A Neural Algorithm of Artistic Style

Leon A. Gatys, Alexander S. Ecker, Matthias Bethge

26 August 2015

Problems Addressed

1. The algorithmic basis of creating new images through interplay of content and style is not well known.
2. This paper introduces a Deep Neural Network architecture capable of generating artistic pictures of high perceptual quality.
3. In doing so, the paper also tries to provide and algorithmic insight to how humans create and perceive artistic imagery.
4. The previous approaches for doing so relied on non-parametric techniques to directly manipulate pixel representation. By using Deep Neural networks, manipulations are carried out in feature spaces that represent the high level content of an image.

Proposed Solution

1. The main idea is to use a Convolutional Neural Network to separate and then recombine content and style of arbitrary images, which is essentially a neural algorithm for creation of artistic images.
    - Image content and style cannot be completely disentangled, however.
    - Therefore, when synthesising an image with matches the content of one and style of another, there usually doesn't exist an image which may satisfy both constraints simultaneously.
    - Therefore, the loss function that is minimized contains two terms - one for content and another for style.
2. The style representation of an image is a multi-scale representation that includes multiple layers of the neural network.
    - The final results comprised of style representations from the entire network hierarchy.
    - Style can be defined more locally by taking smaller number of lower layers. Different choices would lead to different visual results.
    - When matching the style representations up to higher layers in the network, local images structures are matched on an increasingly large scale, leading to a smoother and more continuous visual experience.

Experiments

1. The results presented in this text were obtained using the VGGNet.
    - The used feature space was provided by 16 Convolutional layers and 5 Pooling layers of the network.
    - No fully connected layers were used.

- Max-pooling was replaced by Average-pooling as it improved gradient flow and leads to more visually appealing results.

2. A layer with N_l distinct filters has N_l feature maps each of size M_l, where M_l is the height times the width of the feature map. So the responses in a layer l can be stored in a matrix $F^l \in R^{N_l \times M_l}$ where F^l_{ij} is the activation of the i^{th} filter at position j in layer l. To visualise the image information that is encoded at different layers of the hierarchy (Fig 1, content reconstructions) we perform gradient descent on a white noise image to find another image that matches the feature responses of the original image. So let $\vec{p}$ and $\vec{x}$ be the original image and the image that is generated and P^l and F^l their respective feature representation in layer l. We then define the squared-error loss between the two feature representations:

$$\mathcal{L}_{content}(\vec{p}, \vec{x}, l) = \frac{1}{2} \sum_{i,j} (F^l_{ij} - P^l_{ij})^2$$

   The derivative of this loss with respect to the activations in layer l equals

$$\frac{\partial \mathcal{L}_{content}}{\partial F^l_{ij}} = \begin{cases} (F^l - P^l)_{ij} & if\ F^l_{ij} > 0 \\ 0 & if\ F^l_{ij} < 0 \end{cases}$$

   Thus we can change the initially random image $\vec{x}$ until it generates the same response in a certain layer of the CNN as the original image $\vec{p}$.

   These feature correlations are given by the Gram matrix $G_l \in R^{N_l \times N_l}$, where G^l_{ij} is the inner product between the vectorised feature map i and j in layer l:

$$G^l_{ij} = \sum_k F^l_{ik} F^l_{jk}$$

   To generate a texture that matches the style of a given image (Fig 1, style reconstructions), we use gradient descent from a white noise image to find another image that matches the style representation of the original image. This is done by minimising the mean-squared distance between the entries of the Gram matrix from the original image and the Gram matrix of the image to be generated. So let $\vec{a}$ and $\vec{x}$ be the original image and the image that is generated and A^l and G^l their respective style representations in layer l. The contribution of that layer to the total loss is then

