Autoencoding *Blade Runner*: Reconstructing Films with Artificial Neural Networks

Terence Broad and Mick Grierson

Terence Broad
Goldsmiths, University of London
8 Lewisham Way
New Cross, London SE14 6NW, U.K.
ma201tb@gold.ac.uk

Mick Grierson
Goldsmiths, University of London
8 Lewisham Way
New Cross, London SE14 6NW, U.K.
m.grierson@gold.ac.uk

See <www.mitpressjournals.org/toc/leon/50/4> for supplemental files associated with this issue.

ABSTRACT

In this paper, the authors explain how they created *Blade Runner—Autoencoded,* a film made by training an autoencoder—a type of generative neural network—to recreate frames from the film *Blade Runner*. The autoencoder is made to reinterpret every individual frame, reconstructing it based on its memory of the film. The result is a hazy, dreamlike version of the original film. The authors discuss how the project explores the aesthetic qualities of the disembodied gaze of the neural network and describe how the autoencoder is also capable of reinterpreting films it has not been trained on, transferring the visual style it has learned from watching *Blade Runner* (1982).

Reconstructing videos based on prior visual information has some scientific and artistic precedents. Casey and Grierson [1] present a system for real-time matching of an audio input stream to a database of continuous audio or video, presenting an application called REMIX-TV. Grierson develops on this work with PLUNDERMATICS [2], adding more sophisticated methods for feature extraction, segmentation, and filtering. Mital, Grierson, and Smith [3] extend this approach further to synthesize a target image using a corpus of images. The image is synthesized in fragments that are matched from the database extracted from the corpus based on shape and color similarity. Mital uses this technique to create a series called "YouTube Smash Up" [4], which synthesizes the week's most popular video on YouTube from fragments of other trending videos. Another, somewhat related approach is the research in reconstructing what people are watching while in an MRI scanner, solely from recorded brain scans [5].

During the development of *Blade Runner—Autoencoded*, we were influenced by this earlier research as we pursued the same goal, while taking advantage of the recent advances in deep generative models (detailed in the next section). The film *Blade Runner* (1982) was chosen as the visual material on which to anchor this research, because of its relation to the themes of perception, artificiality, and artificial intelligence.

Technical Background

Research in deep learning, specifically in the field of computer vision, has been accelerating in recent years, particularly since Krizhevsky, Sutskever, and Hinton's [6] breakthrough in the 2011 ImageNet competition, where they used a sole convolutional neural network to place images into 1,000 possible classifications. Prior to this, all competing entries were a combination of carefully engineered visual features, in tandem with more rudimentary machine-learning algorithms to do classification. This was the first successful approach of a system that learned everything end-to-end in this kind of real-world image-classification scenario.

While it was possible to have powerful image-recognition capabilities using a convolutional neural network, it was not thought possible to reverse this kind of system so that it could be used as a generative model for images. As a result, these systems were often referred to as "black box" systems, partially because there was a certain level of skepticism as to whether these kinds of

   doi:10.1162/LEON_a_01455

systems were seeing things the way humans do. This skepticism was evidenced by the observation that such networks could easily be fooled into incorrectly classifying images that had been subtly manipulated using carefully crafted patterns of noise imperceptible to humans [7]. In response to such observations, there was a drive in the research community towards developing generative models capable of generating realistic natural images, because if a network is capable of generating realistic natural images, it has a greater understanding—or at least we can be more confident it does—of the subject that it is representing.

An autoencoder is one such type of network that can be used as a generative model. It can be thought of as two networks: one that takes input (such as an image) and *encodes* it into a latent (numerical) representation; the other network, which is symmetrical in design, *decodes* the latent representation back into the original data space (reconstructing the image). The network is given images from the dataset to reconstruct and is trained to minimize the loss, which is calculated by the per-pixel difference between the images. An extension to this is the variational autoencoder (VAE) [8,9] that combines this network structure with a variational Bayesian approach to training, which makes strong assumptions concerning the distribution of latent variables (by assuming a Gaussian prior). This forces the autoencoder to use the latent space more efficiently, leading to more robust reconstructions and better generalization.

