Neural Style based Comics Photo-Caption Generator

Hetvi Choksi[a], Smriti Das[a], Tasneem Gorach[a], Mahima Kriplani[a]*, Bhushan Inje[b]*

[a]Student, Computer Engineering Department, NMIMS MPSTME, Shirpur 425405, Maharashtra, India
[b]Assistant Professor,Computer Engineering Department, NMIMS MPSTME, Shirpur 425405,Maharashtra,India

---***---

ABSTRACT - *In this paper, we propose a solution to generate comic poetry strips, i.e. comic styleimages with poetic captions, given ordinary images as input. For the image stylization, we use a novel technique called comic style transfer, which is based on the neural style transfer algorithm proposed recently, which has been used to transfer style of fine arts and paintings onto ordinary photographs. We use a CNN model to separate the content and style of two images, such that the style of one image is combined with the content of another image to create a comixified image. We also compare various existing style transfer methods and their performance with respect to transferring comic style to another image. The stylized image is further processed to generate poetic captions describing the comics images, using a multiadversarial GAN, so as to generate a comics poetry strip based on ordinary input photograph images.It is a step in the direction of trying to achieve complete automation of the comic book creation process.*

KEYWORDS - Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Bilingual Evaluation under Study (BLEU), Recurrent Neural Network Language Model (RNNLM), Long Short-Term Memory (LSTM), Recurrent Entity Network (REntNet), Amazon Web Service (AWS), Visual Geometry Group (VGG), Generative Adversarial Network (GAN).

1. INTRODUCTION

Comic books have been an active interest for decades, in people belonging to all age groups and nationalities. The USA alone has a comic book industry of 1 Billion USD a year. Comics, cartoons, graphic novels and other such forms of artistic expression are in wide demand by the youth. Generation of comic books has been highly digitalized in the past few years, however most digital comics generating tasks are manual, with the requirement of professional illustrators and designers to generate the comics as well as storyline writers to create a story.

The automation of the comic generation task can revolutionize a major form of printing media. Since, not only the time, effort and money spent on drawing and developing digital illustrations can be saved, but also the ability to convert ordinary real life images into a comixified effect, can help illustrators to develop a series of comic images very easily [2019].

JOURNAL OF STATISTICS AND MANAGEMENT SYSTEMS

Neural Style Transfer was a technique introduced by Gatys et al [2015], which suggested that style and content in an image are separable. Hence, the style of one image can be transferred to another image. A similar approach can be used to transfer comic style from one image to another, as a result, comicifying the image.

Generating stories from images is another task that has been carried out since the past few years. However, comic books constitute a storyline which is contextual, as a result, the images as well as the stories or dialogues describing them are related. The task of deriving context across various images is something that is beyond the scope of neural networks in the present world scenario. In this paper, we aim to propose a model which generates a comic styled effect on images, along with captions which can also be further integrated into a stylized caption model.

Hence, the style of the image as well as caption would be captured, which can further be used to derive context, and generate commercialized comic books.

2. RELATED WORK

Neural Style Transfer was proposed by Gatys, et al, in 2016, to recompose an image in the style of another artist. Convolutional neural networks were used to separate the content and style of an image and then recombine the content of one image with the style of another image using neural representations. The output of the convolutions is feature maps which are differently filtered versions of input image. Hence, artistic images can be created using ordinary content images [2015].

After Gatys etal.'s proposal of Neural Style Transfer in 2015, approaches to increase the speed of computation and resolution of images were proposed [2016]. J. Liao [2016] et al introduced semantic style transfer that can find semantically meaningful components between images that look visually very different, i.e., different in appearance. 'Semantic' for images refers to identifiable objects which represent the high level content of the image. Jing Liao et al. suggested a method known as visual attribute transfer: the transfer of visual information such as color, texture between images. For example, one image can be painting and the other is a photograph of real scene, both depict the same scene.

