Attention 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?

What are transformers?

Definition

A transformer is a neural network architecture module that explicitly 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?

Figure: Ilya Repin, An Unexpected Visitor, 1884.

How do we look at an image?

Figure: Experiments on visual attention [Yarbus, 1967]

Tasks:

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

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

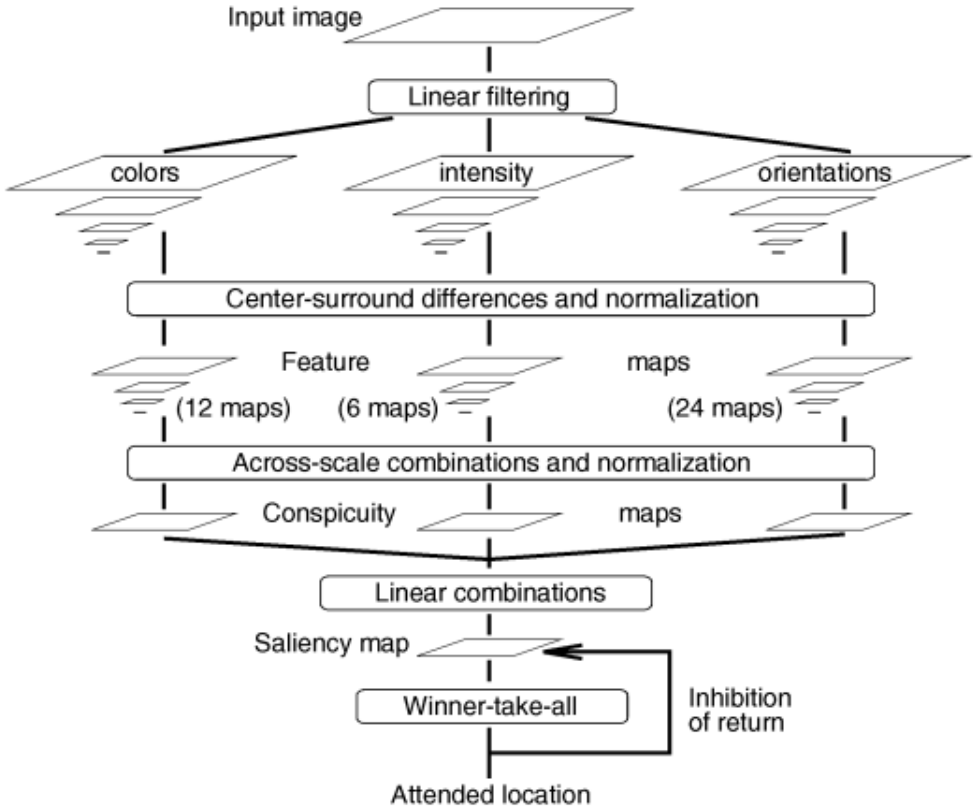
\+ Some degree of invariance

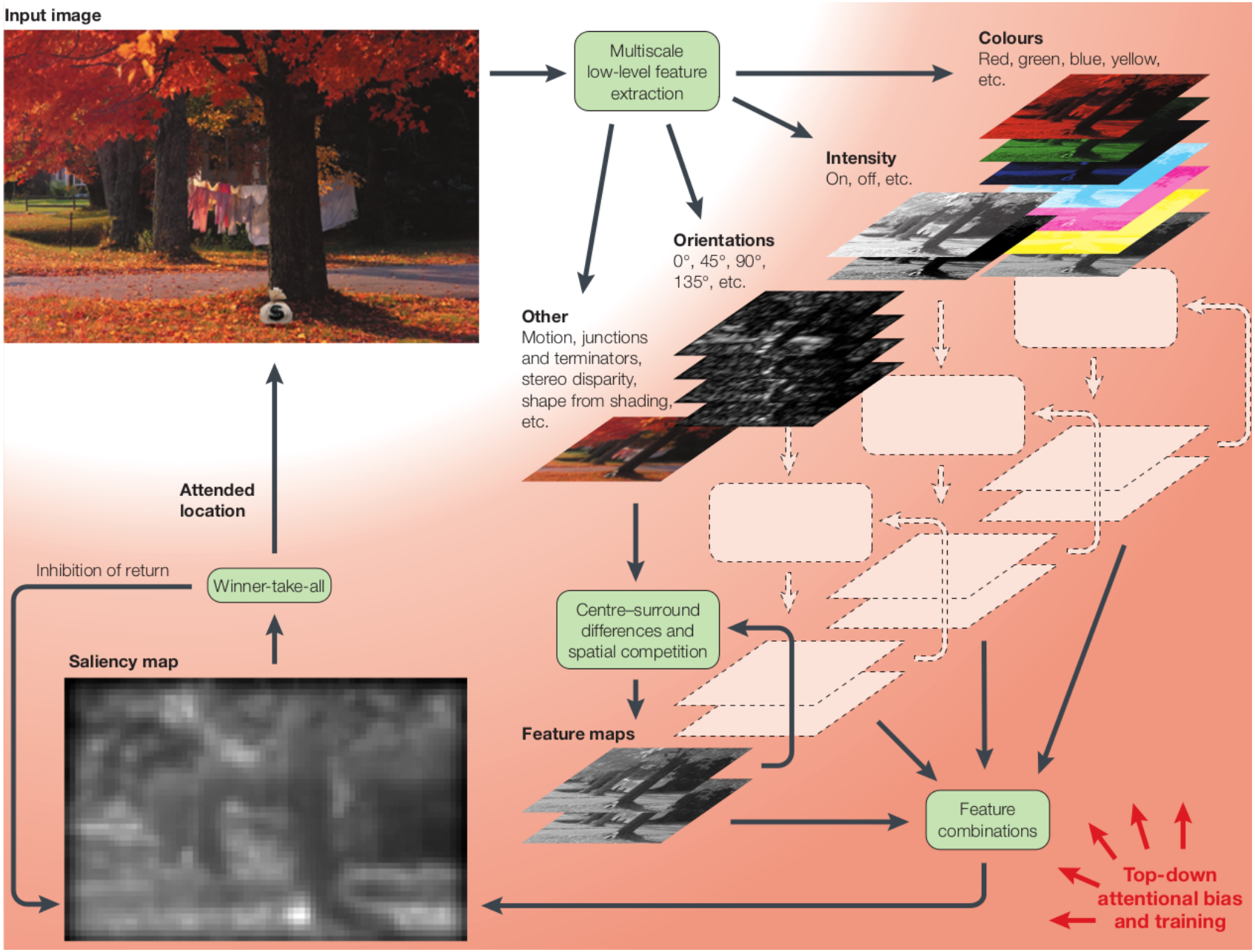
\+ Interpretability

Contents

A classical bottom-up model

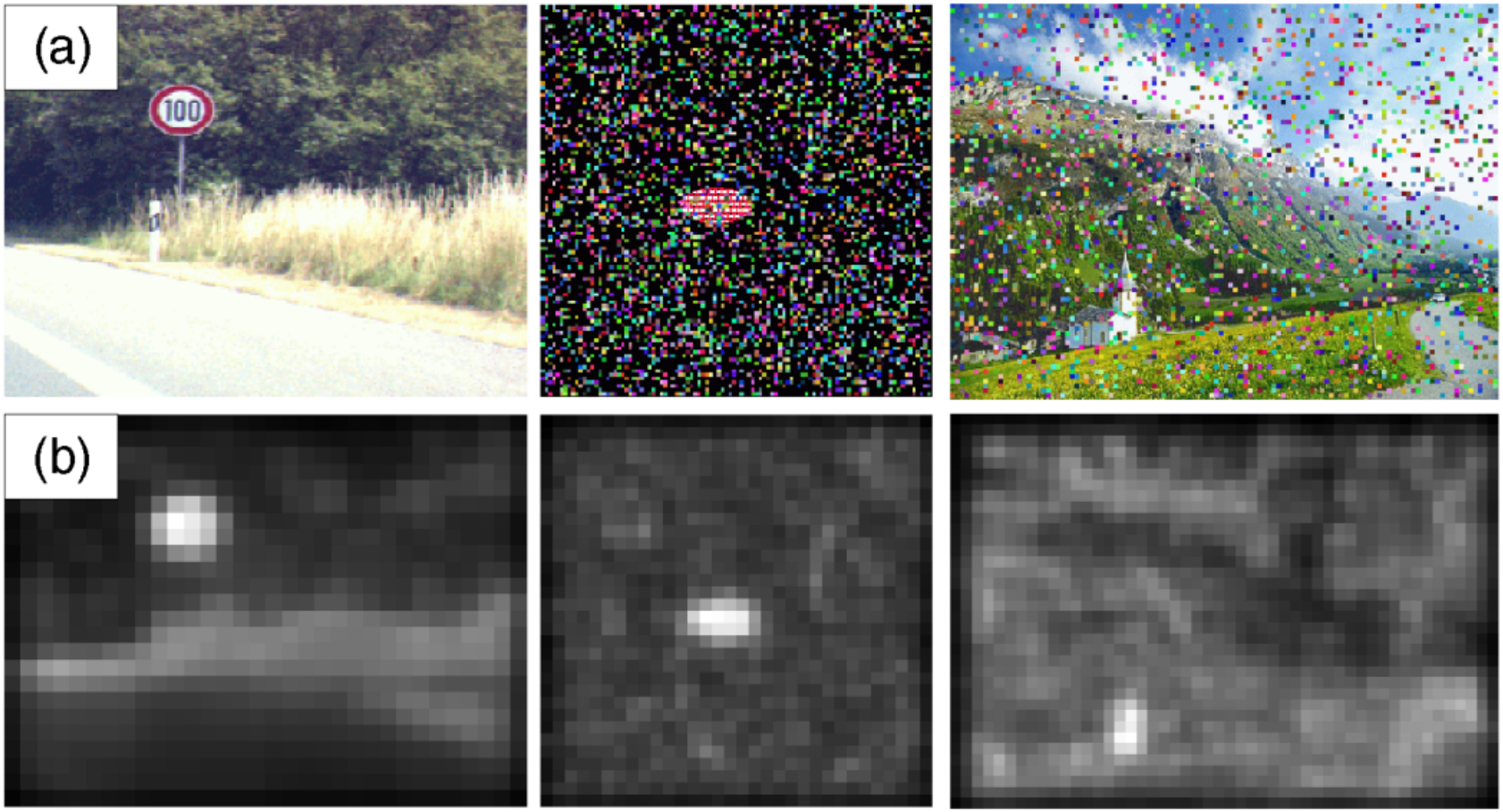