$$E_l = \frac{1}{4N_l^2 M_l^2} \sum_{ij} (G^l_{ij} - A^l_{ij})^2$$

   and the total loss is

$$\mathcal{L}_{style}(\vec{a}, \vec{x}) = \sum_{l=0}^{L} w_l E_l$$

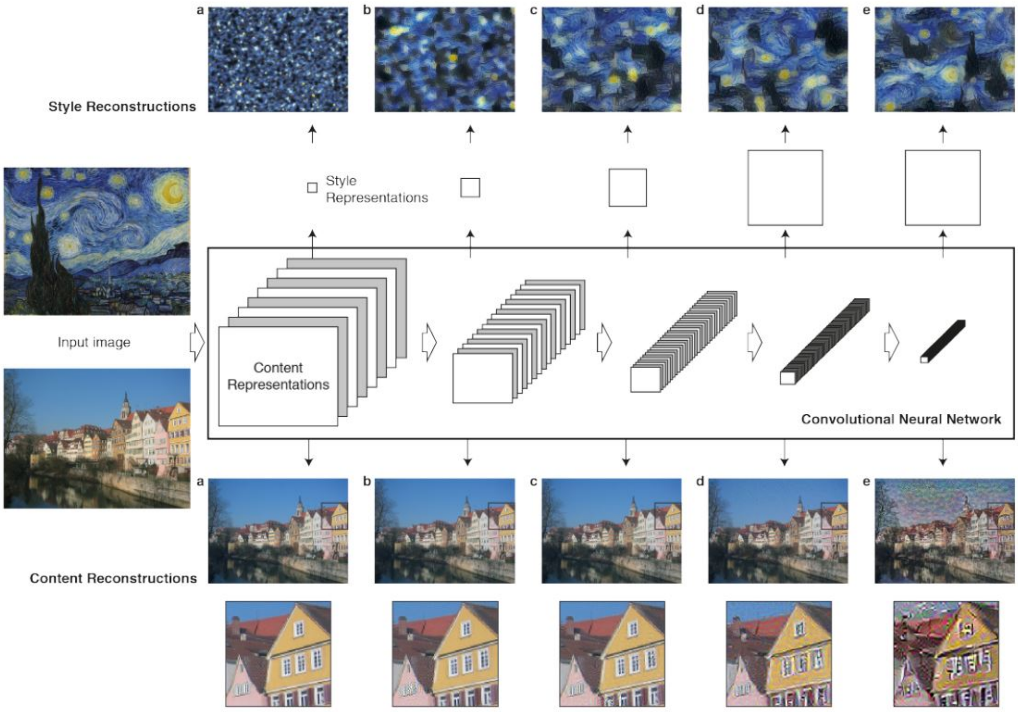
   where w_l are weighing factors of the contribution of each layer to the total loss.

Results

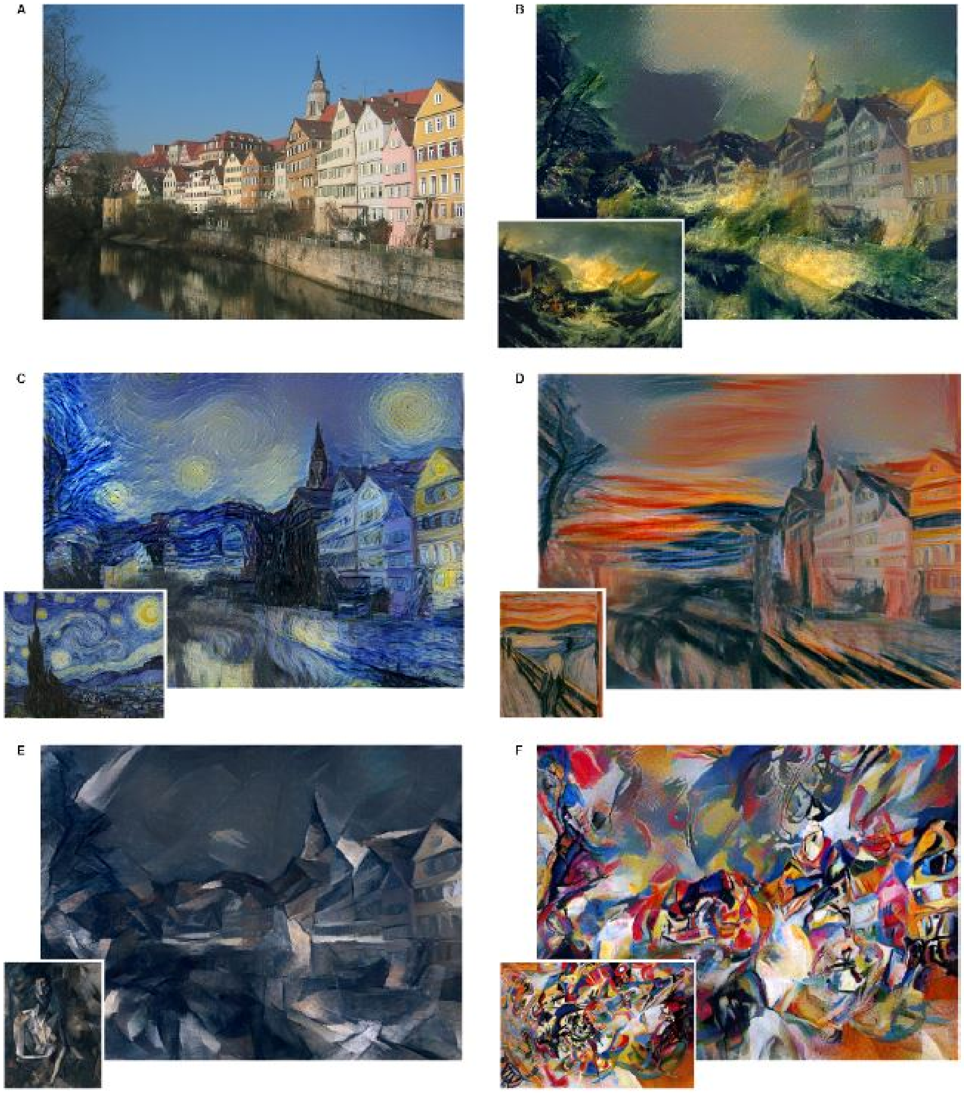
1. The representations of content and style in a Convolutional Neural network are separable. This allows independent manipulation of both representations to produce new images.

