Autoencoding Blade Runner: Reconstructing Films With Artificial Neural Networks

Author Draft: To be presented at SIGGRAPH '17 Art Papers

ABSTRACT

'Blade Runner—Autoencoded' is 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 project explores the aesthetic qualities of the disembodied gaze of the neural network. The autoencoder is also capable of representing images from films it has not seen based on what it has learned from watching Blade Runner.

Introduction

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 synthesis a target image using a corpus of images. The image is synthesised in fragments that are matched from the database extracted from the corpus based on shape and colour similarity. Mital uses this technique to create a series of artworks called 'YouTube Smash Up' [4], synthesising the week's most popular video on YouTube from fragments of other trending videos on the platform. Another, somewhat related approach (and key influence to this project) is the research in reconstructing what people are watching while in an MRI scanner, solely from recorded brain scans [5].

SIGGRAPH '17 Art Gallery / Art Papers, July 30 - August 03, 2017, Los Angeles, CA, USA
ACM 978-1-4503-4998-7/17/07.
http://dx.doi.org/10.1145/3072940.3072964

This project was set out as a continuation in-kind of the previously described research, pursuing the same goal, while taking advantage of the recent advances in deep generative models (detailed in the next section). The film Blade Runner was chosen as the visual material for which to anchor this research, because of its relation to the themes perception, artificiality and artificial intelligence.

Technical Background

Research in deep learning, specifically in the field of computer vision, has been increasingly accelerating in recent years, particularly since Alex Krizhevsky's et al. [6] breakthrough in the 2011 ImageNet competition, where they solely used a single convolutional neural network to classify images into 1000 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 systems were seeing things in the way humans do. This skepticism was evidenced by the observation that such networks could easily be fooled into incorrectly classifying images which had been subtly manipulated using specific patterns of noise that were imperceptible to humans [7]. In response to such observations there was a drive in the research community towards developing generative models that were capable of generating realistic natural images. The reasoning being that if a network is capable of generating realistic natural images, it has a greater understanding—or at least we can be more confident it has—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 an 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 (reconstructs 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 [8,9] combines this network structure with a variational Bayesian approach to training, which makes strong assumptions concerning the distribution of latent variables (a Gaussian prior). This forces the autoencoder to use the latent space more efficiently, leading to more robust reconstructions and better generalisation.

Generative adversarial networks (GAN) [10] are an altogether different approaching 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 maximise the probability of correctly labeling the image as real or generated, the generator is trained to try and *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. The was significant as 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 small resolution. (They did this by replacing the traditional structure of convolutions alternating with pooling layers with strided convolutions, and fractionally-strided backwards convolutions for the generator network.)

In 2016, Larsen et al. [12] elegantly combined the GAN approach with a variational autoencoder (VAE). They use the strided convolution architecture popularised by Radford et al. and combined the training routines of the two approaches. They add 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 more simple a pixel-by-pixel basis. This significantly increases the generative capability of the VAE, optimising 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 that was implemented and used as the basis for this project.

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 dataset of images (usually of all the same subject matter i.e. handwritten digits [ref] or faces [ref]). Using these datasets restricts the complexity of what the model needs to represent and allows a direct comparison to be made between the visual fidelity of 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 the data that is 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 in the models generative capabilities.

Figure 1. Sample of a 64 frame mini-batch of reconstructed samples from the network trained on Blade Runner after 1 epoch at a resolution of 96x64.

Initially the model was only trained at a resolution of 96x64 (64x64 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 256x144 (coincidentally the smallest resolution allowed on YouTube). By increasing the size of the model, training was made a lot slower and more precarious, making it more likely that one of the three networks (that all have to learn in unison) would fail, resulting in a sharp degradation in the quality of images. Forcing the process to be started again from the beginning. It took approximately 3 days for the model to be trained on all the frames from the film once. (One complete cycle through the dataset is referred to as one epoch.)

After some trial and error, a set of hyperparameters were 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 1, 3 and 6 epochs. One novel technical

improvement made to this training procedure as part of this project was to reduce the amount of noise injected into the latent space over the course of training (by reducing the standard deviation of the Gaussian prior), in order for the model to better differentiate between frames that were similar (a more detailed, technical account of this training procedure can be found in the original technical report [13]).

