Challenges Associated with Generative Forms of Multimedia Content (Keynote Talk)

Alan F. Smeaton
Insight Centre for Data Analytics
Dublin City University
Glasnevin, Dublin 9, Ireland
Email: Alan.Smeaton@dcu.ie

***Abstract*—This short paper presents what is currently the main challenge associated with using Generative Adversarial Networks (GANs) to generate visual media. That challenge is around automatically determining the authenticity of an image/video, which is of increasing importance as fake videos and images start to proliferate on social media, especially when associated with political campaigning. The paper introduces GANs, outlines how they are used to generate visual media and summarises this major challenge.**

***Index Terms*—Generative multimedia, GANs,**

I. Introduction

Since the very first time in the mid-1980s when images, audio and then video became available in digital format to a large consumer base, we have always sought to develop computational techniques to automatically analyse such media, in order to determine its content. In the case of images and video, the earliest approaches were to represent content in terms of low level features like colour and texture, and these features were then encoded into a standardised MPEG-7 format [1]. This was quite restricted as the semantic content of colour or texture-based representations is very limited.

Throughout the 1990s and 2000s we made slow and incremental progress in media analysis using machine learning, in particular using Support Vector Machines (SVMs) [2] which were particularly popular. Computer vision as a challenge and an application domain pushed the development of machine learning and this drove the incremental progress as we saw in challenge benchmarks like Imagenet [3]. We also saw the developments in machine learning which were initially in computer vision, trickle into other application areas.

Then, in 2012, we had a needle shift in the technology with the introduction of convolutional neural networks in the Imagenet challenge [3] which saw a huge leap in performance [4]. Since then, within the last 5 years, we have seen extreme progress in the accuracy and reliability of automatic content analysis of visual media as well as a huge adoption of machine learning in areas like transport, finance, education, personal health and wellness, and others. We are now at the point where the images and video clips we upload and share on social media are automatically analysed and tagged for content descriptors thus making them more accessible and discoverable. This shows that the machine learning techniques used are scalable and robust, having been built on progress in the development of machine learning.

Yet while we can collectively bask in this achievement as representing progress in terms of science or perhaps it is progress just in engineering, we are now faced with an even greater challenge and that is how to cope with a scenario where media, in particular visual media, is automatically generated.

This paper is a summary of a keynote paper presented at the Content Based Multimedia Conference (CBMI) in September 2019. In the next section we

II. Machine Learning and GANs

Current machine learning has attracted lots of attention from researchers across the spectrum and there is a strong focus on deep learning which is attractive because of its sometimes accurate performance in prediction and classification tasks, especially in computer vision. The processing this involves replicates some of the neural processing that is carried out in the human brain. Yet machine learning has several issues and challenges that machine learning applications and researchers wrestle with. These include the following:

- model building, taking a real world problem with real world data and choosing and developing a computational model for it, adjusting hyper-parameters and fine tuning an implementation so it provides some function;
- new architectures for machine learning like capsule networks replicating the (human) brains neural structures [5] and exploring configurations beyond simple connections, like how neurotransmitters in the brain work;
- energy costs for computation, moving from general purpose energy-sapping GPUs to custom-built hardware, driven by [6];

AS is part-funded by Science Foundation Ireland under grant number SFI/12/RC/2289 (Insight Centre)

978-1-7281-4673-7/19/$31.00 ©2019 IEEE

- explainable AI which is about opening up decision-making by systems so that we can understand, and ultimately build trust in systems which use machine learning [7];
- ethical AI, using AI for good and being aware of and taking cognisance of bias in data, all brought together and highlighted in the *Ethics Guidelines for Trustworthy Artificial Intelligence* a document prepared by the European Union's High-Level Expert Group on Artificial Intelligence, set up by the European Commission in June 2018[1]

Since developments in AI and machine learning kicked into a higher gear in about 2012 or 2013, developments in the last 5 years have gone in several directions, which has been summarised in Figure 1.

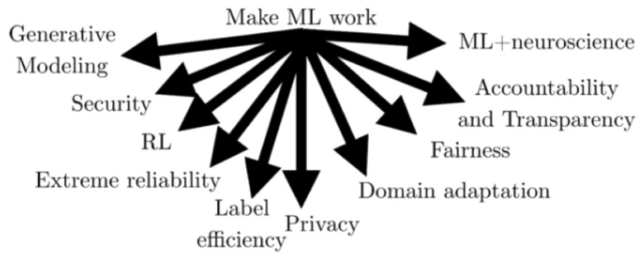

Fig. 1. Five years of Machine Learning development (image credit to Ian Goodfellow)

One of the most interesting of these is the development of generative models, which are a new flavour of machine learning which can generate new, virtual content from a set of training data, and here is where the challenges are for multimedia content.

III. GANs to Generate Visual Media

The *automatic* generation or automatic editing of media content is not a new concept. Examples vary from automatic generation of video summaries [8] to automatic "beautification of facial images [9].

Generative Adversarial Networks are a recent development [10] in which machine learning is used to generate rather than classify or predict. It learns probability distributions from some collection of training data and then generates or synthesises new example data from those distributions. Images are a great natural example for GANs, illustrated in Figure 2. On the left side of the graphic we have a collection of images, of cats for example, and we can use a neural network to learn or encode the characteristics and build a model of what a cat should look like. Then we use a second neural network to generate an instance of a cat, derived from the model of the training data. We test the conformance of the generated instance, and depending on how close it is to the distribution, we refine it and iteratively move towards a generated instance which passes the test of conformance to the model.