- Itti et al. [Itti et al., 1998] proposed a model inspired by the primate visual system.
- It only uses low-level information.





Credits: [Itti and Koch, 2001]

Examples [Itti et al., 1998]



Top-down attention models

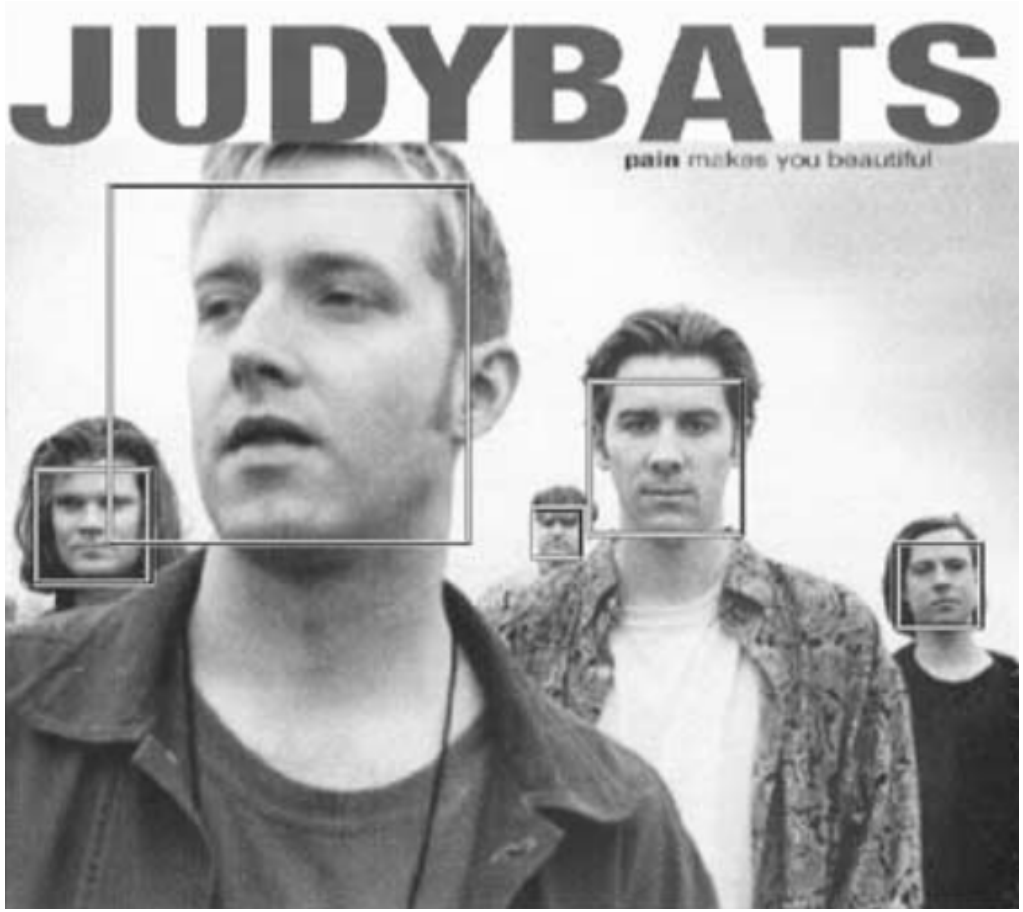
- These are task-dependant.
- Note that all detection methods can be considered as task-oriented attention methods