Generative adversarial networks (GANs) [10] are an altogether different approach to developing a deep generative model. This approach borrows a concept from game theory for the training regime; in this case, two networks are set against each other in a minimax game. One network, the "generator," tries to generate images that fit the distribution of images in the dataset. The second, a "discriminator" network, looks at images (both real and generated) and attempts to maximize the probability of correctly labeling the image as real or generated. Conversely, the generator is trained to try to *fool* the discriminator into thinking it is creating real images. Radford et al. [11] build upon this work by using the same training regime to train deep convolutional neural networks to generate images. This was the first time a convolutional neural network had been effectively inverted and used as a generative model, creating images almost indistinguishable from photographs at low resolutions. (This was accomplished by replacing the traditional structure of the generator network—convolutions alternating with pooling layers—with strided convolutions and fractionally strided backwards convolutions.)

In 2016, Larsen et al. [12] elegantly combined the GAN approach with a variational autoencoder. They used the strided convolution architecture popularized by Radford et al. and combined the training routines of the two approaches. They added a discriminator network to the VAE framework to create a consortium of three networks (encoder, decoder, and discriminator). The discriminator network is used to determine how similar each generated image is to the real image, as opposed to comparing these images on a simple pixel-by-pixel basis. This significantly increases the generative capability of the VAE, optimizing the network to produce images that are perceptually similar, reducing the tendency of the autoencoder framework to generate blurry images. This adversarial-variational autoencoder, trained with a learned similarity metric, was the model used as the basis for this project [13].

Learning the Distribution of Imagery in *Blade Runner*

The standard practice for evaluating deep generative models is to train them on a standard, widely used set of images (usually all of the same subject matter, e.g. handwritten digits [14] or faces [15]). Using these datasets restricts the complexity of what the model needs to represent and allows a direct visual comparison to be made between the results from different models. Taken as a complete set, the frames from *Blade Runner* contain much more variety in terms of subject

matter and perspective than the sort of datasets commonly used to train and evaluate these generative models. Therefore we were initially concerned the model would not be able to represent such a diverse range of imagery with any great efficacy, but after seeing some initial results (Figure 1) we were reassured by a single model's diverse generative capabilities.

Figure 1. Sample of a 64-frame minibatch of reconstructed samples from the network trained on *Blade Runner* after one epoch at a resolution of 96x64. (© Terence Broad)

Initially the model was only trained at a resolution of 96×64 pixels (64×64 was the standard in research at the time). The size of the model was increased to be able to create a video that was watchable online, with the largest possible model that could be represented on a single GPU being 256×144 (coincidentally the smallest resolution allowed on YouTube). By increasing the size of the model, training became much slower and more precarious, and it was more likely that one of the three networks (they all had to learn in unison) would fail, resulting in a sharp, irrevocable degradation of image quality, forcing the process to be started again from the beginning. It took approximately three days for the model to be trained one time on all the frames of the film. (One complete cycle through the dataset is referred to as one epoch.)

After some trial and error, a set of hyperparameters was found that allowed all three networks to learn in a balanced and sustained manner over a long period of time. As shown in Figure 2, there is a gradual improvement in image fidelity after one, three, and six epochs. One novel technical contribution made to this training procedure was to reduce the amount of noise injected into the latent space over the course of training (by gradually reducing the standard deviation of the Gaussian prior) in order for the model to better differentiate between similar frames. (A more detailed, technical account of this training procedure can be found in the original technical report [16].)