Figure 2. Samples after training the model on frames from Blade Runner for 1 epoch (top row), 3 epochs (middle row) and 6 epochs (bottom row) at a resolution of 256x144.

Reconstructing Blade Runner, one frame at a time

After training, the autoencoder is then made the reinterpret each from from the film in order, then the reconstructed frames a 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 static scenes are represented so well, as it has, in effect, seen those scenes many more times than six times. In essence the model is simply overfitting to the training data (caused in most part 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 is an interesting outcome, especially 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 where there is a fixed background into a single representation, even if there is some movement in the scene (see Figure 4). This tendency was rectified somewhat by gradually reducing the noise injecting into the latent representations over the course of training, but not completely. Ultimately, this is a consequence of the images being so similar, they share nearly the same point in latent space, therefore cannot be differentiated by the generator network. Without some training procedure to enforce 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.

One curious outcome is the model's inability to represent completely black frames. When asked to recreate a black frame, it instead produces an image with a greenish haze (reminiscent of the phenomenon of seeing colours 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 lots of black images, but this was not done as it was deemed an interesting outcome.

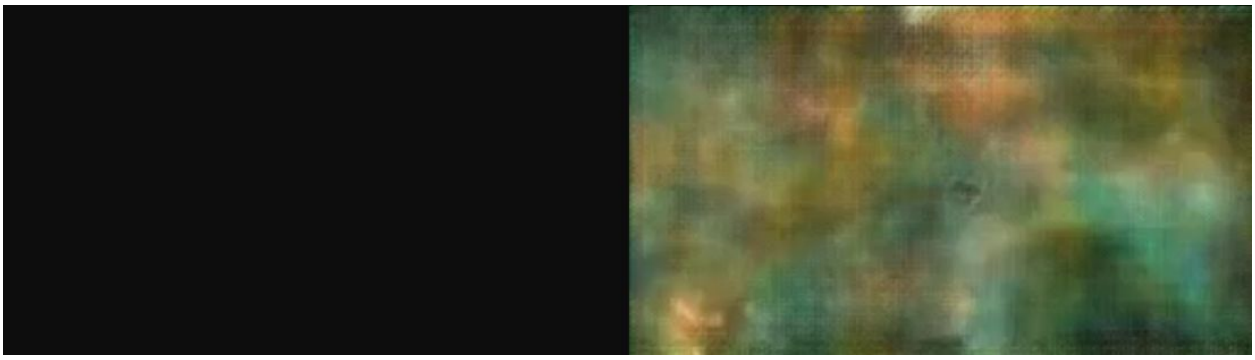

Figure 4. Left: A completely black image. Right: The model trained on Blade Runner interpretation of the completely black image.

Reconstructing other films with the Blade Runner model

Once trained, it is possible to get the autoencoder to process frames from any film. The model reinterprets any given set of images from what it has learned from Blade Runner, thus transferring the distinctive 'neo-noir' aesthetic onto any video presented to the model. Figure 7 shows the 1929 Documentary "Man with a Movie Camera" reinterpreted by the model. The film is black and white but the output from the model is in colour and is consistent with the visual style of Blade Runner.

Figure 5. Top row: Frames from the 1929 film "Man with a Movie Camera". Bottom row: Reinterpreted frames from the model that has been trained on Blade Runner. Images from

Dziga Vertov's "Man with a Movie Camera" are sourced from Wikimedia Commons and are in the public domain.

The reconstructions of other films are aesthetically interesting and unpredictable, but it is difficult to really make out what is being represented most of the time. Since this project was carried out, research has been published using a conditional adversarial encoder-decoder network to translate images from one domain into another [REF]. Providing a more formally defined and effective method to do this kind of image translation.

Why Blade Runner?

The film Blade Runner is adapted from Philip K. Dicks novel 'Do Androids Dream of Electric Sheep?' [14]. Set in a post-apocalyptic dystopian future, Rick Deckard 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. These 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 Voight-Kampff tests in order to distinguish them from humans. In the process he has to ask increasingly difficult moral questions about human and animal suffering with the intention of eliciting an empathic response in humans, but not androids. With the technological advances of the Nexus-6 replicants, it makes it increasingly difficult for Deckard to determine what is human and what is not, with Deckard himself having the growing suspicion that he himself may not be human.