2. The synthesised images simultaneously represent the content information of the photograph and the style information of the artwork.

3. The global arrangement of the image providing the content is preserved - only the colours and local structures that compose the scenery are provided by the artwork (spatial information of the artwork is lost.)



4. The five content reconstructions in Fig 1 are from layers 'conv1 1' (a), 'conv2 1' (b), 'conv3 1' (c), 'conv4 1' (d) and 'conv5 1' (e)


A
B
C
D
E
F

Claims

1. Features from Deep Neural Networks trained on object recognition have been previously used for style recognition in order to classify artworks according to the period in which they were created.

    - There, classifiers are trained on top of the raw network activations, which we call content representations.
    - The authors of this paper conjecture that a transformation into a stationary feature space such as their style representation might achieve even better performance in style classification.

2. The authors also envision that this will be useful for a wide range of experimental studies concerning visual perception ranging from psychophysics over functional imaging to even electrophysiological neural recordings.

3. The mathematical form of our style representations generates a clear, testable hypothesis about the representation of image appearance down to the single neuron level.

    - The style representations simply compute the correlations between different types of neurons in the network.
    - Extracting correlations between neurons is a biologically plausible computation that is, for example, implemented by so-called complex cells in the primary visual system (V1).

- Results suggest that performing a complex-cell like computation at different processing stages along the ventral stream would be a possible way to obtain a content-independent representation of the appearance of a visual input.

4. A neural system, which is trained to perform one of the core computational tasks of biological vision, automatically learns image representations that allow the separation of image content from style.

   The explanation could be that when learning object recognition, the network has to become invariant to all image variation that preserves object identity.

   Hence, Representations that factorise the variation in the content of an image and the variation in its appearance would be extremely practical for this task.

Bibliography

1. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, 1097–1105 (2012). URL http://papers.nips.cc/paper/4824-imagenet.

2. Taigman, Y., Yang, M., Ranzato, M. & Wolf, L. Deepface: Closing the gap to human-level performance in face verification. In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, 1701–1708 (IEEE, 2014). URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6909616.

3. G¨uc¸l¨u, U. & Gerven, M. A. J. v. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. The Journal of Neuroscience 35, 10005–10014 (2015). URL http://www.jneurosci.org/content/35/27/10005.

4. Yamins, D. L. K. et al. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the National Academy of Sciences 201403112 (2014). URL http://www.pnas.org/content/early/2014/05/08/1403112111.

5. Cadieu, C. F. et al. Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. PLoS Comput Biol 10, e1003963 (2014). URL http://dx.doi.org/10.1371/journal.pcbi.1003963.

6. K¨ummerer, M., Theis, L. & Bethge, M. Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet. In ICLR Workshop (2015). URL /media/publications/1411.1045v4.pdf.