Figure 2. Samples after training the model on frames from *Blade Runner* for one epoch (top row), three epochs (middle row), and six epochs (bottom row) at a resolution of 256x144. (© Terence Broad)

Reconstructing *Blade Runner*, One Frame at a Time

After training, the autoencoder is then made to reinterpret (perform a forward pass) on each frame of the film. The reconstructed frames are then resequenced back into a video. The resulting sequence is very dreamlike, drifting in and out of recognition between static scenes that the model remembers well, to fleeting sequences—usually with a lot of movement—that the model barely comprehends. It is no surprise that static scenes are represented so well, as the model has, in effect, seen those scenes many more than six times. In essence, the model is simply overfitting to the training data (caused mostly by training on a highly skewed dataset), something that machine-learning practitioners normally go to a great deal of effort to avoid. In this case though, the aesthetic result of this overfitting is interesting, especially when viewed in contrast to the parts of the film the model struggles to represent.

The flaws in the reconstruction are in and of themselves aesthetically interesting and revealing with respect to the model. An obvious flaw is that the model has a tendency to collapse long sequences with a fixed background into a single representation, even if there is some movement in the scene (Figure 3). This tendency was rectified somewhat by the novel training procedure of gradually reducing the amount of noise injecting into the latent representation over training, but not completely. Ultimately, this is a consequence of the images being so similar that they share nearly the same point in latent space, therefore they cannot be differentiated by the generator network. Without some training procedure to enforce the difference between frames, this will always be a problem.

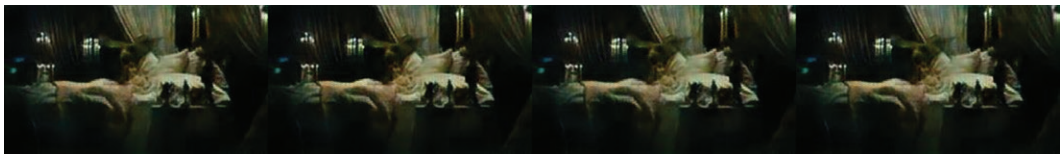

Figure 3. Samples from the reconstruction of *Blade Runner* where the network has collapsed one long sequence with some movement into a single representation. (© Terence Broad)

One curious outcome is the model's inability to represent completely black frames (Figure 4). When asked to recreate a black frame, it instead produces an image with a greenish haze (reminiscent of the phenomenon of seeing colors when one's eyes are closed). This is likely due to the dataset containing very few completely black frames and could certainly be rectified by appending the training dataset with many black images; this was not done, however, as the existing outcome was deemed interesting.

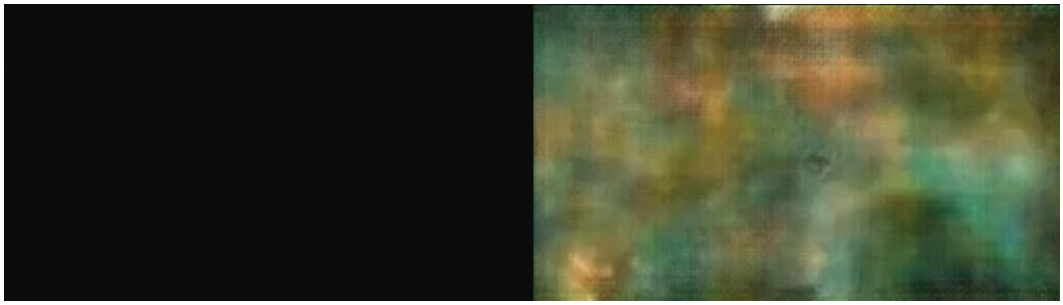

Figure 4. Left: A completely black image. Right: The *Blade Runner*–trained model's interpretation of the completely black image. (© Terence Broad)

Reconstructing Other Films with the *Blade Runner* Model

Once trained, the autoencoder can process frames from any film. The model reinterprets any given set of images based on what it has learned from *Blade Runner*, thus transferring the distinctive "neo-noir" aesthetic onto any video. Figure 5 shows frames from Dziga Vertov's 1929 documentary *Man with a Movie Camera* as reinterpreted by the model. The film is black and white, but the output from the model is in color and is consistent with the visual style of *Blade Runner*.

The reconstructions of other films [17] are aesthetically interesting and unpredictable, but it is difficult to make out what is being represented most of the time. Since this project was carried out, research has been published that uses conditional adversarial networks to translate images from one domain into another [18], providing a more formally defined and effective method to do this kind of image translation.

