Attention and transformers

E. Decencière

MINES ParisTech
PSL Research University
Center for Mathematical Morphology

Contents

Contents

Transformers: a new revolution in deep learning?

- Transformers [Vaswani et al., 2017] have brought a break-through in natural language processing
  - Bidirectional Encoder Representations from Transformers (BERT, by Google [Brown et al., 2020])
  - Generative Pre-trained Transformer 3 (GPT-3, by OpenAI [Devlin et al., 2019]): 175 billion parameters.
- They contribute to the development of new natural language processing applications (translation, voice assistants, etc.)
- Will they do the same in image analysis?

NB: Our aim through this lesson is to (hopefully!) review the main ideas on attention and transformers, through selected examples. However, this overview is not exhaustive.

What are transformers?

Definition

A transformer is a neural network architecture module that allows the network to adaptively focus its attention on certain regions of the data.

What are transformers?

Definition

A transformer is a neural network architecture module that allows the network to adaptively focus its attention on certain regions of the data.

Transformers today

Nowadays, when people refer to the transformer, they generally mean the architecture proposed by Vaswani et al. in 2017 [Vaswani et al., 2017].

Contents

Contents

How do we look at an image?

Credits: Ilya Repin, An Unexpected Visitor, 1884.

How do we look at an image?

Tasks:

- Age of the characters?
- How long has the visitor been away?
- Memorize the objects in the scene.

Credits: Experiments on visual attention [Yarbus, 1967]

Information used by human visual attention

- Bottom-up:
    - local features (orientation, intensity, junctions, colour, motion, etc.)
    - local features contrast
    - context
- Top-bottom: task related
- Construction of a single *saliency map*

Exploring the image

- Winner-takes all! We focus on the maximum of the saliency map.
- Inhibition of return: We explore the following maxima, at first avoiding those that have already been inspected

Why has visual attention evolved?

- Photoreceptor cells are expensive
- Processing power is limited
- Solution: concentrate the cells in a given region and use the gaze to optimize their use

Why has visual attention evolved?

- Photoreceptor cells are expensive
- Processing power is limited
- Solution: concentrate the cells in a given region and use the gaze to optimize their use

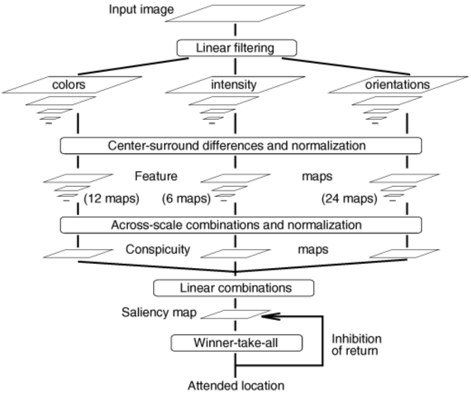
- The same arguments apply to artificial visual systems

+ Some degree of invariance
+ Interpretability

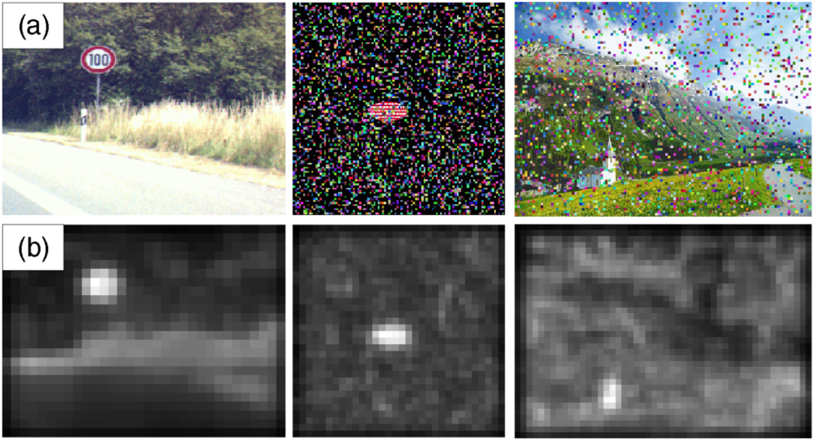
Contents

A classical bottom-up model

- Itti et al. [Itti et al., 1998] proposed a model inspired by the primate visual system.
- It only uses low-level information.



