Synthetic Video Generation

Why seeing should not always be believing!

Alex Adam
Image Tampering

Historically, manipulated imagery has deceived people

Off the shelf software (e.g. Photoshop) exists to do this now

Has become standard in tabloids/social media

Public have become somewhat numb to it - it’s no longer as impactful/shocking
How does machine learning fit in?

Advent of machine learning has made image manipulation even easier

Video manipulation is now also tractable with enough data and compute

Can make good synthetic videos using a gaming computer in a bedroom

Public are largely unaware of this and the danger it poses!
for the Home Department should be in this chamber
Part I: Faceswap
In 2017, reddit (/u/deepfakes) posted Python code that uses machine learning to swap faces in images/video

‘Deepfake’ content flooded reddit, YouTube and adult websites

Reddit since banned this content (but variants of the code are open source https://github.com/deepfakes/faceswap)

Autoencoder concepts underlie most ‘Deepfake’ methods
Faceswap Algorithm

Image 1

Feature maps 1

Decoder 1

Result image 1

Encoder

Feature maps 2

Decoder 2

Result image 2

Image source https://medium.com/@jonathan_hui/how-deep-learning-fakes-videos-deepfakes-and-how-to-detect-it-c0b50f7cb9
Inference

Image source https://medium.com/@jonathan_hui/how-deep-learning-fakes-videos-deepfakes-and-how-to-detect-it-c0b50fbf7cb9
• Faceswap model is an autoencoder.

• **Encoder**: This converts an input image into a compressed representation.

• **Decoder**: This reconstructs the original input image from this representation.

• Two autoencoders are trained. One for each face. Separate decoders, but a shared encoder.

• During training alternate one gradient descent step of each.
Faceswap: How do you do it?

- Split source and target videos into frames

- **Detect** faces in the source/target frames (filtering out unwanted faces)

- Train faceswap autoencoders on the extracted faces

- Used trained autoencoder to swap target input face - paste back into target frame
Faceswap: Example
• The encoder/decoder paradigm in the original faceswap underlies more advanced techniques

• Can swap more than the face too (e.g. the whole head).

• Can also do facial-reenactment (e.g. re-sync a video of one individual) to match a source video

• [https://github.com/deepfakes/faceswap](https://github.com/deepfakes/faceswap) architecture is tuned for swapping the face region

• More challenging to gather training data for e.g. the whole head
Part II: Generative Adversarial Networks (GANs)
Early 2018 - Shortly after the pure CNN faceswap implementations, GAN implementations started to appear.

https://github.com/shaoanlu/faceswap-GAN

Has potential to generate much more realistic faces (due to adversarial loss)

Inherently fools attempts to classify generated content
Generative Adversarial Networks

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio

(Submitted on 10 Jun 2014)

We propose a new framework for estimating generative models via an adversarial process, in which we simultaneously train two models: a generative model $G$ that captures the data distribution, and a discriminative model $D$ that estimates the probability that a sample came from the training data rather than $G$. The training procedure for $G$ is to maximize the probability of $D$ making a mistake. This framework corresponds to a minimax two-player game. In the space of arbitrary functions $G$ and $D$, a unique solution exists, with $G$ recovering the training data distribution and $D$ equal to 1/2 everywhere. In the case where $G$ and $D$ are defined by multilayer perceptrons, the entire system can be trained with backpropagation. There is no need for any Markov chains or unrolled approximate inference networks during either training or generation of samples. Experiments demonstrate the potential of the framework through qualitative and quantitative evaluation of the generated samples.

Subjects:  Machine Learning (stat.ML); Machine Learning (cs.LG)
Cite as:  arXiv:1406.2661 [stat.ML]
(or arXiv:1406.2661v1 [stat.ML] for this version)

Submission history
From: Ian Goodfellow [view email]
[v1] Tue, 10 Jun 2014 18:58:17 UTC (1,257 KB)
The ‘Adversarial’ in GANs

- Conventionally in machine learning optimise by minimising difference between model predictions and true values.

- Examples are minimising ‘mean squared error’ or ‘reconstruction loss’

- Suppose your model generates data (e.g. faces). Turns out you can improve the quality of the model by tuning it with a discriminator

- The discriminator is trained to spot real data from generated data
GAN Visualisation

The generator and discriminator compete with each other until an equilibrium is reached.