Few attempts have even been made to convert an image into a comixified effect. The style transfer technique used here is dubbed as comic style transfer, since the style transferred to the content image is that of a comic book image. Pesko and Trzciński [2018], studied different types of style transfer models to find the best model which can convert an input image into a comic style. The various models compared comprised of Gatys' original style transfer model, Huang and Belongie's Adaptive Instance Normalization [2017], Li et al's Universal Style Transfer technique[2017], as well as Photorealistic Stylization(Li et al)[2018]. Their review recognized the Adaptive Instance Normalization technique as working best for transferring the style of 20 different comic images for 20 specific content images downloaded from the internet.

In [2017], Y. Chen at al, proposed a comic CNN which converted photos to comics using a CNN which was trained on 1482 comic images by 10 different artists, compared to the VGG model used by Gatys et al. In 2018, Y. Chen further proposed a model called Cartoon GAN [2018], which converts input images into an image with a cartoonish effect using a generative adversarial network with 2 novel losses, namely, semantic loss and edge preserving adversarial loss. On the other hand, a comixGAN was introduced by Pesko et al, [2018] which extended the concept of style transfer to videos by using keyframe extraction along with highlight score and image aesthetic scoring. Another approach in comixifying videos was taken by Google by introducing Storyboard, an Android app.

3. COMIC STYLE TRANSFER

In neural style transfer, the content reconstructions from higher layers of the network represent high level content but the detailed pixel information is lost. For style reconstructions, the feature correlations amongst multiple layers in a feature space are used. This way the texture information is captured but not the global arrangement. This is the basis of neural style transfer where the CNN develop a hierarchical representation of features [2015]. Comic Style Transfer can be used to transfer the style of one image into another content image such that the result appears to be comixified. Hence, the higher pixel information, obtained from the style image, can be used to properly merge the texture of the style image with the original ordinary photograph.

3.1 EXISTING APPROACHES

Some models proposed by different papers with variations in the style transfer have previously been explored to perform comic style transfer (ComixGAN, comic case study)

3.1.1Gatys

In Gatys' approach, 16 convolutional and 5 pooling layers out of the VGG 19 network are used to represent the style and content information of the images. Three images constitute the input- content, style image and white noise image [2015].

The total loss is calculated using:

$$L_{total} = \alpha * L_{content} + \beta * L_{style}$$

This total loss is minimized using L-BFGS optimization technique.

A major drawback of the network is very less speed of computation. Style transfer of a 512x512 pixel image takes about 1 minute on even the current GPU architectures such as Titan X.

3.1.2. Adaptive Instance Normalization.

The major drawback in Gatys' approach was the very slow speed of computation. This problem was solved by models proposing a feed forward synthesis. However, it compromised the generalization capability of the model to new styles. Thus Adaptive Instance Normalization network was introduced by Huang et al [2017], which reduced the computation speed along with good generalization to arbitrary new styles.

It consists of two networks- a style and a loss network. The loss network is uses backpropogation to minimize the total loss generated. The style network consists of simple encoder- decoder architecture. While, the encoder consists of few layers of the VGG 19 network, the decoder is a mirrored version of the encoder. An ADAIn layer exists between the encoder and decoder which align the mean and variance of the content feature maps to style feature maps. The output of AdaIn is is mapped to the image space to get a stylized image by training a randomly initialized decoder using the loss network.

3.1.3 UST-WC

To represent image style using a gram matrix and a covariance matrix are equally effective, so in this approach instead of using Gram matrix, a covariance matrix is used. It is similar to AdaIn method, the only difference being that instead of aligning mean and variance of the content and style feature maps, the covariance matrices of feature maps are matched. Another difference is instead of using AdaIn layer it uses Whitening and Coloring transform (WCT)[2013]. During training, style transfer which is represented using WCT layer is not used, and only the input content image is reconstructed. The actual style transfer takes place in the WCT layer.

For Image reconstruction purpose, five different encoder-decoder networks are trained. Each of the encoder and decoder consist of VGG-19 layers while WCT is placed as an intermediate between encoder and decoder. The encoder extracts vectorized feature maps f_c and f_s of style and content image after which WCT transforms the $f_c$to match the covariance matrix of f_s and there after the decoder reconstructs the image.

