The Power of Deep Learning Models: Applications

Sameerunnisa Sk, J. Jabez, V. Maria Anu

Abstract: ***The extraordinary research in the field of unsupervised machine learning has made the non-technical media to expect to see Robot Lords overthrowing humans in near future. Whatever might be the media exaggeration, but the results of recent advances in the research of Deep Learning applications are so beautiful that it has become very difficult to differentiate between the man-made content and computer-made content. This paper tries to establish a ground for new researchers with different real-time applications of Deep Learning. This paper is not a complete study of all applications of Deep Learning, rather it focuses on some of the highly researched themes and popular applications in domains such Image Processing, Sound/Speech Processing, and Video Processing.***

Index Terms: Deep Learning,

I. INTRODUCTION

Artificial Intelligence was founded in 1956 as an academic discipline. It thrived for a decade and then lost its hype as the computing capability of the era was not sufficient to achieve the expected results. Main part of the research in this field has always focused on achieving human-like intelligence in machines trying to mimic the human intelligence. Machine Learning is a part of Artificial Intelligence which focuses on teaching machine to learn and take decisions. It is a group of computer algorithms trained to recognize patterns and perform predictions. Deep Learning has existed in the domain of Machine Learning for a long time but could not gain its high status until lately when scientists realized how Graphics Processing Units (GPUs) could be of great use in teaching machines about our (human) world. Deep Learning is a special class of algorithms which are trained to extract features from large corpus of examples unlike supervised machine learning algorithms. Although unsupervised machine learning algorithms existed before Deep Learning algorithms, they took lot of time on traditional CPUs. Deep Learning exploits the parallel processing of GPUs to speed up the feature extraction and so the overall learning.

II. APPLICATIONS

Deep Learning is being applied in various domains for its ability to find patterns in data, extract features and generate intermediate representations. It is proved to be immensely powerful in mimicking human skills such as seeing, hearing and very recently speaking.

A. Image Processing

The ImageNet project is a large visual database with more than 14 million hand-annotated images. It has more than 20,000 categories consisting of several hundred images. The ImageNet project runs an annual software contest, the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), famously known as *The ImageNet Challenge*. In 2012, AlexNet, a Convolutional Neural Network (CNN) won the ImageNet 2012 challenge with 10.9% top-5 test error points lower than that of the runner up [1]. This was possible due to the utilization of Graphics Processing Units (GPUs) during training, which later became the inception of Deep Learning revolution.

Colorization of grayscale images is a highly creative task which requires the knowledge of different aspects like culture, time, weather and lighting conditions of any given grayscale image. The process of colorizing a grayscale image by a human takes days of manual work using an Image Manipulation tool. There has been some research done to automatically colorize the grayscale image [2][3][4][5]. With the latest advancements in Deep Learning, a deep neural network (DNN) can colorize a grayscale image in a matter of seconds without any human intervention [6][7].

In 2014, an exciting new paper was published by Ian Goodfellow et. al. named Generative Adversarial Networks (GANs) [8]. This paper was the beginning of plethora of possibilities in the area of DNNs. GANs are composed of two separate networks: a) Generative model: generates new real-looking fake data from the training data fed into it; b) Discriminative model: acts as a binary classifier and checks whether the data generated by Generative model is real or fake. Both the networks are fighting against each other; trying to fool each other in the Zero-Sum Game Framework. Generative model improves over time based on the Discriminative model's loss output, to the point where generated fake data is indistinguishable from original data of training dataset.

Different kinds of GANs used for this study:

1. GAN – Generative Adversarial Networks (2014) [8]
2. CGAN – Conditional Generative Adversarial Nets (2014) [19]
3. DCGAN - Unsupervised representation learning with deep convolutional generative adversarial networks (2015) [36]
4. SGAN - Stacked Generative Adversarial Networks (2017) [35]

Revised Version Manuscript Received on 16 September, 2019.
Sameerunnisa Sk, Department of CSE, VJIT, Hyderabad, India.
Dr. J. Jabez, Department of IT, SIST, Chennai, India.
Dr. V. Maria Anu, Department of CSE, SIST, Chennai, India

5. StackGAN - StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks [16]
6. CoGAN - Coupled Generative Adversarial Networks (2016) [34]
7. WGAN - Wasserstein Generative Adversarial Networks (2017) [33]
8. CycleGAN - Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks (2017) [18]
9. ISGAN – Invisible Steganography via Generative Adversarial Networks [41]

Another extraordinary work by NVIDIA's AI research team developed a GAN which takes CelebA image dataset [9] as input and generates completely fake celebrity faces [10]. The Fig 1 (reproduced from [11]) is an example of some fake faces generated by the network. This work demonstrates how their GAN starts generating 4x4 images and able to generate a resolution of 1024 x 1024 with a mere training of 20 days.

