TEXT TO IMAGE GENERATION USING HYBRID ATTENTION GENERATIVE ADVERSARIAL NETWORK

Bhavya Bordia[1], Shaswat Patel[2], Biju R Mohan[3], Supreeth G[4]

[1,2,3,4]Department of Information Technology NITK Surathkal, Karnataka, India 575025

Received: 14 March 2020 Revised and Accepted: 8 July 2020

ABSTRACT: Visualization is essential because it makes learning, creating, planning much easier, and to visualize we need a description that provides meaning to the visual. Due to the advancement in Machine Learning Algorithms, they can help in translating the descriptions to visuals. Generative Adversarial Networks (also known as GANs) can be used to create a set of images from the text which are a form of descriptions. Generative models algorithms come under unsupervised machine learning. GANs have many applications and they can be used in visualizing and creating models according to the need just by describing it. They can be used in architect planning, home designing. They can also be used in creating animations, rather it can make the process faster and hence can be used to make virtual games. GANs can be used also be used in the apparel industry and advertising sectors. But these all are the advance use of the algorithms which may be implemented in the future. Our project aim is just to create a set of images from captions given by users in text format, which are semantically matching with the text and seems too realistic. We will be using different generative models for this purpose.

KEYWORDS—Generative Adversarial Networks, Text-to- Image Generation, Image Re-description, Unsupervised Machine Learning

I. INTRODUCTION

Text-to-Image generation means that to translate text to images which are visually realistic and has same semantic meaning as the text used to describe the image. Because of this reason it can have wide range of applications but it needs a good research to make the algorithm better so that it can feasible to use. Due to its challenging nature and vast application it has become a crucial field for research in machine learning communities. Generative Adversarial Networks are combination of Generators which are trained to generate sample images more likely as true image and Discriminators which are used to distinguish between the generated images and true image distribution. The Generator try to fool the discriminator and discriminator try to catch it and hence they compete with each other. This is used for

calculating the loss which is further provided as feedback to generator.

Due to the rapid growth of machine learning in recent times, many areas of research previously considered to be extremely difficult and complex have now become a reality. Text to image generation is one such area that has seen enormous progress, largely due to the advancements of machine learning algorithms. Image generation from just text has its uses in multiple fields and the impact this can have is huge. With the invention of fast and sophisticated machinery along with the increasing rate of researches, the generation of images is bound to become simpler. The number of applications this can actually influence is enormous, right from schools where teachers can use this to help their students visualize to critical military machinery.

Generating images just by description or set of captions can help in creating animation, creating visuals and graphics for game development which still requires a team of designers. With this automated image generation, the creation of animation can be done in lesser time and with lesser people. Nowadays GANs are used in face aging apps and Image editing apps. Moreover, GANs can be used in Architectural design and virtual home decorating.

Even with all the advancements in technology, working with GANs has a few difficulties. The structure of the network is generally complex and training such a network is a really hard task. The process of training the network has to take care of balancing the generator and discriminator which can cause over-fitting when not done the right way. Without adequate knowledge of the learning rates of the discriminator and generator, the image generation process becomes ineffective.

II. LITERATURE SURVEY

Before GANs, text to image generation was possible by using algorithms like PixelCNN[5] and Deep Recurrent Attention Writer (DRAW)[1]. In the former algorithm, an image is synthesized from captions with a multi-scale model structure, whereas the latter algorithm mainly focuses on filtering out important words from the caption and iteratively draws image patches. PixelCNN is an auto regressive model which uses standard convolutional layers to capture a bounded receptive field and compute features for all pixel positions at once. It learns the pixel distribution function and later builds images from the distribution. Training the dataset is time consuming and the obtained results aren't of a great quality either.[5] Whereas DRAW follows the replication behaviour like an artist. Instead of creating an image instantly, it shift towards a more natural form of image construction, in which parts of a scene are created independently from others, and approximate sketches are successively refined. But as a typical recurrent network it takes a lot of time for training.
[1] After thorough studies, it has been found that Generative Adversarial Networks outperforms every algorithm on the basis of generating better image and performance. Generative adversarial networks (GANs) are algorithmic architectures that use two neural networks, pitting one against the other (thus the "adversarial") in order to generate new, synthetic instances of data that can pass for real data. GANs network consists of generator and the discriminator network which compete against each other to give the desired result. In GANs, the generator tries to fool the discriminator by generating fake images whereas the discriminator tries to identify the fake image generated by generator. So there is trade of accuracy between one another,i.e. if one performs good then other will perform bad and other will then try to improve showing the competitive nature.[4]

Generative Adversarial Networks have also proved to ef- fective in Image to Image translation[3].Cycle GANS[12] are used for Image to Image translations.