GAN Generation Timeline

None of these faces are ‘real’. All generated by GANs

[Links to papers]

arxiv.org/abs/1406.2661
arxiv.org/abs/1511.06434
arxiv.org/abs/1606.07536
arxiv.org/abs/1710.10196
arxiv.org/abs/1812.04948
Part III: Style Transfer (Cycle-GAN)
- 2017 - Researchers demonstrate image-to image translation between unpaired sets of images using Cycle consistent GANs (Cycle-GAN).

- 2018 - See applications of these methods in deepfake generation

- 2018- Variant of Cycle-GAN developed to improve ‘temporal consistency’ (Recycle-GAN)

- Can translate any ‘style’ from input to target. Unsupervised video retargeting
Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks

Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros

(Submitted on 30 Mar 2017 (v1), last revised 15 Nov 2018 (this version, v6))

Image-to-image translation is a class of vision and graphics problems where the goal is to learn the mapping between an input image and an output image using a training set of aligned image pairs. However, for many tasks, paired training data will not be available. We present an approach for learning to translate an image from a source domain $X$ to a target domain $Y$ in the absence of paired examples. Our goal is to learn a mapping $G : X \rightarrow Y$ such that the distribution of images from $G(X)$ is indistinguishable from the distribution $Y$ using an adversarial loss. Because this mapping is highly under-constrained, we couple it with an inverse mapping $F : Y \rightarrow X$ and introduce a cycle consistency loss to push $F(G(X)) \approx X$ (and vice versa). Qualitative results are presented on several tasks where paired training data does not exist, including collection style transfer, object transfiguration, season transfer, photo enhancement, etc. Quantitative comparisons against several prior methods demonstrate the superiority of our approach.

Comments: An extended version of our ICCV 2017 paper, v6 updated the implementation details in the appendix. Code and data: this https URL

Subjects: Computer Vision and Pattern Recognition (cs.CV)

Cite as: arXiv:1703.10593 [cs.CV]

(or arXiv:1703.10593v6 [cs.CV] for this version)

Bibliographic data
[Enable Bibex (What is Bibex?)]
Cycle Consistent GANs

- Paired data is a luxury. Suppose we have unpaired data.

- How do you generate e.g. a ‘Van Gogh version’ of your particular image?

- Clearly paired training data for this is unlikely to exist.

- Unclear what map to learn - need to define one.

Image source: https://hardikbansal.github.io/CycleGANBlog/
Cycle-GAN

Real Image in domain A → $G_{AB}$ → Fake Image in domain B → $G_{BA}$ → Reconstructed Image

$L2$ Loss

$D_B$ → real or fake?

Discriminator for domain B

Real Image in domain B

$G_{BA}$ generates a reconstructed image of domain A.
This makes the shape to be maintained when $G_{AB}$ generates a horse image from the zebra.

Unsupervised Image to Image Translation

Image source: https://junyanz.github.io/CycleGAN/
Youtube creative commons: https://www.youtube.com/watch?v=9reHvktowLY
Recycle-GAN: Unsupervised Video Retargeting

Aayush Bansal, Shugao Ma, Deva Ramanan, Yaser Sheikh

(Submitted on 15 Aug 2018)

We introduce a data-driven approach for unsupervised video retargeting that translates content from one domain to another while preserving the style native to a domain, i.e., if contents of John Oliver's speech were to be transferred to Stephen Colbert, then the generated content/speech should be in Stephen Colbert's style. Our approach combines both spatial and temporal information along with adversarial losses for content translation and style preservation. In this work, we first study the advantages of using spatiotemporal constraints over spatial constraints for effective retargeting. We then demonstrate the proposed approach for the problems where information in both space and time matters such as face-to-face translation, flower-to-flower, wind and cloud synthesis, sunrise and sunset.

Comments: ECCV 2018; Please refer to project webpage for videos – this http URL
Subjects: Computer Vision and Pattern Recognition (cs.CV); Graphics (cs.GR); Machine Learning (cs.LG)
Cite as: arXiv:1808.05174 [cs.CV]
(or arXiv:1808.05174v1 [cs.CV] for this version)

Bibliographic data
[Enable Bibex (What is Bibex?)]