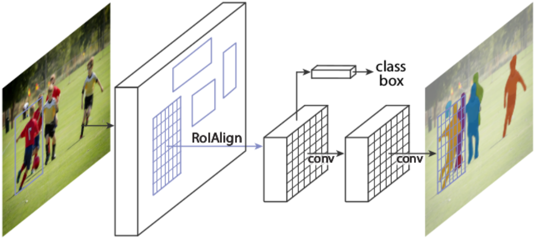
Examples [Itti et al., 1998]



Contents

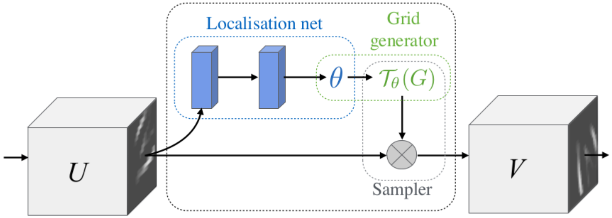
Region proposal networks [Ren et al., 2015]

- Detection and instance segmentation methods use region proposal networks, that can be interpreted as an attention mechanism.
- The region proposal network gives the coordinates of the rectangle and a probability that it contains an object.

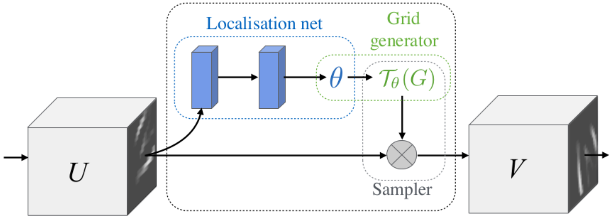


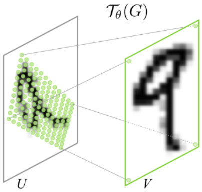
A region proposal module is used by mask R-CNN [He et al., 2017]

Spatial transformers [Jaderberg et al., 2016]

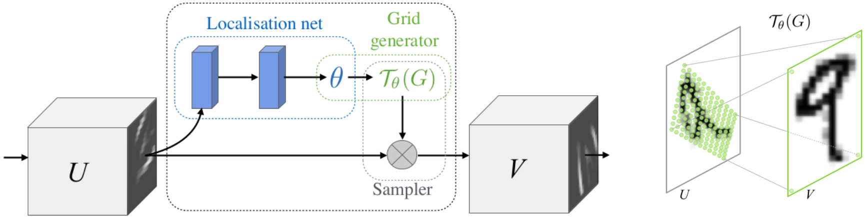


Spatial transformers [Jaderberg et al., 2016]



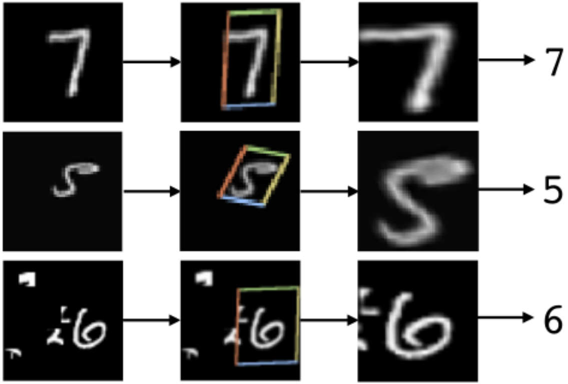


Spatial transformers [Jaderberg et al., 2016]

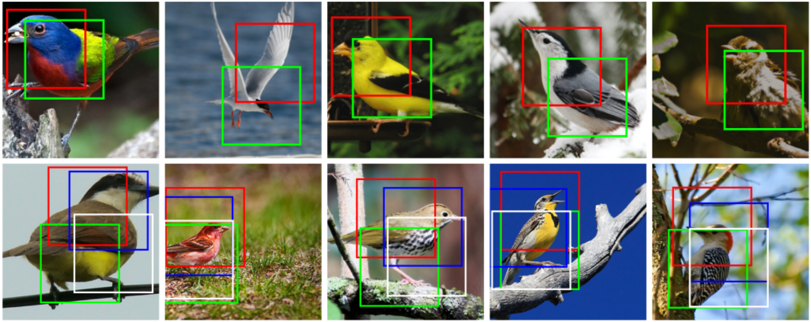


- This module can be added to any convolutional network
- End-to-end learning

Spatial transformers illustration



Spatial transformers with multiple heads



Remarks

- Note that in the first row one transformer tends to focus on the bird's head, while the second is centered on the body
- In the second row, the specialization is less apparent

Contents

Transformer avatars

Some examples

- Graph transformers [Lecun et al., 1998]
- Transforming auto-encoders [Hinton et al., 2011]
- Spatial transformers [Jaderberg et al., 2016]

The transformer [Vaswani et al., 2017].