Fig 1: Fake celebrity faces generated – all of them are fake

Adobe is working on building AI Tools as part of their proprietary image manipulation software, Photoshop; that lets the user instantly edit photos and videos [12]. The tool is named Scene Stitch, that looks into Adobe's stock photos library and snaps in similar scenes and perspectives to whichever image the user wishes to edit. A peculiar work generated flower paintings by the adaptation of WGAN that look like works of van Gogh [13]. The full project along with code is available at [14] and the result is depicted in Fig 2 (reproduced from [14]). Another similar computer generated artwork produced with a variant of GAN called CycleGAN generates abstract forest images [15] and the result is depicted in Fig 3 (reproduced from [15]). This work is derived from the original CycleGAN paper [18] which worked on image to image style transfer and produced some extravagant results like turning a horse to zebra and vice versa, photograph to monet and vice versa, summer image to winter image and vice versa.

This extraordinary paper [16] published in 2016 generates a series of images based on the given textual description. The results of this paper is depicted in Fig 4 (reproduced from [17]). This is an application of SGAN.

pix2pix is another project which is an implementation of a variant of adversarial networks, CGAN. Main eye catchy results of this project are generation of images from line drawings, black and white images to color images, labels to façade, labels to street scene, which is represented here in Fig 5 (reproduced from [20]). The network is trained only on 400 images for 2 hours which could produce some really amazing results [20].

Fig 2: Novel Generation of Flower Paintings

Fig 3: Abstract forest images generated by CycleGAN

Fig 4: Image generated out of a given textural description

[21] uses CoGANs with joint training to solve the UNsupervised Image-to-image Translation (UNIT) problem by a significant improvement of 84.88% to 90.53% in SVHN->MNIST [37][38] task. The results of the respective paper for reference is shown in the Fig 6 (reproduced from [22]).

Google Lens [23] is a AI-powered visual search tool that was introduced by Google in 2017. This tool, a part of smartphone camera when used identifies different objects in a photo shot by the camera. Then the user may choose any one of the objects and the Lens gives dynamic results based on the chosen object. Lens uses image recognition techniques and TensorFlow, Google's open source machine learning framework to give the meaningful results as per the chosen image object.

This tool is really a tool for curiosity [39]. A photo shot in my backyard gave the result as shown in the Fig 7. This tool has been shown to be useful in education settings as well [40].