Example: Face detection with the Viola-Jones method [Viola and Jones, 2001]

- Define weak learners based on integrals on rectangles
- Select learners using AdaBoost
- Apply them in a hierarchical way

Image size: 24×24 pixels

Illustration [Viola and Jones, 2001]

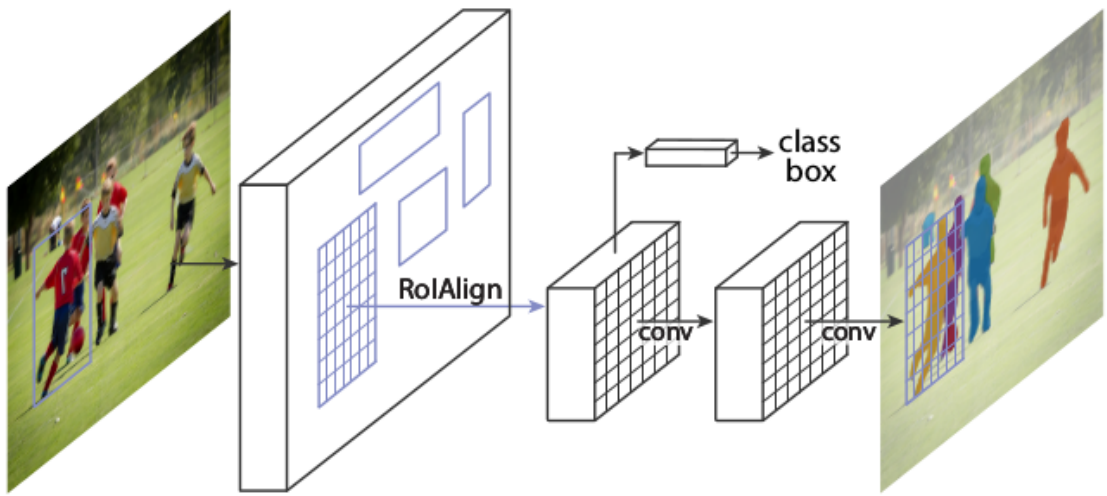


Once attention is focused, the corresponding regions can be further analysed. Here, for identification purposes, for example.