Today, when people refer to the transformer, they generally mean the architecture proposed by Vaswani et al. in 2017.

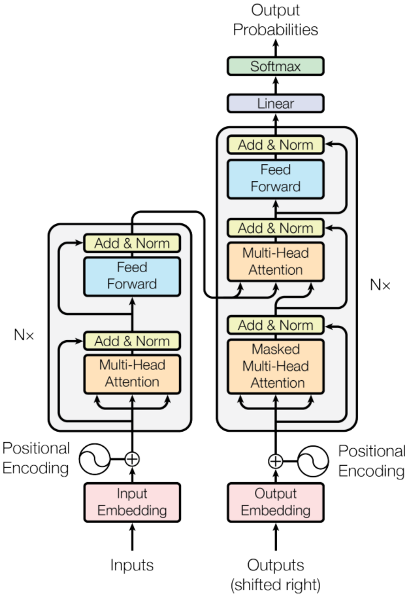
Contents

The rise of transformers

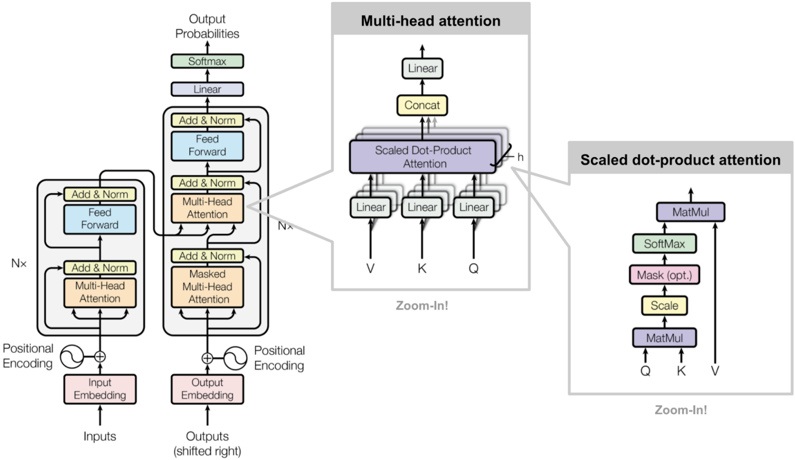
The paper that started it all

Vaswani et al., Attention is all you need, Neurips 2017.

This architecture was developed for text processing.



Architecture [Vaswani et al., 2017]



Credits: https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html

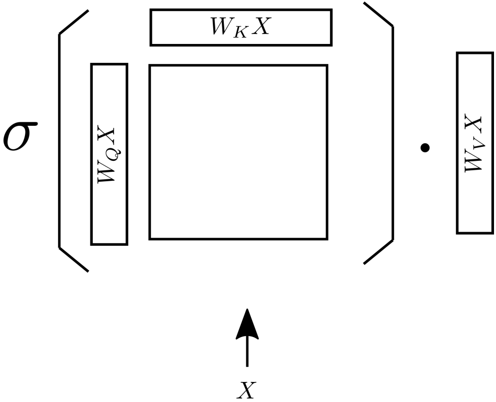
Scaled dot-product attention

Definition

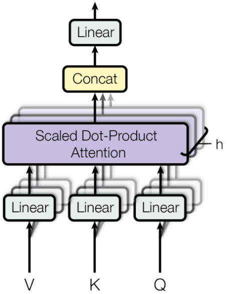
$$Att(Q, K, V) = \sigma\left(\frac{QK^t}{\sqrt{d_K}}\right)V$$

- V: values; K: keys; Q: queries.
- d_K is the length of K.
- σ: row-wise soft-max.