The initial idea of generating images from any text input was first given Self-Attention Generative Adversarial Networks [10]. This paper describes the basic conceptual idea behind the working of generative adversarial networks aka GANs and conditional adversarial networks while focusing mainly on GANs. It consists of two main components(neural networks), the generator network, and discriminator network, where both the networks are competing against each other to outperform the other one. Consider generator model G, for a given sample, which captures the distribution of the data and the discriminator model D, for any given sample, calculates the probability of a certain image belonging to the distribution of data above. These two networks work so that finally, the generator is able to produce good images from the input noise and the discriminator is trained to try and identify the image correctly.

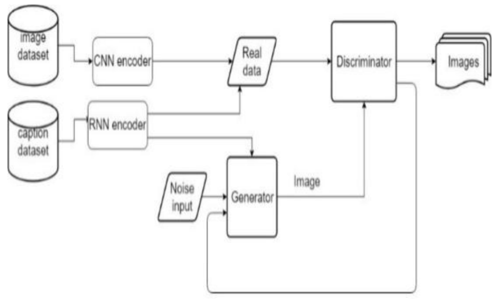

Fig. 1. Initial Approach for generating Image

There are several generative network algorithms that can be used to produce images based on the requirement. Attentional GAN [8] which focuses on building photo-realistic images from text. In most of the text to image generating algorithms, individual words are not considered and the whole sentence is considered while creating a new image, but here individual words are considered and weights are given for each sub- regions so that generated image has all given features from the whole caption.

Tingting et. al have proposed a GAN Model called Mirror GAN [6]. It generates images from captions and then again does image captioning and tries to match the generated caption with the real caption. MirrorGAN consist of three modules: a semantic text embedding module (STEM) which filters out word and sentence level embeddings, Global-Local collaborative attentive module(GLAM) that has cascaded architecture for generating target images and a semantic text regeneration and alignment module (STREAM) which regenerate the text from the generated image and further tries to match with the real text description. Our proposed methodology also uses the same concept of regenerating text from images and calculating loss from it.

One other GAN is StackGAN proposed by Han Zhang et. al. [11] They have generated photo-realistic images by decomposing the hard problem into more manageable sub- problems through a sketch-refinement process. The stage-I GAN draws the low-resolution image with shape and colors of the object based on captions. Then the stage-II takes the text description and the image generated by stage-I as input and generates high-resolution (256 × 256) image.

Also Image completion using structure and texture GAN network proposes a method for the image completion by decomposing image completion into two sub-task: structure (the underlying sketch) completion and texture completion.[2] This method is more interpretable and synthesizes visually more plausible contents in missing regions compared to the baseline.The method of image completion using GANs helps to improve the image resolution which is generated from the text, but not directly convert text to realistic image. [2]
In Image Captioning with Semantic Attention shows the connection between two major artificial intelligence fields: computer vision and natural language processing. The algo- rithm learns to selectively attend to semantic concept proposals and fuse them into hidden states and outputs of recurrent neural networks.[9]As this method combines bottom up and top down approach it becomes more complex, which makes it difficult to implement.[9]