Contents

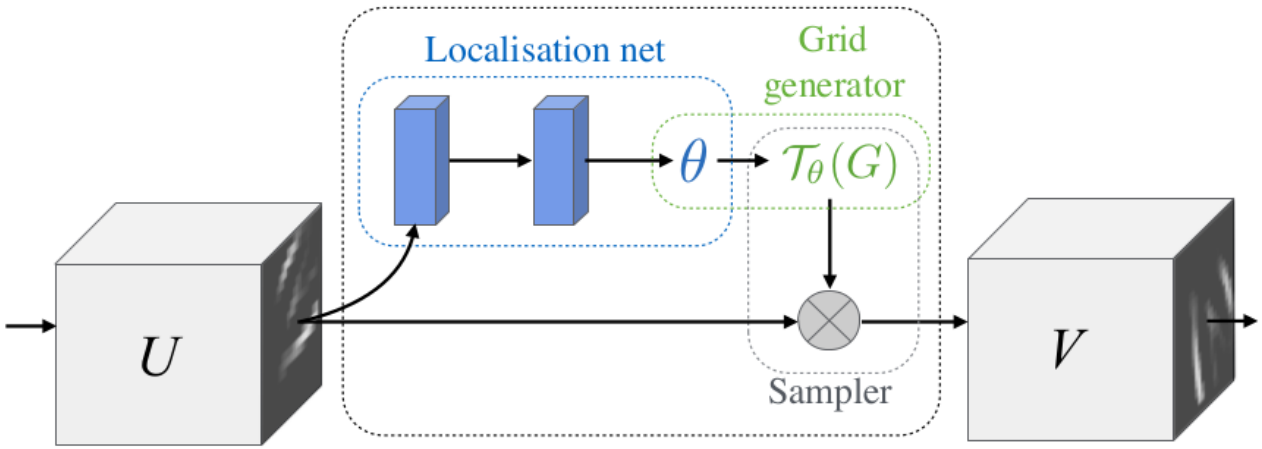
Region proposal networks [Ren et al., 2015]

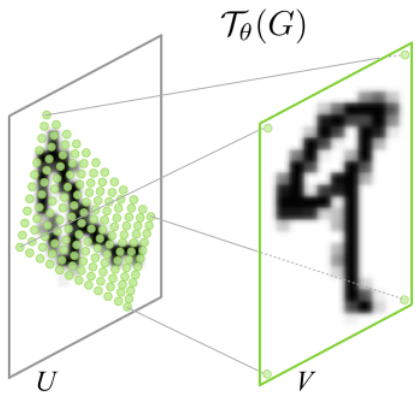
- Detection and instance segmentation methods use region proposal networks, that can be interpreted as an attention mechanism.
- The region proposal network gives the coordinates of the rectangle and a probability that it contains an object.



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

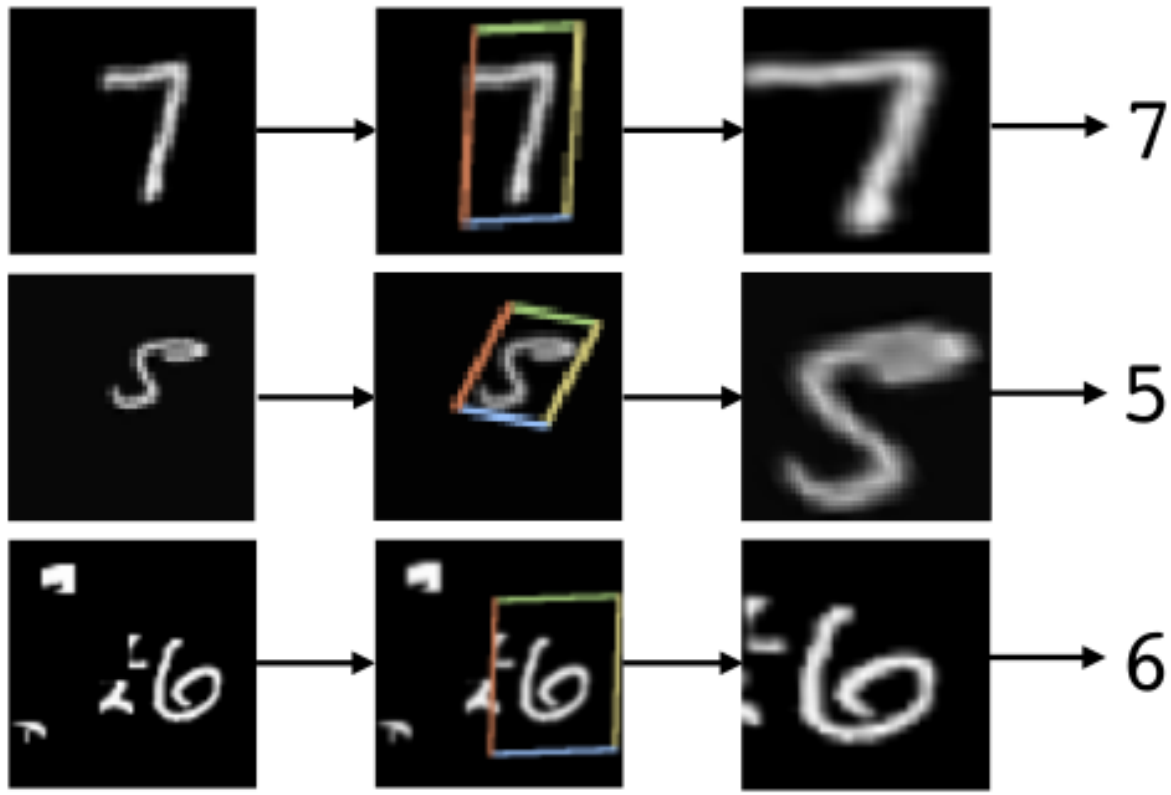
Spatial transformers [Jaderberg et al., 2016]