Self-attention

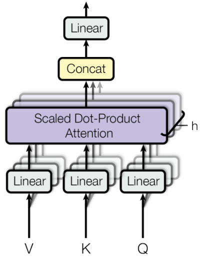


Multi-head attention



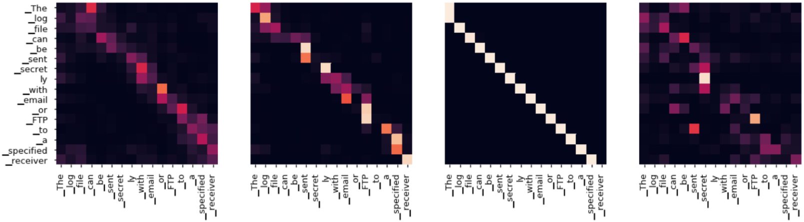
- Matrices W_Q, W_V and W_V are learnable.
- h heads work in parallel.

Dot-product self-attention illustration



In the case of self-attention:

- $V = K = Q = X$



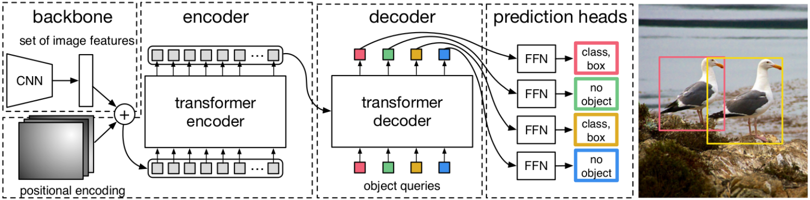
Credits:
https://nlp.seas.harvard.edu/2018/04/03/a

Success of transformers in natural language processing

- Bidirectional Encoder Representations from Transformers (BERT, by Google [Brown et al., 2020])
- Generative Pre-trained Transformer 3 (GPT-3, by OpenAI [Devlin et al., 2019]): 175 billion parameters.

Contents

DETR: detection transformer [Carion et al., 2020]

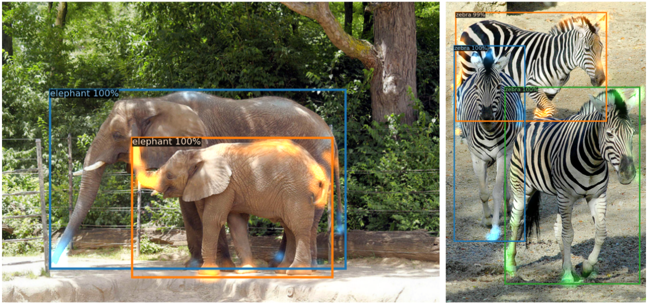


Remarks

- Convolutional layers are used to encode the image
- After a 1×1 convolutional layers, each feature map is flattened and considered as an input for the transformer encoder
- Decoder outputs are processed by a feed-forward network (FFN) to generate the box coordinates and label (possibly $\emptyset$).

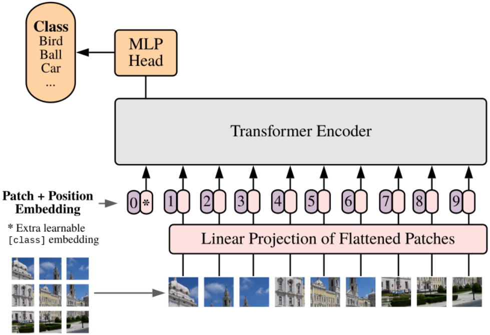
Results and comments

- Similar accuracy and run-time performance to Faster R-CNN on the COCO object detection dataset
- Optimization was apparently difficult (extra losses, for instance)



Contents

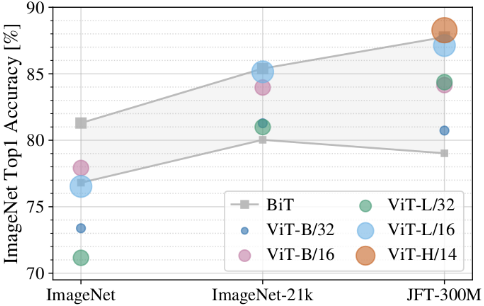
ViT: the vision transformer [Dosovitskiy et al., 2021]



Remarks

- Only uses the transformer encoder
- Directly takes as inputs image patches
- Achieves state-of-the-art results when pre-trained on very large databases (Google's JFT-300M dataset)

ViT results

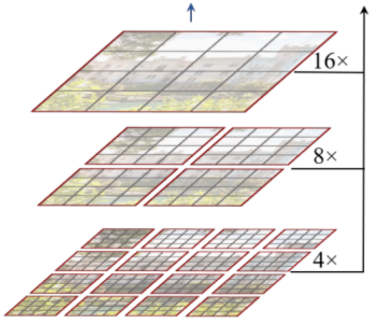


[1]

- ViT-H/14 requires 2500 TPUv3-core-days for pre-training
- But: "Training data-efficient image transformers & distillation through attention" [Touvron et al., 2021].