Submission history
From: Aayush Bansal [view email]
Recycle-GAN: Example

Can smooth ‘standard’ Cycle-GAN to get better smoothness in time:

Image source [https://www.cs.cmu.edu/~aayushb/Recycle-GAN/](https://www.cs.cmu.edu/~aayushb/Recycle-GAN/)
Youtube creative commons [https://www.youtube.com/watch?v=8OqxXt_Y2Ik](https://www.youtube.com/watch?v=8OqxXt_Y2Ik)
Part IV: Advanced Topics (Face Reconstruction, Reenactment)
● Academic research in face manipulation techniques predates the recent interest in ‘deepfake’

● Outline the conceptual ideas underlying face reconstruction & reenactment
- **Monocular Face Reconstruction** - Aim to reconstruct 3D face models from visual data. Optimisation methods fit a 3D template model of inner face. Deep nets can reconstruct the 3D shape.

- **(Video-Based) Facial Reenactment** - Rewrite the face of a target actor by transferring facial expressions from source actor.

- **Visual Dubbing** - Special case of facial reenactment where mouth motion is altered to match new audio track.

- **Image Translation** - Covered this in Section III.
Parametric Face Representation

- Represent faces via a parametric model

- Parameters encode
  - Rigid head pose (rotation, translation)
  - Facial identity (geometry, reflectance)
  - Facial expression
  - (Eye gaze)
  - (Illumination)

- Given an image - can calculate these parameters

*Neural Nets can reconstruct 3D face from still images (monocular facial reconstruction)*
Deep Video Portraits

Siggraph 2018

H. Kim ¹ P. Garrido ¹ A. Tewari ¹ W. Xu ¹ J. Thies ² M. Nießner ² P. Perez ³ C. Richardt ⁴ M. Zollhöfer ⁵ C. Theobalt ¹

¹ MPI Informatics ² Technical University of Munich ³ Technicolor ⁴ University of Bath ⁵ Stanford University

Abstract:

We present a novel approach that enables photo-realistic re-animation of portrait videos using only an input video. In contrast to existing approaches that are restricted to manipulations of facial expressions only, we are the first to transfer the full 3D head position, head rotation, face expression, eye gaze, and eye blinking from a source actor to a portrait video of a target actor. The core of our approach is a generative neural network with a novel space-time architecture. The network takes as input synthetic renderings of a parametric face model, based on which it predicts photo-realistic video frames for a given target actor. The realism in this rendering-to-video transfer is achieved by careful adversarial training. and as a result, we can create modified target videos that
Facial Re-enactment (Transferring the face)

- Extract parameters for source and target frames

- Calculate ‘modified parameters’ that correspond to the source facial pose (head, expression) in the target identity

- Render a face for these modified parameters

- Train a neural network that maps this rendering to video

- *(Aside: Condition on past frames for smoothness. Also render correspondence and eye gaze map in addition the colour rendering. All are input for the NN)*
Overall Model Architecture

Discriminator

Adversarial Training (see Section II)
Deep Video Portraits: Example

- Allows almost complete control of target video portrait!
- Don’t need much training data (~1 minute of video footage)
- Can also transfer e.g. just expressions

Youtube creative commons [https://www.youtube.com/watch?v=3OdYTfDkTAc](https://www.youtube.com/watch?v=3OdYTfDkTAc)
Related Work

- **Face2Face** (Thies et al 2016)
  - Non deep learning
  - Similar reenactment quality
  - Restricted to expression transfer

- **Synthesising Obama - leaning lip-sync from audio** (Suwajanakorn et al 2017)
  - Use RNN to synthesise mouth texture from audio stream
  - Could generate a mouth for a fake audio stream
Conclusions & Questions
● Machine learning will change the way we think about video and ‘truth’

● Many amazing applications of synthetic video (special effects/dubbing)

● Need to also be aware of the dangers

● Outlined key principles underlying faceswap, GANs, style transfer and face reenactment
Thank you

If you’re interested in finding out more about Faculty and our research, get in touch. We are hiring!

https://faculty.ai/

Upcoming events

1st May: Making AI safe and fair today
22nd May: Why we need a better conversation around automation and care

Register at: www.faculty.ai/events
Rate today’s session