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

Spatial transformers illustration



Contents

Transformer avatars

Some examples

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

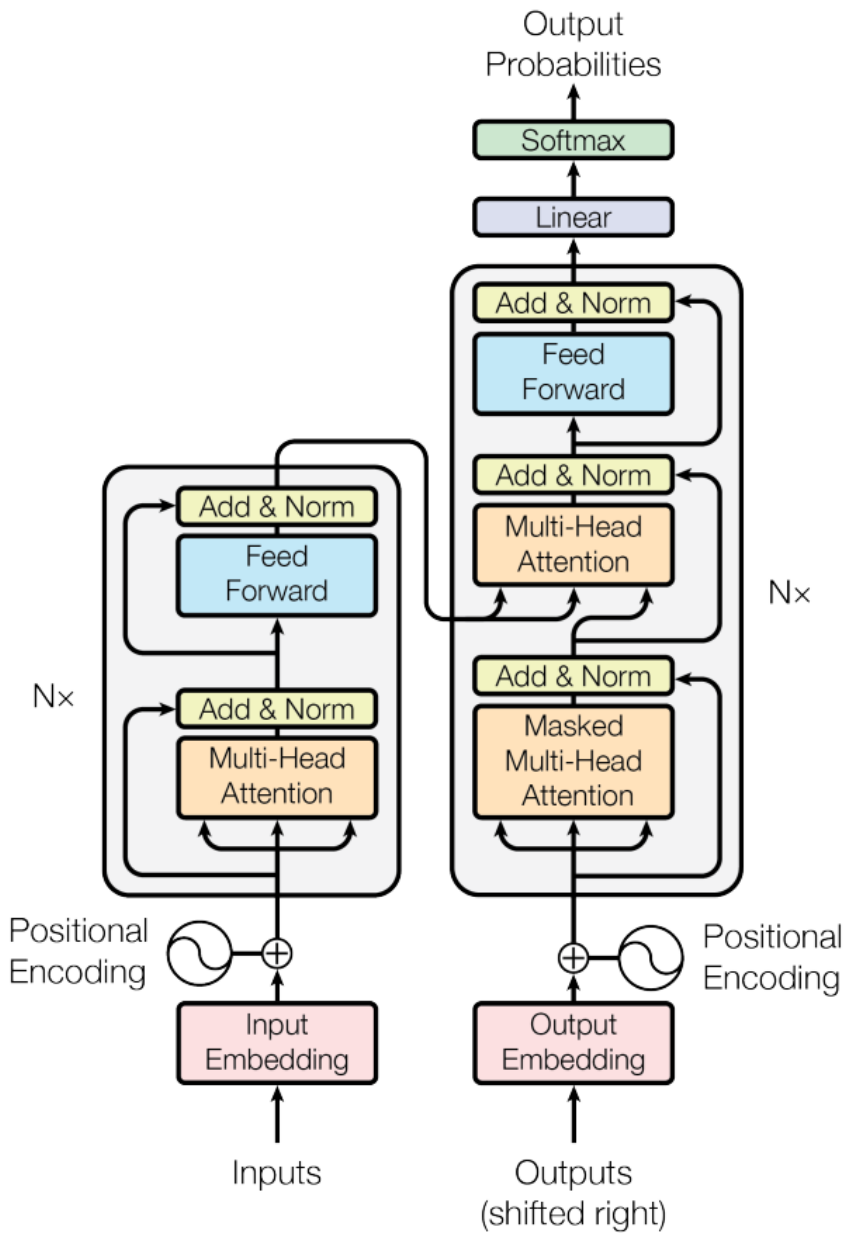
The transformer [Vaswani et al., 2017].

