Paper Summaries

Peter Robertson

1 GANs Based Methods

1.1 Generative Adversarial Network

The original GANs paper proposed by Ian Goodfellow [1]. This paper uses two feed forward neural networks (as proof of concept) as generator and discriminator. The original loss function is used, although it is mentioned that even at this stage issues with it are highlighted. Argument made to say that while learning a probability distribution is great, the goal is to produce samples from it, so generating sample from an implicit distribution is as useful. No need to go into extreme depth as this was covered in the Deep Learning lectures.

1.2 Conditional Generative Adversarial Nets

Conditional GANs [2] extend from Goodfellow's work [1] by concatenating an input to the input noise given the generator. This is performed on the MNIST Dataset [3] by conditioning the generator on a given digit to produce.

1.3 DCGAN

This paper [4] takes the concept of adversarial training of neural networks and applies it with convolutional neural networks on the CIFAR10 dataset [5]. The paper also attempts to make some progress in stabilizing the training of GANs models with the use of Batch Normalization before the input to each subsequent layer. Experimentally this is shown to assist training by helping gradients to flow through the network and poor weight initialization.

1.4 Wasserstein GANs

In an attempt to stabilize the training of GANs models, the Wasserstein or 'Earth-Mover' loss function is proposed [6]. By using the Wasserstein loss function, the generator and discriminator no longer have to alternate training epochs to try to maintain model stability. By using the Wasserstein loss function, the discriminator (critic as it is called in this paper) is trained to an optimum point, at which time the loss provided to the generator is at a maximum, making training the generator easier. The generator can then be trained to an optimum position, at which time, training the critic can continue again. This method has greatly improved stability and is fairly robust to a large range of model architectures which have been tested. The generator and critic also no longer have to have a similar capacity.

To achieve this, the Lipschtiz condition must be met. This is achieved by clipping the weights on the model. The paper points out that this method of achieving the Lipschtiz condition is non-ideal and is a point of further work.

Other points of note, the loss plot produced by the WGAN model seems to correlate with perceived image quality, which could not be said for precious loss functions.

1.5 Improved Training of Wasserstein GANs

To improve the Lipschitz constraint condition to the WGANs [6], an alternate method is proposed by using gradient penalty [7]. Paper examines the distribution of weights with WGAN and found that weights are bunched around clipping limits. With WGAN+GP the weights form a normal distribution, allowing more complex functions to be captured as clipped weights for a bias to simple functions, limiting the model capacity. No batch normalization is used, as this is no longer valid when using gradient penalty. This loss function seems to be able to train any architecture, even ResNet architectures which had currently been unsuccessful when used in GANs models. Training speed is also improved compared with WGAN.

1.6 Progressive Growing of GANs for Improved Quality, Stability, and Variation

Using the WGAN+GP loss function from [7], Nvidia propose a novel training architecture to produce large high quality images [8]. Starting with a very small network which the generator and discriminator can produce and criticize 4x4 images respectively. The model is trained until the discriminator is successfully fooled by the generated images. Then, an additional layer is 'faded' into the networks, allowing 8x8 images to processed. The layer is faded in with a residual connection to the previous layer, slowly weighting the new layer more heavily until the residual connection weighted to zero. This avoids sudden changed in the model architecture.

The two networks are mirror images of each other. Transpose convolution is not used, instead the image in up-sampled with 2x2 element duplication, while down-sampling is performed with average pooling.

While the project at hand is not related to generating images but 3D models, the concept of working from a low dimensionality upwards, adding complexity seems intuitive. In regards to the generation of faces, 4x4 and 8x8 images force the model to learn simple facial structures such as face shape which will be consistent for all samples, while higher resolution images allow the model to add finer degrees of detail to the images, but learning lower level concepts is forced initially. This principle isn't dissimilar to the concept of meta-learning used in [9].