By reinterpreting Blade Runner with an artificial neural networks memory of the film, 'Blade Runner—Autoencoded' seeks to emphasise 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 networks, with aspects of the flaws in it's visual reconstruction reminiscent of the deficiencies of our own, especially regarding memories of dreams.

There is a theory that Philip K. Dick structured the novel "Do Androids Dream of Electric Sheep?" around the working of the great French philosopher René Descartes, with Deckard acting out Descartes philosophical dilemmas [15]. (The name Rick Deckard bearing striking resemblance as an Americanised version of René Descartes.) Descartes emphasised that the senses (the primary source of knowledge) are often erroneous and prone to error. By examining this imperfect reconstruction of Blade Runner, the gaze of a disembodied machine, it becomes easier to acknowledge the flaws in our own internal representation of the world and easier to imagine the potential of other, substantially different systems that could have their own internal representations of the world.

Outcomes

The film 'Blade Runner—Autoencoded' and a report of the project were first published online in May 2016, gaining a great deal of attention on social media (with over 200,000 views on YouTube) and was subsequently written about in several online news articles (most notably by Aja Romano in Vox [A]). After the results were published online the autoencoder was trained for a further 20 epochs and used the create a second version of the film (see Figure 6) which was also upscaled into high resolution to make the work suitable to be viewed 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' that was trained an additional 20 times on the film.

This work was also both exhibited 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 that articulate the shifts that have taken place as technology has altered the way in which space and image are constructed and experienced. The exhibition engages with the fact that we are living in an environment more radically transformed by technology than at any other point in human history, and where cyberspace determines the contours of everything. [B] (*maybe change this last sentence*)

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.' In her opinion, the work 'belongs to the current moment in which artists are engaging with questions of where the boundary between AI and human perception lies.'[C] Iles relates the work to what Hito Steyerl describes as the 'disembodied, post-humanized gaze, outsourced to machines and other objects.' [D]

In the summer of 2017 the work will be included in the exhibition 'Into the Unknown: A Journey through Science Fiction' at The Barbican in London. A broad and encompassing survey of how Science Fiction is engaged by literature, film, music, video games and contemporary art. After being exhibited at The Barbican the show will subsequently go on an international tour.

Most recently, this technique of training on and reconstructing a film using an autoencoder was applied to the film 'Geomancer' (2017) [E], created in collaboration with the artist Lawrence Lek, who was commissioned to make the film for the Jerwood/FVU Awards 2017. 'Geomancer' tells the story of a weather satellite that becomes sentient and lands in Singapore on the eve of the city-state's centennial celebrations in the year 2065. In the film, the section processed by the autoencoder represents the internal mental representation of the AI protagonist during the films penultimate dream sequence.

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) 51–58.

4. P. K. Mital, YouTube Smash Up (2014), available at: http://pkmital.com/home/youtube-smash-up/

5. S. Nishimoto et al., "Reconstructing visual experiences from brain activity evoked by natural movies", Current Biology, 21.19 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) 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) 1278–1286.

10. I. Goodfellow, et al., "Generative Adversarial Nets.", in Advances in Neural Information Processing Systems (2014) 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) 1558–1566.

13. T. Broad and M. Grierson, “Autoencoding Video Frames”, Technical Report (London: Goldsmiths, 2016) available at: http://research.gold.ac.uk/19559/

14. P. K. Dick, “Do androids dream of electric sheep?”, (New York: Random House USA, 1982).

15. J. Brandt, “What defines human?”, http://www.br-insight.com/what-defines-human (2000).

A. 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

B C. Iles, “The Cyborg and the Sensorium,” *Dreamlands: Immersive Cinema and Art, 1905-2016.* (New Haven: Yale University Press, 2016) p. 121.

C. C. Iles, *personal communication* (2017).

D. H. Steyerl, “In Free Fall: A Thought Experiment on Vertical Perspective”, *The Wretched of the Screen* (Berlin: Sternberg Press, 2012) p. 24.

E. L. Lek, “Geomancer” (2017) available at: https://vimeo.com/208910806/5e2e08b486