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

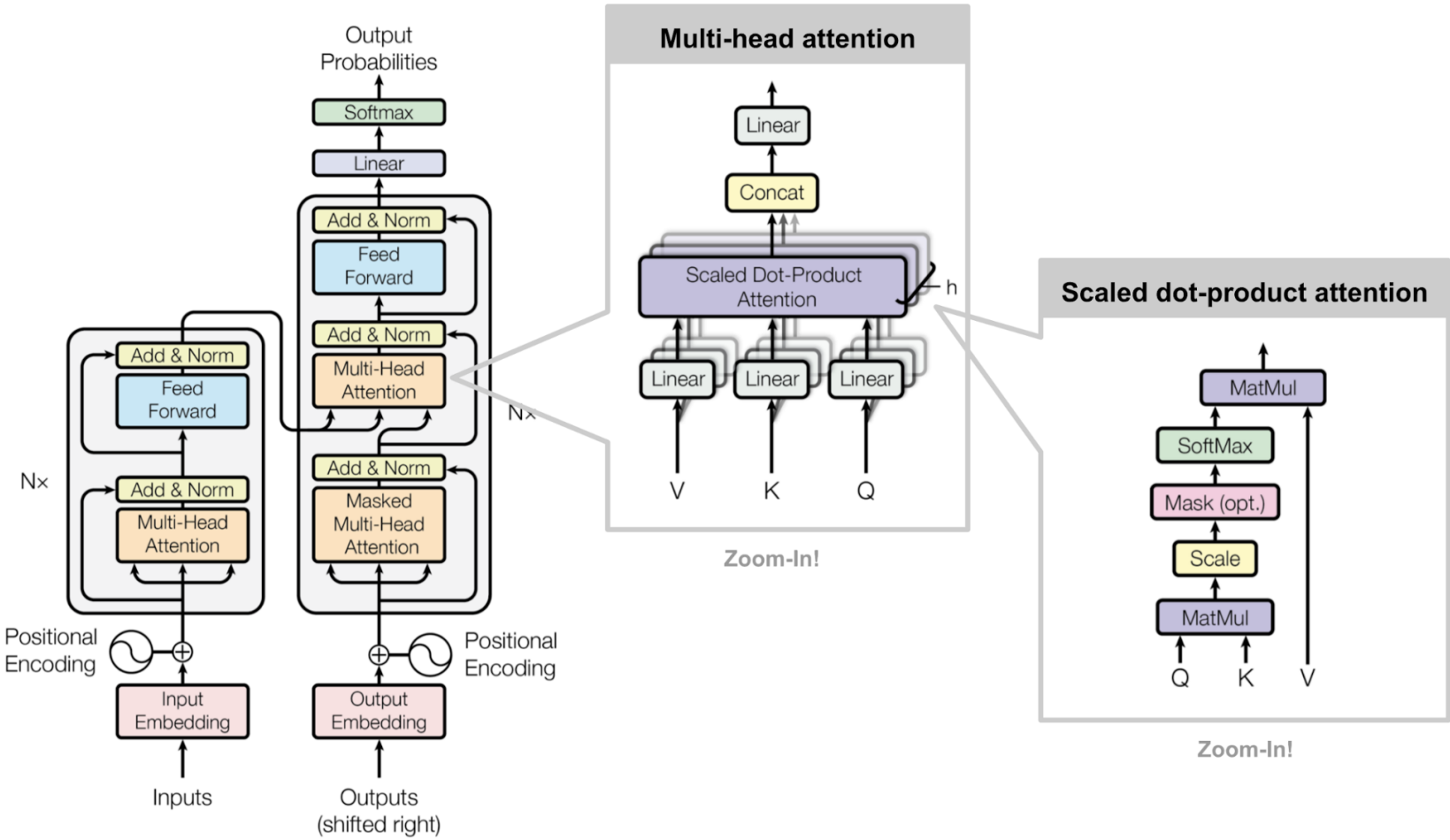
The rise of transformers

The paper that started it all

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



Architecture [Vaswani et al., 2017]