Figure 5. Top row: Frames from the 1929 film *Man with a Movie Camera*. Bottom row: Frames reinterpreted by the model trained on *Blade Runner*. (Images from *Man with a Movie Camera* are sourced from Wikimedia Commons and are in the public domain.)

Why *Blade Runner*?

The film *Blade Runner* was adapted from Philip K. Dick's novel, *Do Androids Dream of Electric Sheep?* [19]. The story is set in a post-apocalyptic dystopian future where Rick Deckard, the main character, is a bounty hunter who makes a living hunting down and killing replicants—built to be used as slaves on outer-world colonies, but not allowed on Earth. Replicants are so well engineered that they are physically indistinguishable from human beings. Deckard is called back from retirement to hunt down a group of Nexus-6 replicants, the newest model of replicant produced by the Tyrell Corporation.

Because replicants are physically indistinguishable from humans, Deckard has to issue the "Voight-Kampff" test in order to distinguish them. The test is a series of increasingly difficult moral questions about human and animal suffering, with the intention of eliciting an empathic response from humans, but not from androids. With the technological advances of the Nexus-6 replicants, it is increasingly difficult for Deckard to determine who is human and who is not; Deckard feels a growing suspicion that he may not be human himself.

By reinterpreting *Blade Runner* with an artificial neural network's memory of the film, *Blade Runner—Autoencoded* seeks to emphasize the ambiguous boundary in the film between replicant and human, or, in the case of the reconstructed film, between our memory of the film and the neural network's. Some of the flaws in its visual reconstruction are reminiscent of the deficiencies of our own, especially regarding memories of dreams. There is a theory that Dick structured his novel around the work of the great French philosopher René Descartes, with Deckard acting out Descartes's philosophical dilemmas [20]. Descartes emphasized that the senses (our primary source of knowledge) are often prone to error. By examining this imperfect reconstruction of *Blade Runner*—as seen through the gaze of a disembodied machine—it becomes easier to acknowledge the flaws in our own internal representations of the world and easier to imagine the potential of other, substantially different systems that could have their own internal representations.

Outcomes

Blade Runner—Autoencoded and a report on the project were first published online in May 2016 and gained a great deal of attention on social media (with over 200,000 views on YouTube). The project was discussed in several online news articles (most notably by Aja Romano in Vox [21]). After the online publication, the autoencoder was trained for a further 20 epochs to create a second version of the film (Figure 6), which was also upscaled into high resolution to make the work watchable on larger screens. This version of the work was shown at Art Center NABI, Seoul, in the exhibition Why Future Still Needs Us: AI and Humanity, a survey of contemporary artworks (all made in 2016) that incorporate modern machine-learning techniques.

Figure 6. A screenshot from the updated version of *Blade Runner—Autoencoded*, created with the autoencoder trained an additional 20 times on the film. (© Terence Broad)

This work was also featured in, and screened as part of the accompanying film program for, the exhibition Dreamlands: Immersive Cinema and Art, 1905–2016, at The Whitney Museum of American Art in New

York. The exhibition brought together the work of artists articulating the shifts that have taken place as technology has altered the ways in which space and image are constructed and experienced, engaging with the fact that we are now living in an environment more radically transformed by technology than at any other point in human history [22]. For Chrissie Iles, the Anne and Joel Ehrenkranz curator at The Whitney, the work "occupies a unique position, as both a work of science and a work of art," and "belongs to the current moment in which artists are engaging with questions of where the boundary between AI and human perception lies" [23]. Iles relates the work to what Hito Steyerl describes as the "disembodied, post-humanized gaze, outsourced to machines and other objects" [24].

The same autoencoding technique was used in the 2017 film *Geomancer*, made in collaboration with the artist Lawrence Lek [25]. Autoencoding was used in the penultimate dream sequence, to visualize the mental representation of the AI protagonist. In the summer of 2017, *Blade Runner—Autoencoded* will be included in the exhibition Into the Unknown: A Journey Through Science Fiction, at The Barbican in London, and in the exhibition's subsequent international tour.