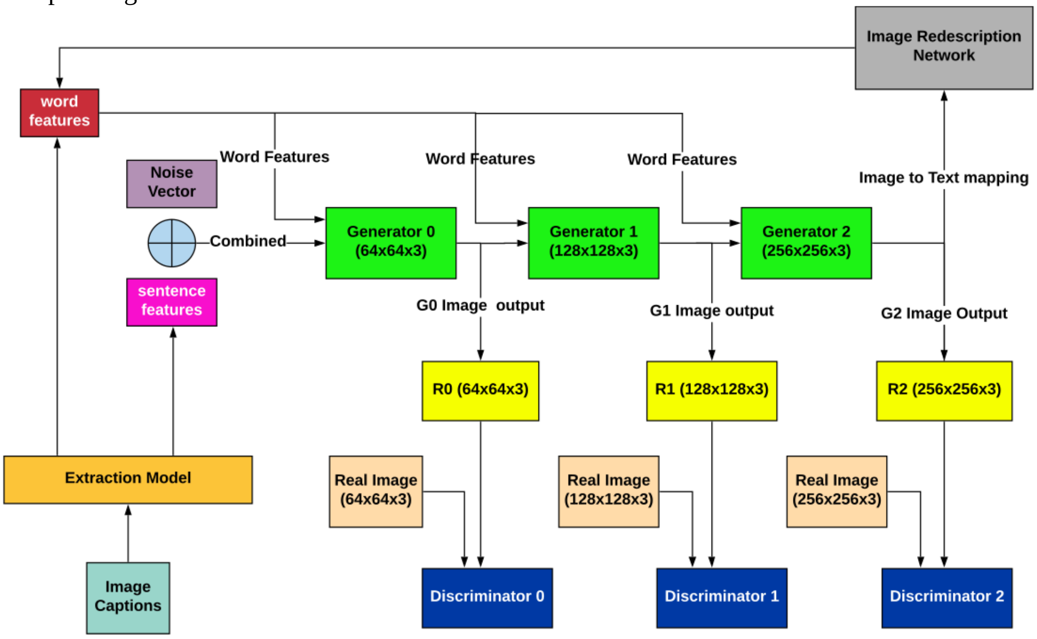

Fig. 2. Proposed Model Architecture

For our proposed model we need the image re-description. It requires the knowledge of image processing and deep neural network. Parth et al[7] have discussed object recognition and machine translation model and how to implement it. According to their proposed model Image is given as input to a fixed dimensional vector and then it is converted to word vector. Recurrent neural network(RNN) decoder is used to generate captions from word vector.

III. PROPOSED METHODOLOGY

Here we are proposing another model for the generation of images. The model can be broadly divided into 3 parts :

- Model for extracting the features from text
- GAN network to generate the images from the captions using self-attention technique
- Image re-description model

As shown in Fig. 2, first the captions as text are given to the network for feature extraction from them. These features are converted to tensors which are then clubbed with the noise tensors to increase the variance in the data. The final clubbed tensors are then passed to the generator networks to give the generated images. The generated images are then re-described using the network in the form of captions and are compared with the original caption to calculated the loss and the loss is propagated back to the network for learning.

A. Extraction of features from Captions

The first step which is followed in the extraction of features from captions is the tokenization of caption sentences and building the vocabulary of the captions which are present. The building of vocabulary is followed by assigning id to the captions based on the occurrence in the caption and then replacing the text caption with the numbers. The numbers are mapped using idtoword and wordtoid dictionary so that it become easy to retrive the actual word associated with the number.

After the captions are converted to numbers, the task is to map the captions with the image. This is done with the help of Recurrent Neural Network (RNN). For the purpose of conversion a bi-directional LSTM is used which is having 1 recurrent layer and having a drop probabilty of 0.5 with 128 hidden neurons. The weights of the model are initialized uniformly in a range of -0.1 to +0.1 and are layers is set trainable. The output of the model are the word embedding and the sentence embedding. The shape of the word embedding is *batch hiddenSize numDirections seqLen* whereas the shape of the sentence embedding is *self.nhidden self.numDirections* where the number of directions is 2 and the number of hidden neurons is 128. The sentence and the word embedding are then transposed according the requirement of the model.

B. GAN Network × × ×

The GAN Network co×nsists for 3 Generator namely $G0$,$G1$, $G2$ for producing 64 64 , 12×8 128, 25×6 256 × respectively and each is having discriminator network $D0$, $D1$ and $D2$ respectively. The output of the previous section is given as the input to the first Generator $G0$ which is used to produce the fake image $F0$. The initial Generator takes the linear input from the previous sample and starts to generate the images of size 4 4. After the first initial generation the generator does the work of upsampling the images to 8 8 followed till 64 64 doubling the size at every layer. After the generation of the image of size 64×64, the generated image along with the real image, real label and fake label is passed to the discriminator $D0$ where it learns to identify the d×ifference between the real and the fake image with the appropriate ×label associated with it. In the learning step the discriminator loss and the generator loss is calculated and the error is propagated backward×s.

We have used NVIDIA Tesla V100 Tensor Core 32 GB configurations which is one of the most advanced data center GPU ever built to accelerate AI, high performance computing (HPC), data science and graphics.

In order to evaluate the models mathematically, the inception score metric has been used. Inception Score(IS) is an objective metric for evaluating the quality of the generated images. The above score seeks to capture two main properties of an image, the quality and diversity. The calculated inception scores of different Inception Score is mentioned in Table-1.

For the next generator $G1$ the input is the first hidden
state of the generator $G0$. Here the generator $G1$ does its routine work of creating a new image(here, the size of the image generated is 128 128 3) or hidden state. Here the generator not only focus on the hidden state output it also pays attention×on the×sentence embedding which are passed. It combines the sentence embedding with the hidden state output to give the generate image, $F1$. Again the generated image $F1$ is passed through the discriminator network along with the real image and the real and fake label. The loss is calculated and propagated backward.