1.7 End-to-End Speech-Driven Facial Animation with Temporal GANs

This paper [10] attempts to generate video from audio, given an image of the subject. The generator is given an image of the subject's head and a short audio clip to generate video for. The model uses two discriminators, one examining individual video frames to see if they are realistic images of the subject's face, the second examines the sequence of frames produced to see if they form a realistic video sequence. The model produces

reasonably impressive results, even more so considering that it is using the original GANs loss function from [1] which has been shown to be unstable, difficult to train and has been greatly improved upon though various other methods.

1.8 Self-Attention Generative Adversarial Networks

The issue of convolutional GANs models [4] failing to produce structural pattens in an image is addressed in [11] by incorporating the attention mechanism into a GANs model. The paper notes that as convolutional networks require large kernels or very deep models to have a large receptive field to capture long range dependencies in an image, models can be improved with attention. Common examples of this issue is generating images of animals which have an incorrect number of limbs, but have realistic fur texture as this is a local dependency. The attention mechanism allows long range dependencies in the image to be modelled by convolutional models more easily. This paper is also co-authored by Ian Goodfellow, the original creator of the GANs architecture.

1.9 Few-Shot Adversarial Learning of Realistic Neural Talking Head Models

This paper uses meta-learning on conjunction with adversarial learning methods [9]. The paper aims to learn an encoding which represents a person's face and then use this to generate video of the subject moving into an unseen position. The model initially learns an encoding based on several frames from an input video from random positions in the video. Once the subject encoding has been learnt, the model continues to learn to construct frames of video of the subject using facial landmarks to drive the positioning. Training is performed with existing positions in the video which the GAN model is tested on if the presented frame is real or generated. After training, the model is capable to producing subjects in positions unseen in the video.

In regards to 3D modelling, a similar method could be used to encode a subject in a very similar manner to what has been performed here. With this subject encoding, a generic neutral mesh template could be replaced by a neutral subject specific mesh template, reducing the work of the generator to just having to capture realistic movement as opposed to movement and reproducing the subject.

2 Other Generative Methods

2.1 Capture, Learning, and Synthesis of 3D Speaking Styles

The aim of [12] is to synthesise video sequences of 3D models speaking given an audio input. A main goal of this paper is to be independent to the speaker. Dataset (VOCASET) collected of 3D head scans of subjects while speaking a total of 480 sequences. Each sequence last 3-4 seconds for a total dataset of 29 minutes. Models are recorded at 60fps, resulting in 100,000 frames in the dataset. A common face template mesh is aligned to all the scans. The model uses the FLAME head model [13] as a template facial model.

Rather than including layers in the model to extract information on the input audio, the model uses the pre-trained Mozilla implementation of the DeepSpeech paper [14], with a few minor improvements. The DeepSpeech model takes an input WAV file and returns un-normalized log probabilities for each letter and some additional characters for

time slices in the audio file. The model outperforms all commercially available speech recognition methods at the time it was published (2014).

Initially, the DeepSpeech model extracts the audio features at 50fps, this is up-sampled to 60fps. Then, an encoder model made up for four convolutional layers and two fully connected layers encode the audio features conditioned on the speaker so that the speaking styles of individual speakers can be learnt and later this latent space can be explored. Finally, a decoder made up of a fully connected layer with a linear activation function is used to output the displacements of the 5023 vertices in the FLAME face model.

Personal experimentation with the completed model has shown mixed results.

2.2 Audio-driven facial animation by joint end-to-end learning of pose and emotion

Nvidia facial animation from speech [15] Attempts to generate 3D facial animation from audio. Output vertices difference from neutral face mesh. Specific to actor, trained on 3-5 minutes of footage for that actor. Model extracts temporal audio features with convolutional net. Then convolutional net to analyse temporal components of speech. Articulation network also accepts emotional state input to aim ambiguities between different facial shape from same audio.