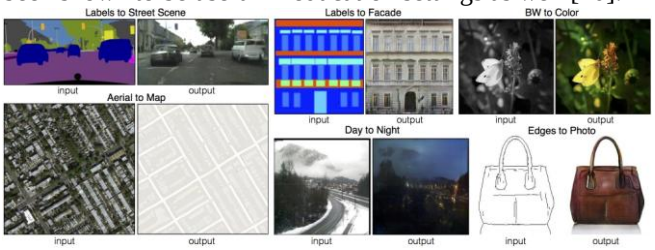

Fig 5: Results produced from pix2pix project

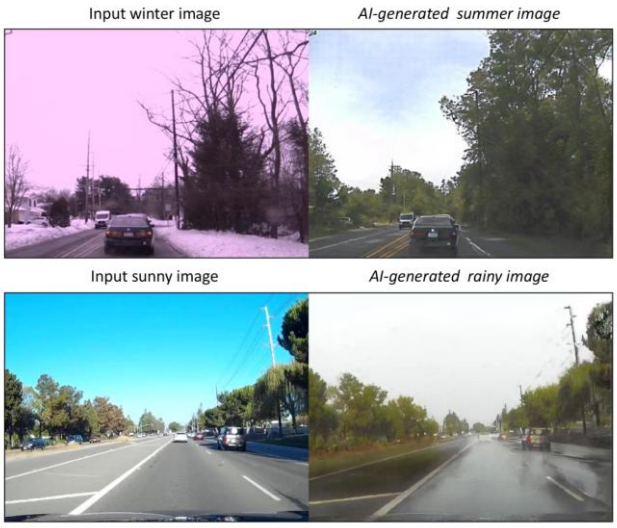

Fig 6: Street Scene Image Translation Results

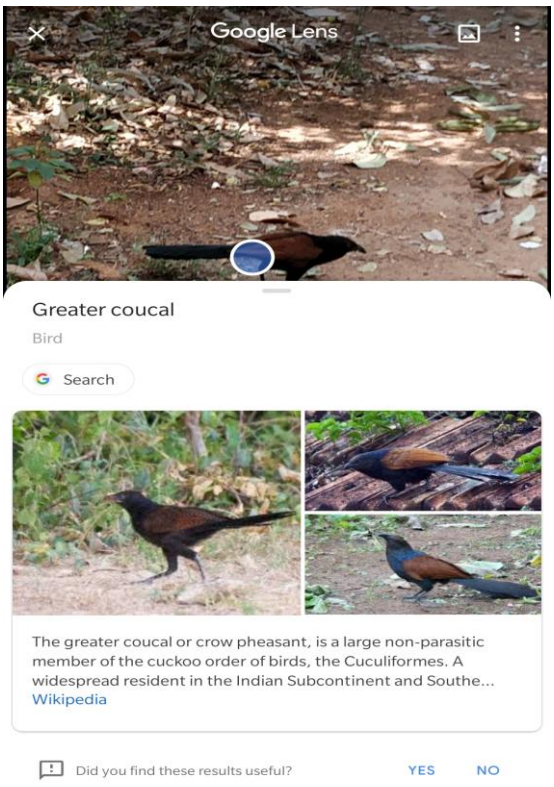

Fig 7: A working example of real-time image search of Google Lens application over the phone

Steganography is an art of covertly communicating a digital message via hiding content. There has been plenty of research in this domain for a long time. In recent times, GANs have been proved to be successful in hiding messages efficiently due to their capability of generating new content from trained data. SteganoGAN [41] makes use of the encoder-decoder concept of GAN to achieve a hidden image payload of 4.4 bpp and also evades standard steganalysis tools. ISGAN [42] is another steganographic technique that improves the security of the secret image by introducing a GAN to minimize the Kullback-Leibler & Jensen-Shannon divergence to avoid detection by steganalysis tools.

B. Speech Synthesis

Speech Synthesis is one of the oldest researches human-computer interaction is trying to master to achieve natural conversation with computers. Text-to-Speech (TTS) systems apply these techniques. Traditional TTS techniques are of two types: Concatenative and Parametric. As the name suggests, the Concatenative TTS method records small audio clips and concatenates them together to form a bigger audio. It is nearly impossible to pre-record all possible words with all possible prosodies, emotions and stress. Parametric TTS on the other hand generates speech based on various parameters like linguistic and phonetic features. Though Parametric TTS looks promising on the front, it could not match natural human speech. In 2016, the researchers at Google's DeepMind created a DNN named WaveNet [24], a new kind of Parametric TTS which outperformed the then best existing TTS systems. WaveNet used Dilated Causal Convolutional Layers to achieve this success. Technically speaking, a new waveform is generated based on the product of conditional probabilities of all previously generated samples. Once the WaveNet is fed with the text that is transformed into a sequence of linguistic and phonetic features, the result is outstanding. Google voice search and Google Assistant uses WaveNet synthesizer. In 2017, a new technique was presented which uses a slightly varied version of Long Short Term Memory Recurrent Neural Network (LSTM-RNN). They presented a technique, Nuance Hierarchical Cascaded Mapping, with the sequence of inputs connected to its respective layers. This technique improved a lot over vanilla version of LSTM-RNN for TTS. In the same year, Google came up with a new method called Tacotron [25]. While WaveNet builds its models based on waveform, Tacotron builds on spectrograms. Though spectrogram is a great representation of speech, it loses information on the phase position in the frame. So, the phase from input spectrogram is estimated to reconstruct the waves with Griffin-Lim algorithm [26]. Their latest work [27] discusses prosody transfer of one speech signal to another and so improving the quality of synthetic speech to be near equivalent to human speech. Google Duplex, a technology that can conduct natural conversations over the phone to carry out "real world" tasks. It is a Recurrent Neural Network (RNN) at its core designed to cope with the challenges faced during interactions with real humans. Google Duplex uses a combination of concatenative TTS engine and synthesis TTS engine (Tacotron and WaveNet) to control intonation depending on the circumstance [28].