The work of the last generator $G2$ is similar to that of $G1$. Here $G2$ produces images of size 256 256 focusing on the sentence embedding using the ×attention.

C. Image Re-Description Model

This is the final phase of the model. In this the output image from the last generator network, $G2$ is of size 256 256 is first converted to features using the cnn e×ncoder. The features extracted from the image are then decoded using the rnn decoder to obtain the caption text. The cnn encoder and the rnn decoder are pre-trained model earlier on the same dataset to caption the image. They are using bidirection LSTM to encode and decode the image. After obtaining the text from the image the caption loss is calculated using the cross entropy loss and then the error is propagated back to the generator network.

IV. RESULTS AND DISCUSSION

The proposed model has been fully implemented and tested with the Caltech UCSD Birds 200 (CUB-200). This is an image dataset with pictures of 200 bird species (mostly found in North America). Along with our proposed model, text-to-image generation has been implemented using three other models to help us evaluate the given model. So we have implemented Deep Convolutional GAN and Attention GAN. Each model has implemented and tested on the above given CUB-200 dataset.

SL No.	Model	IS Score
1.	Dataset	5.265
2.	DCGAN	3.457
3.	Attention GAN	4.436
4.	Proposed GAN	5.089

TABLE IDIFFERENT GANS AND THEIR INCEPTION SCORE

A large number of generated images are classified using Inception Scores. The Inception Score, or IS for short, is an objective metric for evaluating the quality of generated images, specifically synthetic images output by generative adversarial network models.Specifically, the probability of the image belonging to each class is predicted. These predictions are then summarized into the inception score. The score seeks to capture two main properties of the image i.e, image quality and image diversity. The inception score of the CUB ”birds” dataset was found to be 5.265 and the image generated by attentional GAN using the captions has an inception score of 4.215 and IS score obtained using mirror GAN is 4.861. The hybrid model gave satisfactory results with an IS score of 5.089.

The output images generated at 650 epochs for the following captions are in the figure
Caption 1: This bird is white, yellow, and black in color, and has a light blue beak.
Caption 2: This bird is grey with black and has a very short beak.
Caption 3: The tail is big as compared to the bird’s body, the head is small and the bill is short and pointed.
Caption 4: This bird is white and brown in color with a pointy skinny beak and dark eye ring.
Caption 5: A blue bird with a grey rump and wing tips.