3.1.4 Comix GAN

MaciejPesko, Adam Svystun et al. in their paper have proposed a technique of transforming a video into comics. This is done using Generative Adversarial Networks (GANs) based neural style algorithm [2018]. The task of video comixification is performed in two stages:

- A subset of frames is chosen from the video that provide the most appropriate video context and later these frames are filtered.
- A ComixGAN framework based on existing Style Transfer technique is built. This is used to transform the frames into a comic book and give aesthetic results.

The objective of ComixGAN is to provide comic images with natural, uniform colors and distinct, clear edges. Data used for training the model consists of real images obtained from the MS-COCO dataset and comic images which are key frames of different cartoons.

4. CAPTIONING

Comic books often consist of dialogues in speech bubbles which accompany the comics images. This, however comprises not only the detection of different characters in the comic books, but also generating different story lines in the context of a certain character, along with the sentences appearing to be in the semantics of a dialogue, with the speaking style of a particular dialogues. Also, the dialogue made by one character detected in a scene directly affects the dialogue of another character present in the same. Hence, generating dialogues of such a kind are currently outside the scope of neural networks.

However, apart from dialogues, comic books consist of captions which can generally be perceived as the narrator's voice describing the chain of events. Both the captions and the images in the comic books should be in continuation, such that context of the captions from the previous images are maintained. Hence, the captions generated should be conceptual, carry a particular style as well as have context.

4.1 Automatic Image Caption Generator

Automatic writing of descriptions of images, with the proper language semantics has been a task that is challenging but widely pursued by researchers. Though state of the art results have been obtained in object recognition tasks [2013] from images, forming sentence descriptions with proper grammar and language semantics such that all the objects recognized in the image are related together, is a further more challenging task. Most attempts in these fields have taken place by combining two subtasks-object recognition and text generation gibe input words.

Vinyals et al,[2015] proposed a single joint model such that a text caption was generated, given an image as input with maximizing the probability of generating a correct sequence of words. The images were represented using a CNN, while an

LSTM based sentence generator was used to generate the captions.Xu et al [2016], further proposed a caption generation model which incorporated an attention mechanism, which solved the fixed length vector encoding problem of RNNs. Two types of attention mechanisms were used- a "soft" deterministic attention mechanism which was trained using back propagation learning and a hard "stochastic" attention mechanism which used REINFORCE algorithms.

4.2 Conceptual Captioning

However, most captions generated in the above methods were factual in nature, i.e., constituted a basic sentence describing the events of an image. Hence, a model was proposed by [2014], where the main focus of the model was to generate human like sentences which gave importance to the context of the image and not the objects in it. A model was proposed in [2014] for generating image description in which the description generated captures commonsense knowledge and language model trained is based on concept of maximum entropy. With MELM the words with maximum occurrence are included in every sentence.

Piyush Sharma, Nan Ding et al [2018], further introduced 'Conceptual Captions', a new dataset of image caption annotations. It contains more images as compared to MS-COCO dataset with a variety of image caption styles. The Alt-text descriptions are automatically processed into Conceptual Captions. Two models- an RNN and a Tensor2Tensor model were trained with Conceptual Captions dataset, and did a better task compared to when trained on MS COCO dataset. Conceptual Captions has an accuracy of ~90% according to ratings given by humans.

4.3 Stylized Caption Generation

The captions generated in a comic book mostly consist of a particular style or a mood of the story. The story may proceed to be factual narrative, where only the events in the image, i.e., the scene are described, or could be emotional, i.e., the character's internal dialogue or emotions could be expressed via the captions of the image. The captions may hence need to have a style such that it may be romantic, humorous, positive or negative. This is termed as stylized image captioning.

Existing image captioning methods describe the content of an image with a neutral or an objective caption without any style. This is known as a 'factual caption'. Tianlang Chen and Zhongping Zhang et al. in their paper [2018], have proposed a variation of the existing LSTM model known as the style-factual LSTM model. The caption generated using this model has a specific style such as (romantic, positive, humorous) and at the same time it also describes the semantic content of the image. This makes the caption more expressive and attractive.