2.3 End-to-end Learning for 3D Facial Animation from Raw Waveforms of Speech

[16]

2.4 Synthesising 3D Facial Motion from "In-the-Wild" Speech

Model is independent of speaker and capture rig. Create dataset by recording talking 4D faces speaking words from LRW dataset. Register 3D meshes and learn 3D blendshapes. Use DCAW to align speech with meshes to give ground truth. Train a model to predict 3D facial motion from speech under uncontrolled conditions. Use audio from LRW dataset to train with. Imperial Paper [17]

3 3D Modelling

3.1 Direct-Manipulation of Blendshapes

This paper [18] provides a reasonable overview of what blendshapes are. It's a slightly older paper from 2010, but the techniques in the field of animation haven't changed. Animating one second of video may take a skilled animator an hour using blendshape models. Blendshapes are linear weighted sum of blendshape targets. These targets can be facial expressions, or approximations of facial muscles. A blendshape can have 100+ degrees of freedom, while facial meshes have many many more. Limiting the degrees of freedom to a human animator to 100 is still huge, but this may oversimplify the problem to a computer. Facial mesh may have 5000-10000 vertices, perhaps using PCA to reduce this down is smart. Or maybe using a technique similar to Nvidia's Progressively growing GAN [8] may allow the DOF to be slowly increased.

3.2 Learning a model of facial shape and expression from 4D scans

This paper creates the FLAME head model which has been designed for animation [13]. With the aim of being computationally efficient and compatible with game engines, this model only contains 5023 vertices. Aim is to be a good middle ground between current simple models which are unrealistic and high end models which are very time consuming. Model uses very large dataset made up of multiple datasets. Current comparison is FaceWarehouse Does some work to address the issue of registering 4D scans.

3.3 A Survey on Shape Correspondence

The problem of finding the correspondence between two shapes is well discussed in [19]. Correspondence is also referred to as registration, alignment and matching. The paper outlines the main differences between types of correspondence problems such as full and partial correspondence, dense and sparse correspondence.

3.4 Modelling and Correspondence of Topologically Complex 3D Shapes

[20]

4 Lip Reading Papers

4.1 Lip Reading in the Wild

This paper first creates a new extended dataset, later named Lip Reading Words (LRW) which allowed for a CNN architecture based on the VGG-M architecture to be trained to lip read video [21]. The model is able to greatly outperform the state of the art models of the time. LRW contains over 1000 hours of spoken text with over 1000 speakers, a corpus of over 1000 words and over a million word utterances.

4.2 LipNet: Sentence-level Lipreading

Collaboration between Oxford and Google DeepMind [22] on the GRID dataset. Model makes use of Spatiotemporal convolutions which convolve across multiple video frames. Feature maps are passed to two GRU layers and followed by a linear softmax layer for character distributions. CTC loss is used on output layer. The model is the first sentence level model, previous models have been word level. Able to take variable length sequences.

4.3 Lip Reading Sentences in the wild

[23] builds on previous work to construct a model which can work on sentence level rather than just word level. Paper describes process of constructing a new dataset; Lip Reading Sentences (LRS). Model is capable of multiple inputs and is trained on both audio and video to recognise characters. To prevent the model from being reliant on one of the two inputs (namely the audio), the inputs are systematically distorted or removed. Video is passed through a CNN architecture, followed by LSTM layers. Audio is processed as

MFCC then input to LSTM layers. The two are combined with then attention mechanism and further LSTM layers and an output fully connected layer with softmax activation for character distributions. As LSTMs are challenging to train on long sequences, curriculum learning is used. Initially the model is trained on short single word sequences, as training continues, the length of the input sequences increases. This accelerates training and reduces model overfitting.

4.4 Deep Lip Reading: a comparison of models and an online application

[24] adapts various models which have had success for translation and automatic speech recognition (ASR) problems.

4.5 Large-Scale Visual Speech Recognition

As opposed to character level predictions, [25] predicts phonemes, combined with a language model to predict words. The paper describes a new dataset (LSVSR) which is larger than all previous datasets. Due to the size of the datasets and the model, 64 GPUs are used in parallel to train with a batch size of 64. The model uses five spatiotemporal convolution layers, followed by three bidirectional LSTM layers and a fully connected layer to predict phonemes. This model again achieves the state of the art lip reading results.

