Attention transformers

E. Decencière

MINES ParisTech
PSL Research University
Center for Mathematical Morphology

Contents

Contents

Transformers: a new revolution in deep learning?

- Transformers 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?

A fist quick definition

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

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?

- 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?

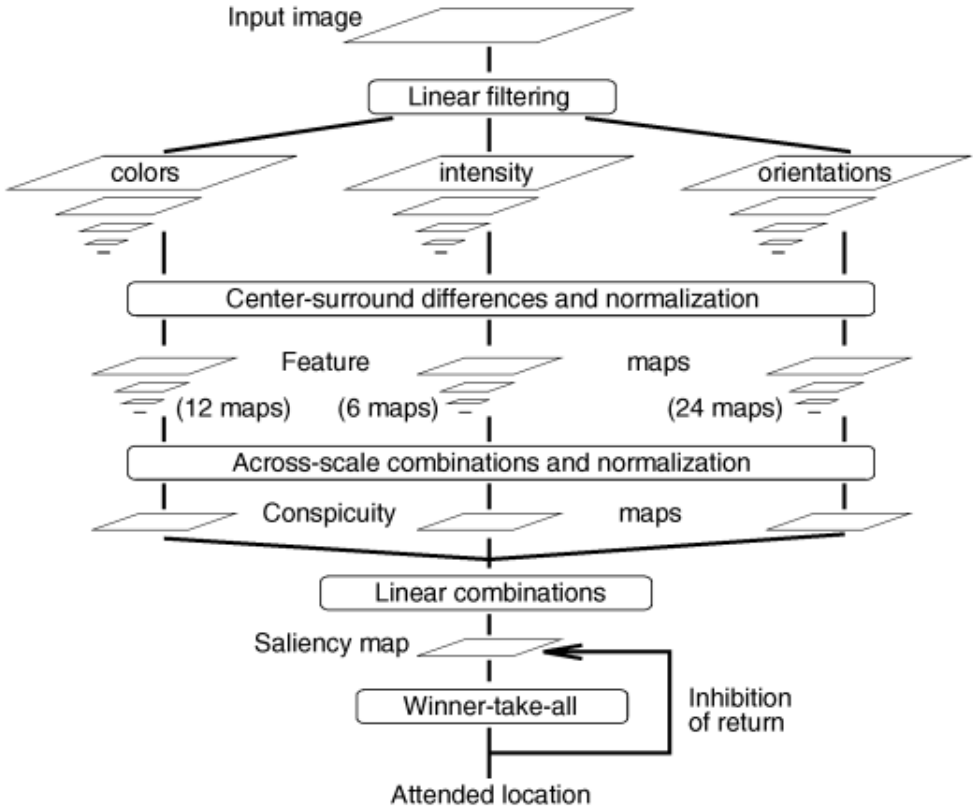
- Photoreceptor cells are expensive
- Processing power is limited
- Solution: concentrate the cells in a given region and use the gaze to optimize their use