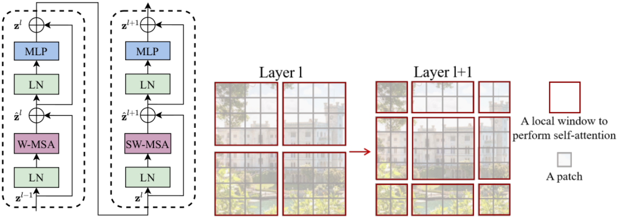
[1] BiT: Big transfer [Kolesnikov et al., 2020]

Shifted window (SWIN) transformer [Liu et al., 2021]



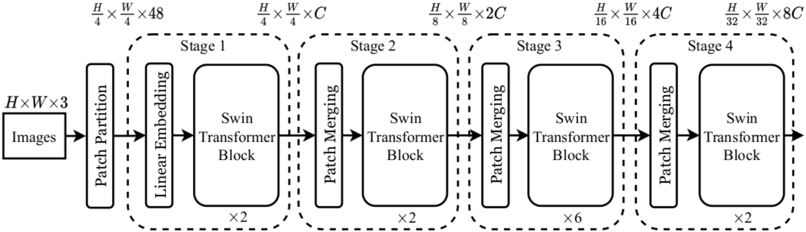
- The transformer modules are applied within each window
- Hierarchical approach: patches are merged at some levels
- The model uses **shifted windows**

SWIN blocks



- Multi-headed self-attention with regular (W-MSA) and shifted (SW-MSA) windowing configurations are applied alternatively

SWIN architecture



Results

The Swin transformer obtains better results than previous methods on:

- ImageNet 1k and 22k classification
- COCO object detection and image segmentation
- ADE20k semantic segmentation

Contents

Discussion

Convolutional neural networks

- Convolutional networks are based on two inductive biases:

Transformers

Discussion

Convolutional neural networks

- Convolutional networks are based on two inductive biases:
    - Locality

Transformers

Discussion

Convolutional neural networks

- Convolutional networks are based on two inductive biases:
  - Locality
  - Translation equivariance

Transformers

Discussion

Convolutional neural networks

- Convolutional networks are based on two inductive biases:
  - Locality
  - Translation equivariance
- These hypothesis allow simplifying the models, but may also limit their generality

Transformers

Discussion

Convolutional neural networks

- Convolutional networks are based on two inductive biases:
  - Locality
  - Translation equivariance
- These hypothesis allow simplifying the models, but may also limit their generality
  - Long range interactions are difficult to take into account

Transformers

Discussion

Convolutional neural networks

- Convolutional networks are based on two inductive biases:
  - Locality
  - Translation equivariance
- These hypothesis allow simplifying the models, but may also limit their generality
  - Long range interactions are difficult to take into account
  - Translation equivariance is not always welcome

Transformers

Discussion

Convolutional neural networks

- Convolutional networks are based on two inductive biases:
  - Locality
  - Translation equivariance
- These hypothesis allow simplifying the models, but may also limit their generality
  - Long range interactions are difficult to take into account
  - Translation equivariance is not always welcome

Transformers

- Transformers, like fully connected layers, do not make any assumptions on the data structure

Discussion

Convolutional neural networks

- Convolutional networks are based on two inductive biases:
  - Locality
  - Translation equivariance
- These hypothesis allow simplifying the models, but may also limit their generality
  - Long range interactions are difficult to take into account
  - Translation equivariance is not always welcome

Transformers

- Transformers, like fully connected layers, do not make any assumptions on the data structure
  - Localization is brought by a positional encoding