[1]https://ec.europa.eu/futurium/en/ai-alliance-consultation

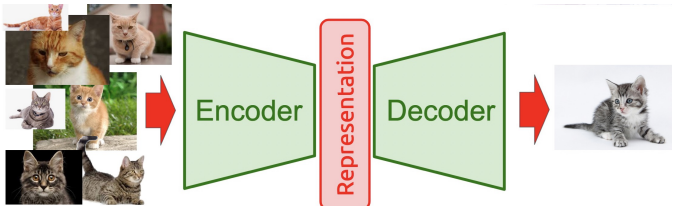

Fig. 2. Operation of GANs

Generating facial images of celebrities were the first large scale application and are a great natural example for GAN generation since the faces are usually front-facing and there are many of them. After an initial focus on celebrity faces, using GANs to generate images moved on to other targets, including the 1,000 object classes in the ImageNet dataset [11] with rapid improvements in image quality. In 2018, a GAN-generated *painting* trained on existing impressionist paintings and called "Edmond de Belamy, sold for $435,000 at auction in Christies.

Within the last couple of years, GANs have been used for generating realistic examples of paintings, images videos, text, 3D models (for replacement teeth !), DNA sequences and each of these is an honourable and useful application. the popularity of using GANs in research has increased dramatically and in 2018 a survey paper identified and summarised the contribution of 322 other papers on GANs [12].

One of the most eye-catching, and potentially threatening use of GAN is to generate fake videos of people, especially when they are used in political campaigns on social media. The CTRL-SHIFT-FACE channel on YouTube is currently probably the best (public) example of how powerful and accurate deepfake videos can be though these are all watermarked as being deepfakes and thus this is a responsible use of the technology.

IV. The Challenges with Generated Visual Media

Because automatically generated media is "fake", created to emulate genuine media to great accuracy, it naturally brings into question how to determine whether it is authentic or not.

Media is now so important not just as broadcasting for our entertainment but for informing and influencing citizens [13], especially on social media. One approach to detecting its authenticity is to create signatures for individuals who are the subject of such deepfakes, using large volumes of video footage and to learn a model for each individual, modeling their favoured body movements, their personal twitches, mannerisms, the way and angle they hold their face, facial reactions, how they use their hands when speaking, if at all, and more. In this way we see that signatures for individuals can be build up and then candidate videos compared against these signatures. The problem with this approach is that deepfakes themselves model the characteristics of individuals drawn from models of those individuals which are created from large amounts of visual data of those individuals. At present these models are build around faces, facial characteristics and

reactions, but the next step is to go beyond just the face and to incorporate body movements, at which point the signature approach fails.

Another approach is to replicate our own natural insights as to what might be fake or real, by replicating human neural processing when we see fake videos. In a way this is digitising what is known as the "uncanny valley", our ability to use our perception to determine a mis-match between what is real and what is fake. In one approach which focus on detecting faces, this is shown to have promise [14] but as with the subjects' signatures approach as generated videos get more realistic in nature, our own perception may not be able to determine the difference between real and fake.

In summarising the challenges that generated visual media presents the following points can be made

- Generated visual media is here, now, and it is relatively easy to create using off-the-shelf hardware including GPUs, and open source software. Some deepfake support systems even include starter kits to make generating your first fake video, easy.
- Generated visual media can be a good thing, like generating interactive video in chatbot applications for mental health well-being.
- As with all new technology, there can and will be darker sides which exploit it for wicked or criminal purposes and we are seeing this in social media already.
- Automatically determining the difference between fake and real visual media will probably end up as a "cold war ... as forensic video scientists discover techniques to detect fake videos, deepfakers will counter with ways to avoid detection, and the never-ending circle will go around again.

REFERENCES

[1] T. Sikora, "The MPEG-7 visual standard for content description-an overview," *IEEE Transactions on circuits and systems for video technology*, vol. 11, no. 6, pp. 696–702, 2001.

[2] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," in *Proceedings of the ninth ACM international conference on Multimedia*. ACM, 2001, pp. 107–118.

[3] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[5] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Advances in neural information processing systems*, 2017, pp. 3856–3866.

[6] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks," in *Advances in neural information processing systems*, 2016, pp. 4107–4115.

[7] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv preprint arXiv:1708.08296*, 2017.

[8] S. Rudinac, T.-S. Chua, N. Diaz-Ferreyra, G. Friedland, T. Gornostaja, B. Huet, R. Kaptein, K. Lindén, M.-F. Moens, J. Peltonen *et al.*, "Rethinking summarization and storytelling for modern social multimedia," in *International Conference on Multimedia Modeling*. Springer, 2018, pp. 632–644.

[9] J. Li, C. Xiong, L. Liu, X. Shu, and S. Yan, "Deep face beautification," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 793–794.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[11] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[12] Z. Wang, Q. She, and T. E. Ward, "Generative adversarial networks: A survey and taxonomy," *CoRR*, vol. abs/1906.01529, 2019. [Online]. Available: http://arxiv.org/abs/1906.01529

[13] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of economic perspectives*, vol. 31, no. 2, pp. 211–36, 2017.

[14] Z. Wang, Q. She, A. F. Smeaton, T. E. Ward, and G. Healy, "Neuroscore: A brain-inspired evaluation metric for generative adversarial networks," *CoRR*, vol. abs/1905.04243, 2019. [Online]. Available: http://arxiv.org/abs/1905.04243