C. Sound Generation

In 2016, researchers from MIT's CSAIL published a curious application named "Visually Indicated Sounds" [29]. This paper shows how a software application produces realistic sounds for given mute video samples. This application uses an algorithm which is a combination of both Convolutional Neural Network (CNN) and LSTM-RNN. The fake sounds produced by the algorithm sounded so real they fooled humans alike [30].

[31] produced an entire sound track out of samples generated by DNN. The track was named "NeuralFunk" by the producer. A TensorFlow implementation of WaveNet was trained over 62000 samples which produced a decent sound track.

D. Video Processing & Generation

In 2018, [32] worked on Video to Video synthesis, a relatively unexplored area compared to its image counterpart. This vid2vid method is developed by NVIDIA research team and they even open-sourced the vid2vid code which is based on PyTorch. With the pretext of GANs being already well researched in the image-to-image synthesis problems, the same architecture is used for the development of this method. The researchers designed a feed-forward network to learn a mapping from the past P source images and the past P-1 generated images to a newly generated output image. They used Guassian Mixture Models and a feature embedding scheme to propose a generative method to output multiple videos with different visual appearances depending on sampling of various feature vectors. The different datasets that were tested on include Cityscapes, Apolloscape, Face video dataset, Dance video dataset.

III. DATASETS USED IN THE REFERRED PAPERS IN THE STUDY

1. CelebA Celebrity Face Image Dataset
2. SVHN
3. MNIST
4. Cityscapes, Apolloscape, Face Video, Dance Video

IV. SUMMARY AND ANALYSIS

The clear winner among the Deep Neural Networks is GAN, due to its generative capability. Most of the applications used the generative ability of GANs while varying the basic model to achieve near-human accuracy. Image-to-image synthesis along with image style transfer has been the focus of GAN applications in many of the works. Application of GANs in Steganography is a promising direction to consider GAN for serious work rather than just some fun experimental artwork. Video2Video Synthesis by NVIDIA research team is a beautiful work while it being a less explored problem.

While Image Synthesis was extensively taking advantage of GANs, the applications of Sound Synthesis and Generation were still developed with vanilla DNN algorithms like CNN, RNN, LSTM, and VAE (Variational Autoencoder).

Sl. No.	DNN Model	Paper / Project	Year of Publication	Application
Image Processing based Applications				
1	DNN	Learning Representations for Automatic Colorization	2016	Grayscale Image Colorization
2	GAN	Progressive Growing of GANs for Improved Quality, Stability, and Variation	2018	Fake Celebrity Face Generator
3	WGAN	GANGogh: Creating Art with GANs	2018	Van Gogh look-alike flower paintings generator
4	CycleGAN	Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks	2017	Style transfer of horse to zebra and vice versa
5	CycleGAN	CycleGANs to Create Computer-generated Art	2018	Abstract forest images generator
6	SGAN	Stacked Generative Adversarial Networks	2017	Image generation from given text description
7	CGAN	Image-to-Image Translation with Conditional Adversarial Networks	2017	Pix2pix
8	CoGAN	Coupled Generative Adversarial Networks	2016	Unsupervised Image-to-image Translation Networks
9	Image Recognition DNN & TensorFlow	Google Lens	2017	Google Lens

10	SteganoGAN	SteganoGAN: High Capacity Image Steganography with GANs	2019	Image Steganography with 4.4 bpp cover image generation
11	ISGAN	Invisible steganography via generative adversarial networks	2019	Image Steganography
Sound Synthesis & Generation based Applications				
12	CNN	WaveNet: A Generative Model for Raw Audio	2016	Google WaveNet
14	RNN	Tacotron: Towards End-to-End Speech Synthesis	2017	Google Tacotron
15	WaveNet & Tacotron g	Google Duplex	2018	Google Duplex
16	CNN, LSTM-RNN	Visually Indicated Sounds	2016	Sound Synthesis for given video
17	WaveNet, TensorFlow, VAE	NeuralFunk: Combining Deep Learning with Sound Design	2018	NeuralFunk: Sound Track composition from NN generated sound samples
Video Synthesis & Generation based Applications				
18	GMM, GAN	Video-to-Video Synthesis	2018	Vid2Vid by NVIDIA