Discussion

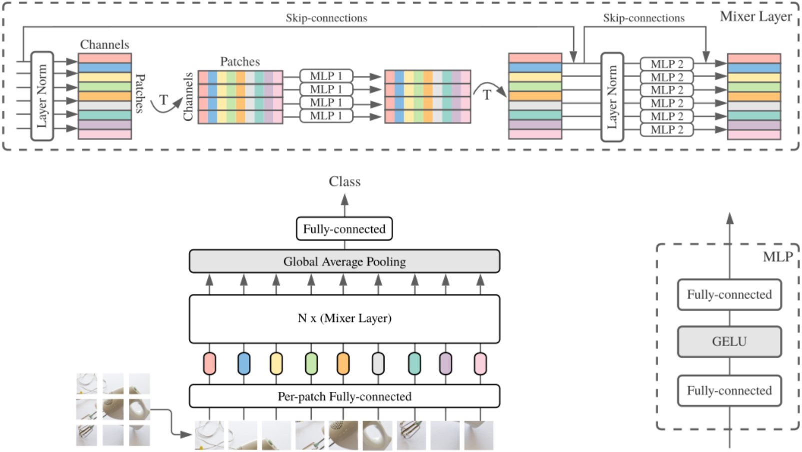
Convolutional neural networks

- Convolutional networks are based on two inductive biases:
  - Locality
  - Translation equivariance
- These hypothesis allow simplifying the models, but may also limit their generality
  - Long range interactions are difficult to take into account
  - Translation equivariance is not always welcome

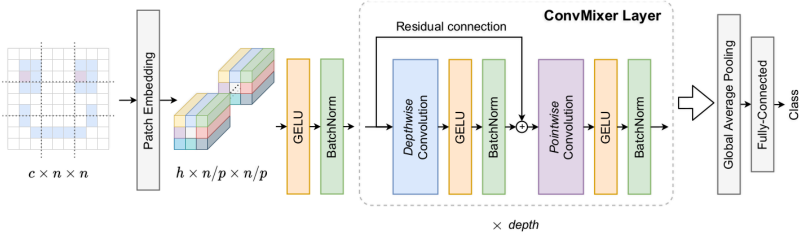
Transformers

- Transformers, like fully connected layers, do not make any assumptions on the data structure
  - Localization is brought by a positional encoding
- Are transformers a smart way of analysing images with fully connected layers?

MLPs is all you need [Tolstikhin et al., 2021]



Patches are all you need? [Trockman and Kolter, 2022]



References I

[Brown et al., 2020] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*. arXiv: 2005.14165.

[Carion et al., 2020] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-End Object Detection with Transformers. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J.-M., editors, *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pages 213–229, Cham. Springer International Publishing.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. arXiv: 1810.04805.

[Dosovitskiy et al., 2021] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *arXiv:2010.11929 [cs]*. arXiv: 2010.11929.

References II

[He et al., 2017] He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask R-CNN. *arXiv:1703.06870 [cs]*. arXiv: 1703.06870.

[Hinton et al., 2011] Hinton, G. E., Krizhevsky, A., and Wang, S. D. (2011). Transforming Auto-Encoders. In Honkela, T., Duch, W., Girolami, M., and Kaski, S., editors, *Artificial Neural Networks and Machine Learning – ICANN 2011*, Lecture Notes in Computer Science, pages 44–51, Berlin, Heidelberg. Springer.

[Itti et al., 1998] Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[Jaderberg et al., 2016] Jaderberg, M., Simonyan, K., Zisserman, A., and Kavukcuoglu, K. (2016). Spatial Transformer Networks. *arXiv:1506.02025 [cs]*. arXiv: 1506.02025.

[Kolesnikov et al., 2020] Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. (2020). Big Transfer (BiT): General Visual Representation Learning. In *European Conference on Computer Vision*.

[Lecun et al., 1998] Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

References III

[Liu et al., 2021] 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. pages 10012–10022.

[Ren et al., 2015] Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Advances in Neural Information Processing Systems*, 28:91–99.

[Tolstikhin et al., 2021] Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., Lucic, M., and Dosovitskiy, A. (2021). MLP-Mixer: An all-MLP Architecture for Vision. *arXiv:2105.01601 [cs]*. arXiv: 2105.01601.

[Touvron et al., 2021] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. *arXiv:2012.12877 [cs]*. arXiv: 2012.12877.

[Trockman and Kolter, 2022] Trockman, A. and Kolter, J. Z. (2022). Patches Are All You Need? *arXiv preprint arXiv:2201.09792*.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, 30.

References IV

[Yarbus, 1967] Yarbus, A. L. (1967). Eye Movements During Perception of Complex Objects. In Yarbus, A. L., editor, *Eye Movements and Vision*, pages 171–211. Springer US, Boston, MA.