5 Speech Recognition

5.1 Deep Speech: Scaling up end-to-end speech recognition

The DeepSpeech model [14] at the time of publishing greatly outperformed all commercially available speech recognition models. The model uses no hand crafted components and is very robust to background noise due to the augmented training data. The model from the original paper was constructed and trained with efficiency in mind due to the belief that being able to train a simple but effective model for an extended period of time would yield greater results than that of a more complex model with a reduced training time. The model takes speech spectrograms as input and splits these into slices. The first three layers are fully connected, followed by a single bi-directional recurrent layer. The fifth layer is an additional fully connected layer, while the output layer is a softmax function.

This simple model was extended by Mozilla in their publicly available implementation which made use of LSTM layer as the recurrent layer as well as using Mel-frequency cepstral coefficients (MFCC) as opposed to spectrograms.

The training data was extended with data augmentation by adding noise to the audio. For 1000 hours of audio, 1000 hours of noise is required.

6 Attention Based Mechanism

6.1 Attention is All you Need

Google Attention paper [26]

6.2 Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

Attention applied to images [27]

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[2] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.

[3] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

[4] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[5] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[6] Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *CoRR*, abs/1701.07875, 2017.

[7] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. *CoRR*, abs/1704.00028, 2017.

[8] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017.

[9] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. *arXiv preprint arXiv:1905.08233*, 2019.

[10] K. Vougioukas, S. Petridis, and M. Pantic. End-to-End Speech-Driven Facial Animation with Temporal GANs. *arXiv e-prints*, May 2018.

[11] Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *CoRR*, abs/1805.08318, 2018.

[12] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael Black. Capture, learning, and synthesis of 3D speaking styles. *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[13] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017.

[14] Awni Y. Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. Deep speech: Scaling up end-to-end speech recognition. *CoRR*, abs/1412.5567, 2014.

[15] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Trans. Graph.*, 36(4):94:1–94:12, 2017.

[16] Hai Xuan Pham, Yuting Wang, and Vladimir Pavlovic. End-to-end learning for 3d facial animation from raw waveforms of speech. *CoRR*, abs/1710.00920, 2017.

[17] Panagiotis Tzirakis, Athanasios Papaioannou, Alexander Lattas, Michail Tarasiou, Björn W. Schuller, and Stefanos Zafeiriou. Synthesising 3d facial motion from "in-the-wild" speech. *CoRR*, abs/1904.07002, 2019.

[18] J. P. Lewis and K. Anjyo. Direct manipulation blendshapes. *IEEE Computer Graphics and Applications*, 30(4):42–50, July 2010.

[19] Oliver van Kaick, Hao Zhang, Ghassan Hamarneh, and Daniel Cohen-Or. A survey on shape correspondence. *Computer Graphics Forum*, 30(6):1681–1707, 2011.

[20] Ibraheem Alhashim. Modelling and correspondence of topologically complex 3d shapes. *CoRR*, abs/1506.06855, 2015.

[21] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In Shang-Hong Lai, Vincent Lepetit, Ko Nishino, and Yoichi Sato, editors, *Computer Vision - ACCV 2016 - 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II*, volume 10112 of *Lecture Notes in Computer Science*, pages 87–103, 2016.

[22] Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, and Nando de Freitas. Lipnet: Sentence-level lipreading. *CoRR*, abs/1611.01599, 2016.

[23] Joon Son Chung, Andrew W. Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3444–3453. IEEE Computer Society, 2017.

[24] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Deep lip reading: a comparison of models and an online application. *CoRR*, abs/1806.06053, 2018.

[25] Brendan Shillingford, Yannis M. Assael, Matthew W. Hoffman, Thomas Paine, Cían Hughes, Utsav Prabhu, Hank Liao, Hasim Sak, Kanishka Rao, Lorrayne Bennett, Marie Mulville, Ben Coppin, Ben Laurie, Andrew W. Senior, and Nando de Freitas. Large-scale visual speech recognition. *CoRR*, abs/1807.05162, 2018.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6000–6010, 2017.

[27] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2048–2057. JMLR.org, 2015.