7. Khaligh-Razavi, S.-M. & Kriegeskorte, N. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. PLoS Comput Biol 10, e1003915 (2014). URL http://dx.doi.org/10.1371/journal.pcbi.1003915.

8. Gatys, L. A., Ecker, A. S. & Bethge, M. Texture synthesis and the controlled generation of natural stimuli using convolutional neural networks. arXiv:1505.07376 [cs, q-bio] (2015). URL http://arxiv.org/abs/1505.07376. ArXiv: 1505.07376.

9. Mahendran, A. & Vedaldi, A. Understanding Deep Image Representations by Inverting Them. arXiv:1412.0035 [cs] (2014). URL http://arxiv.org/abs/1412.0035. ArXiv: 1412.0035.

10. Heeger, D. J. & Bergen, J. R. Pyramid-based Texture Analysis/Synthesis. In Proceedings of the 22Nd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '95, 229–238 (ACM, New York, NY, USA, 1995). URL http://doi.acm.org/10.1145/218380.218446.

11. Portilla, J. & Simoncelli, E. P. A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients. International Journal of Computer Vision 40, 49–70 (2000). URL http://link.springer.com/article/10.1023/A%3A1026553619983.

12. Tenenbaum, J. B. & Freeman,W. T. Separating style and content with bilinear models. Neural computation 12, 1247–1283 (2000).
URL http://www.mitpressjournals.org/doi/abs/10.1162/089976600300015349.

13. Elgammal, A. & Lee, C.-S. Separating style and content on a nonlinear manifold. In Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, vol. 1, I–478 (IEEE, 2004).
URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1315070.

14. Kyprianidis, J. E., Collomosse, J., Wang, T. & Isenberg, T. State of the "Art": A Taxonomy of Artistic Stylization Techniques for Images and Video. Visualization and Computer Graphics, IEEE Transactions on 19, 866–885 (2013).
URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6243138.

15. Hertzmann, A., Jacobs, C. E., Oliver, N., Curless, B. & Salesin, D. H. Image analogies. In Proceedings of the 28th annual conference on Computer graphics and interactive techniques, 327–340 (ACM, 2001).
URL http://dl.acm.org/citation.cfm?id=383295.

16. Ashikhmin, N. Fast texture transfer. IEEE Computer Graphics and Applications 23, 38–43 (2003).

17. Efros, A. A. & Freeman, W. T. Image quilting for texture synthesis and transfer. In Proceedings of the 28th annual conference on Computer graphics and interactive techniques, 341–346 (ACM, 2001).
URL http://dl.acm.org/citation.cfm?id=383296.

18. Lee, H., Seo, S., Ryoo, S. & Yoon, K. Directional Texture Transfer. In Proceedings of the 8th International Symposium on Non-Photorealistic Animation and Rendering, NPAR '10, 43–48 (ACM, New York, NY, USA, 2010).
URL http://doi.acm.org/10.1145/1809939.1809945.

19. Xie, X., Tian, F. & Seah, H. S. Feature Guided Texture Synthesis (FGTS) for Artistic Style Transfer. In Proceedings of the 2Nd International Conference on Digital Interactive Media in Entertainment and Arts, DIMEA '07, 44–49 (ACM, New York, NY, USA, 2007).
URL http://doi.acm.org/10.1145/1306813.1306830.

20. Karayev, S. et al. Recognizing image style. arXiv preprint arXiv:1311.3715 (2013).
URL http://arxiv.org/abs/1311.3715.

21. Adelson, E. H. & Bergen, J. R. Spatiotemporal energy models for the perception of motion. JOSA A 2, 284–299 (1985).
URL http://www.opticsinfobase.org/josaa/fulltext.cfm?uri=josaa-2-2-284.

22. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs] (2014).
URL http://arxiv.org/abs/1409.1556. ArXiv: 1409.1556.

23. Russakovsky, O. et al. ImageNet Large Scale Visual Recognition Challenge. arXiv:1409.0575 [cs] (2014).
URL http://arxiv.org/abs/1409.0575. ArXiv: 1409.0575.

24. Jia, Y. et al. Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the ACM International Conference on Multimedia, 675–678 (ACM, 2014).
URL http://dl.acm.org/citation.cfm?id=2654889.16