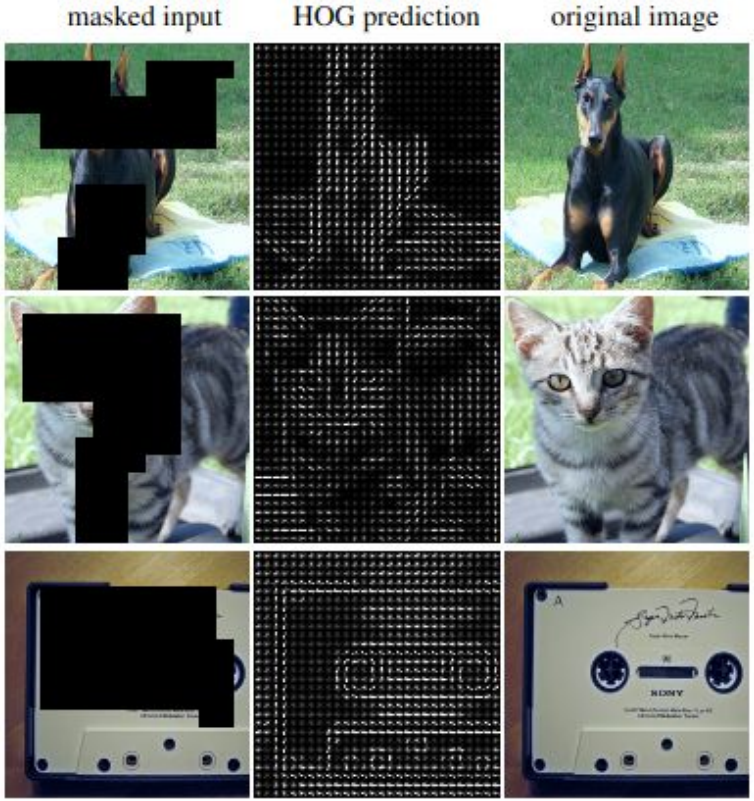


Image-based Joint-Embedding Predictive Architecture (I-JEPA)

Figure 4. Examples of our context and target-masking strategy. Given an image, we randomly sample 4 target blocks with scale in the range $(0.15, 0.2)$ and aspect ratio in the range $(0.75, 1.5)$. Next, we randomly sample a context block with scale in the range $(0.85, 1.0)$ and remove any overlapping target blocks. Under this strategy, the target-blocks are relatively semantic, and the context-block is informative, yet sparse (efficient to process).

MaskFeat

Masked Feature Prediction (MaskFeat)



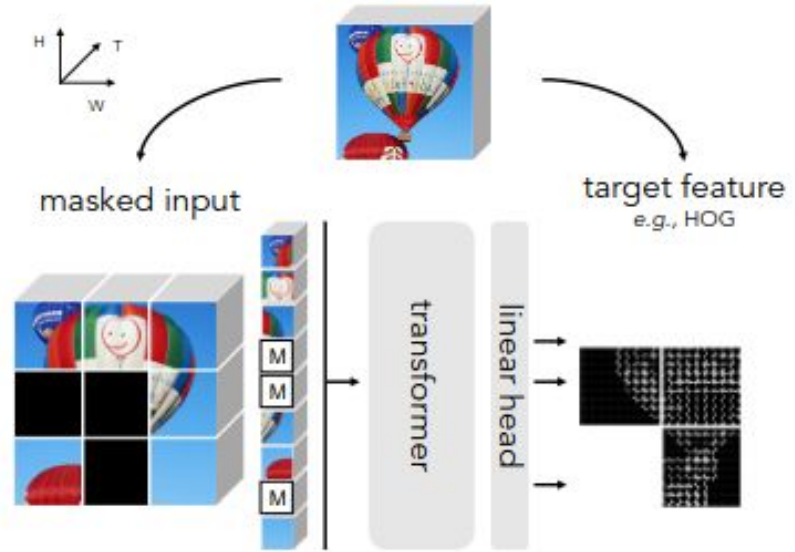
The SSL task is:

- From an image, we get:
  - a masked version of it
  - the HOG feature
- The masked image is the input data
- The HOG feature of the masked parts are the target

Masked Feature Prediction (MaskFeat)

During pre-training:

- The masked image is passed to the transformer encoder
- The model predict the HOG features of the masked patches



References

- C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer, “Masked feature prediction for self-supervised visual pre-training,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14 668–14 678
- R. Zhang, P. Isola, and A. A. Efros, “Colorful image colorization,” in Proc. of the European Conference on Computer Vision (ECCV), 2016
- A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in Proc. of the International Conference on Learning Representations (ICLR), 2021.
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in Proc. of Advances in Neural Information Processing Systems (NeurIPS), 2017
- M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, “Self-supervised learning from images with a joint-embedding predictive architecture,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 15 619–15 629
- G. Camporese, E. Izzo, and L. Ballan, “Where are my neighbors? exploiting patches relations in self-supervised vision transformer,” Proc. of the British Machine Vision Conference (BMVC), 2022
- M. Noroozi and P. Favaro, “Unsupervised learning of visual representations by solving jigsaw puzzles,” in Proc. of the European Conference on Computer Vision (ECCV), 2016
- C. Doersch, A. Gupta, and A. A. Efros, “Unsupervised visual representation learning by context prediction,” in Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV), 2015
- S. Gidaris, P. Singh, and N. Komodakis, “Unsupervised representation learning by predicting image rotations,” in Proc. of the International Conference on Learning Representations (ICLR), 2018