References and Notes

1. M. Casey and M. Grierson, "Soundspotter/remix-tv: fast approximate matching for audio and video performance," Proceedings of the *International Computer Music Conference* (2007).
2. M. Grierson, "Plundermatics: real-time interactive media segmentation for audiovisual analysis, composition and performance," Proceedings of *Electronic Visualisation and the Arts Conference*, Computer Arts Society, London (2009).
3. P.K. Mital, M. Grierson, and T.J. Smith, "Corpus-based visual synthesis: an approach for artistic stylization," Proceedings of the *ACM Symposium on Applied Perception* (2013) pp. 51–58.
4. P.K. Mital, YouTube Smash Up (2014), <http://pkmital.com/home/youtube-smash-up/>.
5. S. Nishimoto, et al., "Reconstructing visual experiences from brain activity evoked by natural movies," *Current Biology* **21**, No. 19, 1641–1646 (2011).
6. A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in Neural Information Processing Systems (2012) pp. 1097–1105.
7. C. Szegedy, et al., "Intriguing properties of neural networks," The International Conference on Learning Representations (2014).
8. D.P. Kingma and M. Welling, "Auto-encoding variational Bayes," The International Conference on Learning Representations (2014).
9. D.J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," The International Conference on Machine Learning (2014) pp. 1278–1286.
10. I. Goodfellow, et al., "Generative Adversarial Nets," *Advances in Neural Information Processing Systems* (2014) pp. 2672–2680.
11. A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," The International Conference on Learning Representations (2016).
12. A.B. Larsen, S.K. Sønderby, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," The International Conference on Machine Learning (2016) pp. 1558–1566.
13. The original source code for this project is available at <https://github.com/terrybroad/Learned-Sim-Autoencoder-For-Video-Frames>.
14. Wikipedia article for the MNIST handwritten digits dataset is available at <https://en.wikipedia.org/wiki/ MNIST_database>.
15. The CelebFaces dataset was created and first discussed by the authors of this paper: Z. Liu, et al., "Deep learning face attributes in the wild," Proceedings of the *IEEE International Conference on Computer Vision* (2015) pp. 3730–3738.
16. T. Broad and M. Grierson, "Autoencoding Video Frames," Technical Report (London: Goldsmiths, 2016), available at <http://research.gold.ac.uk/19559/>.

17. A side-by-side comparison of *Man with a Movie Camera* and its reconstruction using the Blade Runner model, as well as other films such as *A Scanner Darkly and Koyaanisqatsi* are available to watch online at the following YouTube playlist: <https://www.youtube.com/ playlist?list=PLJME4hivCPY_B_MqOyQQGC_kuYUz518-C>.
18. P. Isola, J. Zhu, T. Zhou, and A. Efros, "Image-to-image translation with conditional adversarial networks," arXiv preprint arXiv:1611.07004 (2016).
19. P.K. Dick, *Do Androids Dream of Electric Sheep?* (New York: Random House USA, 1982).
20. J. Brandt, "What defines human?" (2000), <http://www.br-insight.com/what-defines-human>.
21. A. Romano, "A guy trained a machine to 'watch' *Blade Runner*. Then things got seriously sci-fi" (2016), available at <http://www.vox.com/2016/6/1/11787262/blade-runner-neural-network-encoding>.
22. C. Iles, "The Cyborg and the Sensorium," *Dreamlands: Immersive Cinema and Art, 1905–2016* (New Haven: Yale University Press, 2016) p. 121.
23. C. Iles, personal communication, 2017.
24. H. Steyerl, "In Free Fall: A Thought Experiment on Vertical Perspective," *The Wretched of the Screen* (Berlin: Sternberg Press, 2012) p. 24.
25. L. Lek, "Geomancer" (2017), available at <https://vimeo.com/208910806/5e2e08b486>.