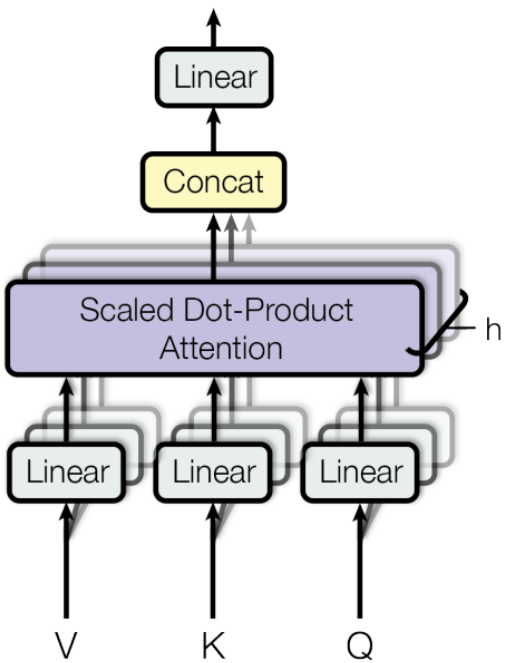
Credits: https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html

Multi-head attention

Dot-product attention

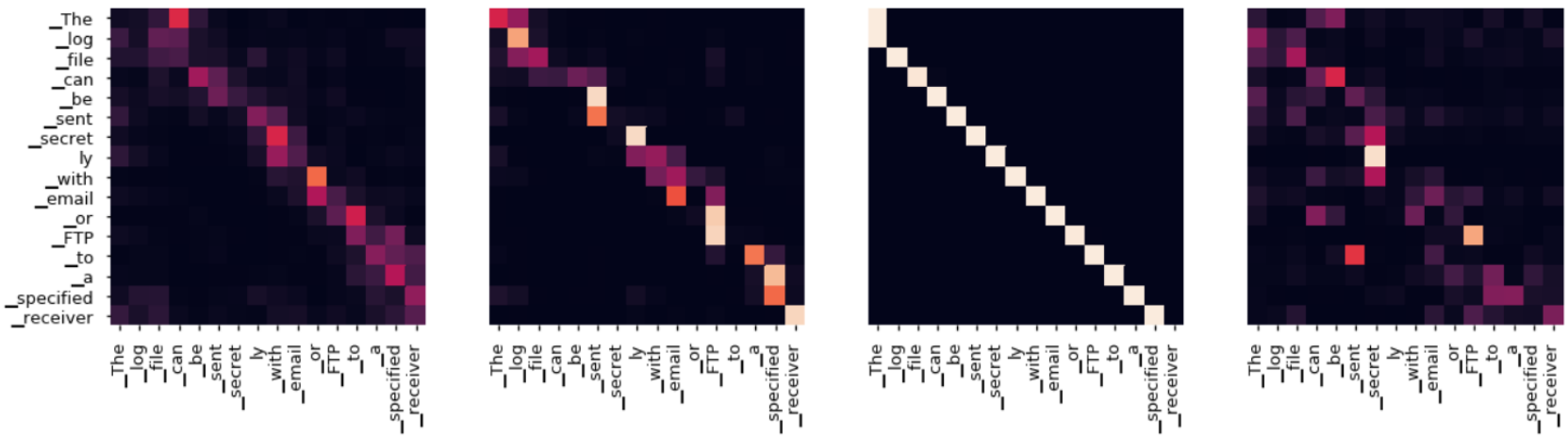
$$Att(Q,K,V) = softmax\left(\frac{QW_Q(KW_V)^t}{\sqrt{d_{K'}}}\right)VW_V$$

$$Att(Q,K,V) = softmax\left(\frac{Q'(K')^t}{\sqrt{d_{K'}}}\right)V'$$



- Matrices W_Q, W_V and W_V are learnable.
- $d_{K'}$ is the length of K'.

Dot-product attention illustration



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

Contents

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.

[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.

[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 and Koch, 2001] Itti, L. and Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203. Number: 3 Publisher: Nature Publishing Group.

References II

[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.

[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.

[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.

[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.

[Viola and Jones, 2001] Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. ISSN: 1063-6919.

References III

[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.