Table 1: Categorical overview of DNN models with application mapping

V. CONCLUSION

In this paper, I have performed a detailed analysis of recent advances in deep-learning based research efforts applied in the domains of Image Processing, Audio Processing and Video Processing. I have identified 25 relevant papers and 17 web resources, examining the particular area and problem they address, models employed, datasets used and the images from respective papers are reproduced with proper citation. My findings indicate there are prominent applications in deep learning in the domain of image processing at large with reducing number of applications in audio and video processing. There is very little research though in other domains of interest like Video Steganography. Though it is not a new domain where machine learning was applied. In recent times DL is also being applied to achieve higher accuracy and more secret message payload capacity. But, there is little research on applying GANs in the area of Video Steganography although GANs look to be promising in such contexts. For future work, I plan to find out further evidence of deep-learning based applications in other not-so-famous areas with more detailed analysis of different techniques being used.

REFERENCES

1. Alex Krizhevsky, Ilya Sutskeve, Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", *Advances in Neural Information Processing Systems 25*, 2012.
2. S. Liu and X. Zhang, "Automatic grayscale image colorization using histogram regression," *Pattern Recogn. Lett. 33* (13), 2012, pp. 1673–1681.
3. Y. Rathore, et al., "Colorization of gray scale images using fully automated approach," *Int. J. Electron. Commun. Technol. 1*, 2010, pp. 16–19.
4. L. F. M. Vieira, et al., "Fully automatic coloring of grayscale images," *Image Vision Comput. 25 (1)*, 2007, pp. 50– 60.
5. T. Welsh, M. Ashikhmin, and K. Mueller, "Transferring color to greyscale images," *ACM Trans. Graph. 21 (3)* , 2002, pp. 277–280.
6. Larsson G., Maire M., Shakhnarovich G. (2016) "Learning Representations for Automatic Colorization", In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – *European Conference on Computer Vision* 2016. *Lecture Notes in Computer Science*, vol 9908. Springer, Cham
7. Emil Wallner. (2017, October, 29). *How to colorize black & white photos with just 100 lines of neural network code* [Online]. Available: https://medium.freecodecamp.org/colorize-b-w-photos-with-a-100-line-neural-network-53d9b4449f8d
8. Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville and Yoshua Bengio. "Generative Adversarial Nets." *Advances in Neural Information Processing Systems*, 2014.
9. CelebA Dataset [Online]. Available: http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html
10. Tero Karras, Timo Aila, Samuli Laine, Jaakko Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation", *International Conference on Learning Representations*, 2018.
11. Shubham Sharma. (2018, August, 4). *Celebrity Face Generation using GANs (Tensorflow Implementation)* [Online]. Available: https://medium.com/coinmonks/celebrity-face-generation-using-gans-tensorflow-implementation-eaa2001eef86
12. James Vincent. (2017, October, 24). *Adobe's prototype AI tools let you instantly edit photos and videos* [Online]. Available: https://www.theverge.com/2017/10/24/16533374/ai-fake-images-videos-edit-adobe-sensei
13. Kenny Jones. (2017, June, 19). *GANGogh: Creating Art with GANs* [Online]. Available: https://towardsdatascience.com/gangogh-creating-art-with-gans-8d087d8f74a1
14. GANGogh Project [Online]. Available: https://github.com/rkjones4/GANGogh
15. Zach Monge. (2019, February, 18). *CycleGANs to Create Computer-Generated Art* [Online]. Available: https://towardsdatascience.com/cyclegans-to-create-computer-generated-art-161082601709
16. Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, Dimitris Metaxas, "StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks", *IEEE International Conference on Computer Vision (ICCV)*, 2017.
17. StackGAN Project [Online]. Available: https://github.com/hanzhanggit/StackGAN
18. Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks", *IEEE International Conference on Computer Vision (ICCV)*, 2017.

19. Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks", *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
20. pix2pix Project [Online]. Available: https://github.com/phillipi/pix2pix
21. Ming-Yu Liu, Thomas Breuel, Jan Kautz, "Unsupervised Image-to-Image Translation Networks", *31st Conference on Neural Information Processing Systems*, 2017.
22. Ming-Yu Liu, Thomas Breuel, Jan Kautz, "Unsupervised Image-to-Image Translation Networks", 31st Conference on Neural Information Processing Systems, (2017) [Online]. Available: https://research.nvidia.com/publication/2017-12_Unsupervised-Image-to-Image-Translation
23. Google Lens: https://lens.google.com/
24. Oord, A.V., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A.W., & Kavukcuoglu, K. "WaveNet: A Generative Model for Raw Audio". *SSW*, 2016.
25. Y Wang et. al., "Tacotron: Towards End-to-End Speech Synthesis", *Interspeech*, 2017.
26. D. Griffin ; Jae Lim, "Signal estimation from modified short-time Fourier transform", 1984.
27. RJ Skerry-Ryan et al. "Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron" 2018.
28. Yaniv Leviathan. (2018, May, 8). *Google Duplex: An AI System for Accomplishing Real-World Tasks over the Phone* [Online]. Available: https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html
29. Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H. Adelson, and William T. Freeman, "Visually Indicated Sounds", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2016).
30. Adam Conner-Simons. (2016, June, 13). *Artificial Intelligence produces realistic sounds that fool humans* [Online]. Available: http://news.mit.edu/2016/artificial-intelligence-produces-realistic-sounds-0613
31. Max Frenzel. (2018, October, 25). *NeuralFunk: Combining Deep Learning with Sound Design* [Online]. Available: https://towardsdatascience.com/neuralfunk-combining-deep-learning-with-sound-design-91935759d628
32. Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro, "Video-to-Video Synthesis", *Neural Information Processing Systems*, 2018.
33. Martin Arjovsky, Soumith Chintala, Léon Bottou, "Wasserstein Generative Adversarial Networks"; *Proceedings of the 34th International Conference on Machine Learning* (PMLR) 70:214-223, 2017.
34. Ming-Yu Liu, Oncel Tuzel, "Coupled Generative Adversarial Networks"; *Advances in Neural Information Processing Systems 29*, 2016.
35. Xun Huang, Yixuan Li, Omid Poursaeed, John Hopcroft, Serge Belongie, "Stacked Generative Adversarial Networks"; *IEEE Conference on Computer Vision and Pattern Recognition*. 2016
36. Alec Radford, Luke Metz, Soumith Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks"; *International Conference on Learning Representations*, 2016
37. The Street View House Numbers Dataset [Online]. Available: http://ufldl.stanford.edu/housenumbers/
38. The MNIST Database of handwritten digits [Online]. Available: http://yann.lecun.com/exdb/mnist/
39. Aparna Chennapragada. (2018, December, 19). *The era of the camera: Google Lens, one year in* [Online]. Available: https://www.blog.google/perspectives/aparna-chennapragada/google-lens-one-year/
40. Shapovalov, Yevhenii B., Zhanna I. Bilyk, Artem I. Atamas and Viktor B. Shapovalov. "The Potential of Using Google Expeditions and Google Lens Tools under STEM-education in Ukraine." *Computing Research Repository* (CoRR) abs/1808.06465. 2018.
41. Kevin A. Zhang, Alfredo Cuesta-Infante, Lei Xu, Kalyan Veeramachaneni, "SteganoGAN: High Capacity Image Steganography with GANs", Computer Vision, 2019.
42. Zhang, R., Dong, S. & Liu, J., "Invisible steganography via generative adversarial networks" *Multimedia Tools and Applications*, Volume 78, Issue 7, pp 8559-8575, 2019.

AUTHORS PROFILE

Sameerunnisa Sk , M.Tech, (Ph.D)
Research Scholar, SIST, Chennai
Working as Assistant Professor, VJIT, Hyderabad,
Published 8 papers in International and National Journals
Research Area: Machine Learning

Dr. J. Jabez , M.E., Ph.D
Published 25 papers in International and National Journals
Research Area: Data Mining, Machine Learning
Membership in CSI

Dr. V. Maria Anu , M.E., Ph.D
One of the authors of "Computer Programming and Numerical Methods", Airwalk Publications, 2017
Published 13 papers in International Journals and 1 paper in National Journal
Published 8 papers in International Conferences
Membership in IAENG, ISRS