- The same arguments apply to artificial visual systems

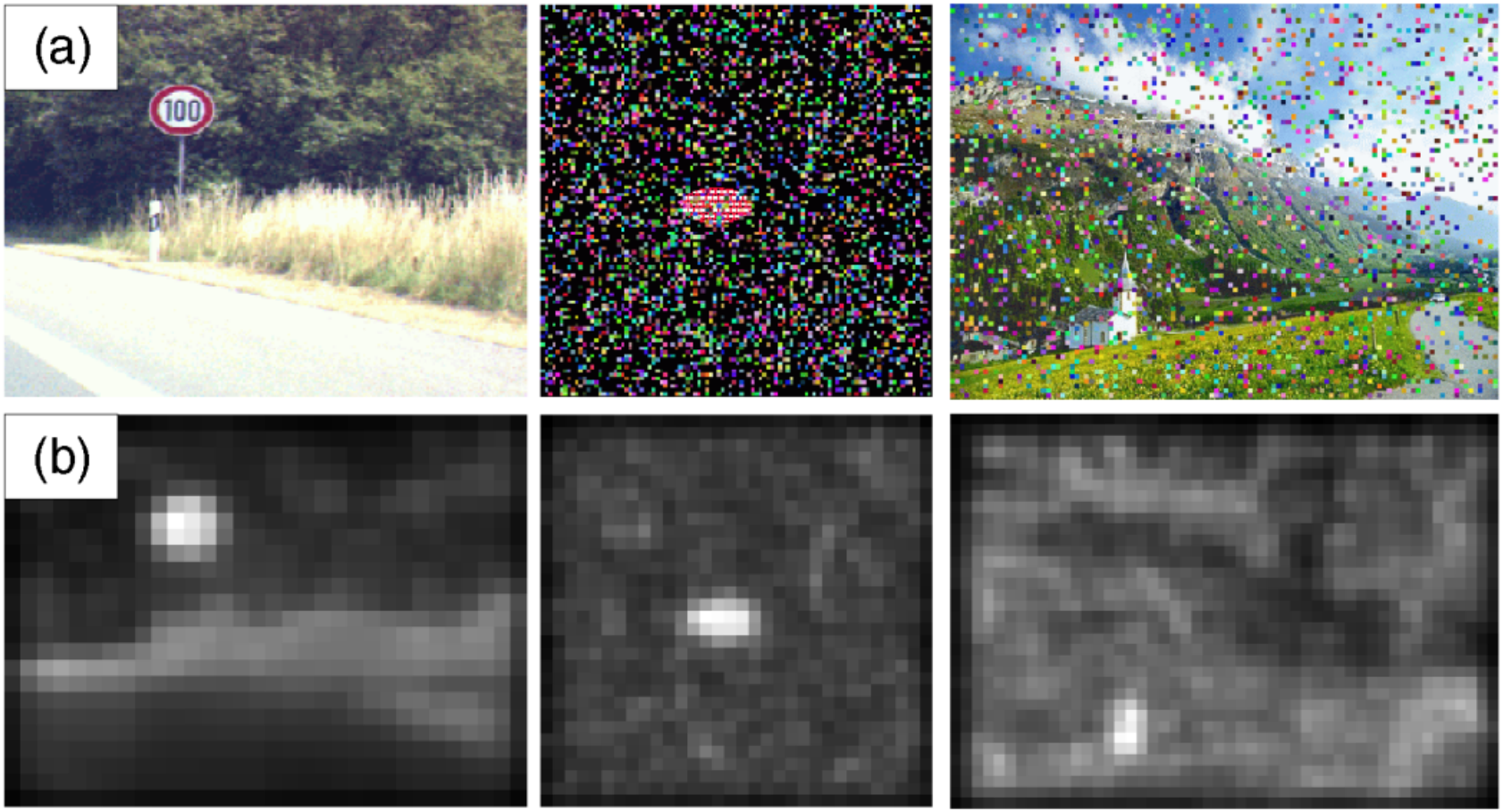
Contents

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]



Top-down attention models

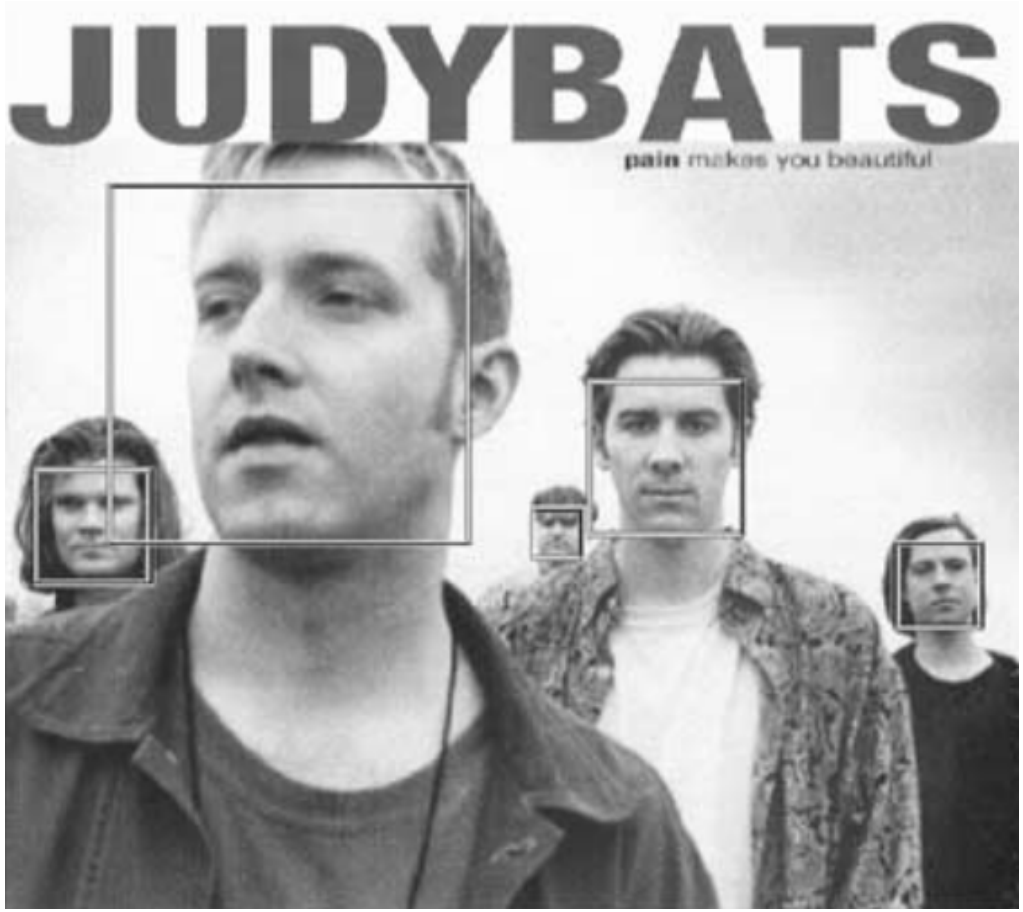
- These are task-dependant.
- Note that all detection methods can be considered as task-oriented attention methods

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

Contents

Historical land-marks

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

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.

[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.

References II

[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.

[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.

[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.