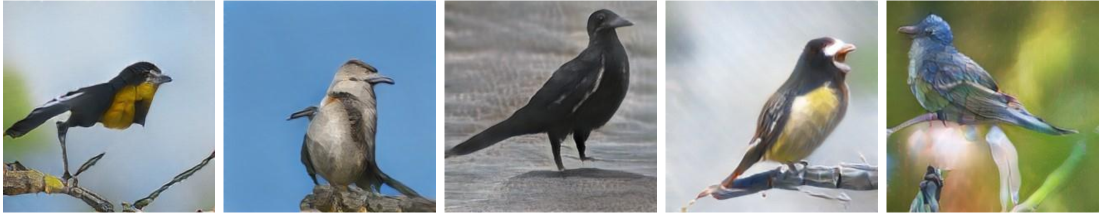

Fig. 3. Birds generated by DC GANs

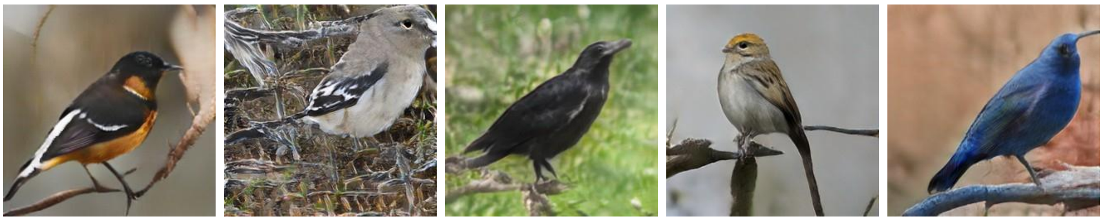

Fig. 4. Birds generated by Attention GANs

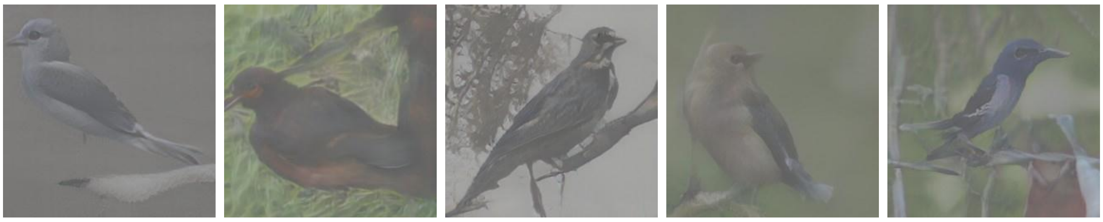

Fig. 5. Birds generated by Mirror GANs

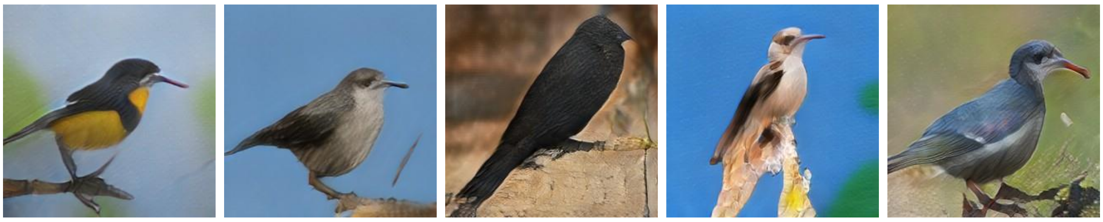

Fig. 6. Birds generated by our Proposed GANs Model

V. CONCLUSION AND FUTURE WORK

The images generated from the proposed model are good in some cases and bad in some cases compared to mirror GANs. The inception score of 5.089 suggests that images generated from the proposed model are distinct and resembles birds closely. More the inception score more is the variance of the generated in images. Also more the inception score suggests that the google inception model is able to classify the image distinctly. The proposed model generate images are less realistic compared to the images generated from AttnGANs and MirrorGANs.

The quality of the images generated from the hybrid model by adding some noise at each level of GANs, this will increase the randomness in the data and will allow the model to learn variety of the features. As the model is based on neural networks so more training data is required to learn. The images generated from the proposed more are more sharp compared to different model, smoothing of the images

would improve the quality of the images generated. Using the skip thoughts model to extract the word features and the sentence features from the captions would improve the generated tensors from the text. Skip thoughts model will provide more contextual similarity to the generated images from the GANs network.

Stacking of GANs (current used is 3) could help to get images of higher quality (current max is 256 256 3) which could be extended to 512 512 3. Currently the model is trained on birds dataset which was collected by CalTech University, the model can be extented to generated even faces of unknown people provided it is trained on proper dataset and could help the society in a good way. Also including the word loss function from the attention model and the caption loss function which uses redescription together in the final layer of the Generator network in an alternate fashion will increase the learning parameters and the learning space for the model.

The scope in the field of generating/creating artificial images using a GAN network is vast and could be explored further to the next level. With the proper dataset for learning the generator network, GANs could perform very well. They will decrease the discrepancy and mistakes in the images which are sketched by the artists from the given description by person. Generative Adversarial Networks are model of future.

VI. REFERENCES

[1] Karol Gregor et al. “DRAW: A Recurrent Neural Net- work For Image Generation”. In: *CoRR* abs/1502.04623 (2015). arXiv: 1502.04623. URL: http://arxiv.org/abs/ 1502.04623.
[2] Jingtao Guo and Yi Liu. “Image completion using struc- ture and texture GAN network”. In: *Neurocomputing* 360 (2019), pp. 75 –84. ISSN: 0925-2312. DOI: https: / / doi . org / 10 . 1016 / j . neucom . 2019 . 06 . 010. URL: http : / / www . sciencedirect . com / science / article / pii / S0925231219308379.
[3] P. Isola et al. “Image-to-Image Translation with Con- ditional Adversarial Networks”. In: *2017 IEEE Con- ference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 5967–5976. DOI: 10 . 1109 / CVPR. 2017.632.
[4] Chris Nicholson. *A Beginner’s Guide to Generative Ad- versarial Networks (GANs)*. 2018 (accessed September 2, 2019). URL: https://pathmind.com/wiki/generative- adversarial-network-gan.
[5] Aaron van den Oord et al. “Conditional Image Gener- ation with PixelCNN Decoders”. In: *NIPS*. 2016.
[6] Tingting Qiao et al. “MirrorGAN: Learning Text-to- image Generation by Redescription”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019).
[7] P. Shah, V. Bakrola, and S. Pati. “Image captioning using deep neural architectures”. In: *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*. 2017, pp. 1–4. DOI: 10.1109/ICIIECS.2017.8276124.
[8] Qiuyuan Huang Han Zhang Zhe Gan Xiaolei Huang Xiaodong He Tao Xu Pengchuan Zhang. “AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks”. In: (2018).
[9] Quanzeng You et al. “Image Captioning with Semantic Attention”. In: *CoRR* abs/1603.03925 (2016). arXiv: 1603.03925. URL: http://arxiv.org/abs/1603.03925.
[10] Han Zhang et al. “Self-Attention Generative Adversarial Networks”. In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Pro- ceedings of Machine Learning Research. Long Beach, California, USA: PMLR, 2019, pp. 7354–7363. URL: http://proceedings.mlr.press/v97/zhang19d.html.
[11] Han Zhang et al. “StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks”. In: *CoRR* abs/1612.03242 (2016). arXiv: 1612.03242. URL: http://arxiv.org/abs/1612.03242.
[12] J. Zhu et al. “Unpaired Image-to-Image Translation Us- ing Cycle-Consistent Adversarial Networks”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 2242–2251. DOI: 10 . 1109 / ICCV. 2017.244.

“References/Citation” to be included in “Research Articles”

[1] D Pandey, V Agarwal, “E-commerce Transactions: An Empirical Study”, International Journal of Advanced Research in Computer Science and Software Engineering, Vol 4, Issue 1, pp. 669-671, 2014
[2] ShwetaSankhwar, V Singh, D Pandey, “Requirement engineering paradigm”, Global Journal of Multidisciplinary Studies, Vo; 3, Isue 3, pp.1-8, 2014
[3] T. J. Ansari, D. Pandey and M. Alenezi, STORE: Security Threat Oriented Requirements Engineering Methodology, JournalofKing Saud University – Computer and Information Scienceshttps://doi.org/10.1016/j.jksuci.2018.12.005
[4] S Sankhwar, D Pandey “Software project risk analysis and assessment: A survey”, Global Journal of Multidisciplinary Studies Vol. 3, Issue 5, pp. 144-160, 2014
[5] D Pandey, U Suman, AK Ramani, “an approach to Information Requirement Engineering”, 2011 International Conference on Information Science and Applications, Korea, pp. 1-4, 2011
[6] D Pandey, U Suman, AK Ramani, “Security Requirement Engineering Framework for Developing Secure Software”, International Journal of Computational Intelligence and Information Security, IJCIIS, Australia, Vol. 1, Issue 8, pp. 55-65, 2010.
[7] V Singh, S Sankhwar, D Pandey “A framework for requirement elicitation”, Global Journal of Multidisciplinary Studies, Vol 1, Issue 1, pp. 1-7, 2014
[8] D Pandey, V Pandey, “”Requirement Engineering: An Approach to Quality Software Development, Journal of Global Research in Computer Science Vol. 3, Issue 9, pp. 31-33, 2012
[9] S Sankhwar, D Pandey , “Defending Against Phishing: Case Studies”, International Journal of Advanced Research in Computer Science Vol. 8, Issue 5, pp. 2605-2607, 2017

[10] S Sankhwar, D Pandey, RA Khan, "A Novel Anti-phishing Effectiveness Evaluator Model", International Conference on Information and Communication Technology for Intelligent Systems, pp. 610-618, 2017

[11] Mahra, A. K. (2019) Management Information Technology: Managing the Organisation in Digital Era . International Journal of Advanced Science and Technology 2005-4238 , 29 (6), 8803-8808.

[12] Mahra, A. K. (2019) Teaching Practices In Management Education: Patliputra Journal of Indology ISSN: 2320-351x , 3 (2), 40-50.

[13] Mahra, A. K. (2019) Application Of Knowledge Management In Management Education Anusandhan Vatika 2230-8938, 3 (3), 6-10.

[14] Mahra, A. K. (2019) A Strategic Approach to Information Technology Management International Journal of Advanced Science and Technology 2207-6360 , 28 (20), 1346-1351.

[15] Mahra, A. K. (2019) A Systematic Literature Review On Risk management For Information Technology International Journal of Advanced Science and Technology 2207-6360 , 28 (20), 1352-1358.

[16] Dwivedi, S. M., & Mahra, A. K. (2013). Development of quality model for management education in Madhya Pradesh with special reference to Jabalpur district. Asian Journal of Multidisciplinary Studies, 1 (4), 204-208.