4.4 Contextual Caption Generation

Capturing contexts over various images, such that caption for two images occur such that the second caption is continued based on the first one, is the most definitive requirement, for generating proper contextual captions such that a comic book may be generated. To capture textual context over image captioning, Zhou et al, proposed an end-to-end trainable text conditional attention model. It is based on a modified gLSTM network proposed by Jia et al[2015] such that it is time variant, known as td-gLSTM [2016]. Though using attention based mechanism in image captioning significantly improves the performance of image captioning models significantly, it only uses visual content and does not incorporate textual context while generating its output.

The conditioning of the image features is done by learning a text-conditional embedding matrix between the image features and the textual context. This text conditional attention enables the system to produce semantic guidance, such that which semantic feature is to be identified in the image based on the word input, such as "clothes" and "dishes", would be identified in the image if previously generated caption consists of the word washing. The CNN weights are fine-tuned with the text-conditional embedding learnt by the model, and hence both the image features and the text features are given as input to the td-gLSTM network.

4.5 Poetic Caption Generation

Even though approaches for context generation, like discussed above are available. The results are still not satisfactory enough such that they can generate text caption within the comics, since the contexts from all previous images and their respective captions can be derived to develop a storyline. Another type of comic books available are comics poetry. Here, the comixified images are accompanied by poetry in the form of a caption. Some examples of commercial comics poetry are *Drawn to Marvel: Poems from the Comic Books* by Bryan D. Dietrich, Marta Ferguson, *Poetry Comics from the Book of Hours* by Bianca Stone, *Krypton Nights: Poems* by Bryan D. Dietrich. Here, the poetry generated can be an anthology and hence context across various images need not be maintained.

5. METHOD

5.1 Experimental Setup

We have used Amazon EC2 instance (Web service) in order to deploy the neural style transfer as well as the generative text writing model. The AWS EC2 instance is c4.large. There are 8 virtual CPUs with every instance having 2.9GHz Intel-Xenon E5-2666 v3 Processor. 15 GiB (16.106 GB) memory is provided. The dedicated EBS bandwidth is 1000 Mbps.

5.2 Neural Style Transfer

Style Transfer has been attempted several times before like discussed in the methods above. However, our aim is to generate images that appear comicified and hence, can be used in the proposed comics poetry generation application.
Most models discussed above did not consider comicification of images as a parameter for their style transfer approach. ComixGAN approach uses generative adversarial networks which are pretrained on a dataset of images drawn by particular comics artists. However, this doesn't take into account the different color distribution, lighting, luminosity as well as brightness of content images. As a result, applying a pre-learned style on all types of content images may not produce desirable results.

We divide the comicification approach in two steps-

1) Applying choice styles to the images, which result in overall comicified output images.
2) Processing the provided content image, so that the edges of the image are preserved. Also, the content details in the image are enhanced.

The style transfer uses a pretrained VGG16 network. Content layer uses higher layers - conv4_2, conv5_2, while style layers use lower layers - conv1_1, conv_1', conv3_1, conv4_1, conv5_1. The network takes three input images- a content image, a style image and an initial image. Generally, the initial image is a plain white noise image. However, we use white noise blended with the content image, with a default ratio of 0.2. This speeds up the stylization process. Also, the layer normalization is used to reduce the time of processing.

Gram matrices are used to extract the texture required from the style image that has to be applied on the blended white noise image.

$$G_{ij}^{l} = \sum_k F_{ik}^{l} F_{jk}^{l}$$

The content and style losses are calculated using an L2 loss function between the activation features of the blended white noise image and the content image, style image respectively.

$$L_{content} = \frac{1}{2}\sum_{i,j}(F_{ij}^{l} - P_{ij}^{l})^2$$

$$L_{style} = \frac{1}{2}\sum_{l=0}^{L} w_l E_l$$

$$E_t = \frac{1}{4N^2M^2}\sum_{i,j}(G_{ij}^{l} - A_{ij}^{l})^2$$

The total loss is then minimized using Adam optimization technique, to perform gradient descent on the blended white noise image.

$$L_{total} = \alpha * L_{content} + \beta * L_{style}$$

The obtained output images were then processed using edge preserving techniques. As a result of which, the edges in the images are more pronounced, which is a required trait for comixification of the images. As a result of this, the contour lines of the content images are enhanced, with more prominent figures and objects in the image. The images are further enhanced using detail enhancement methods, which enhance color and vibrance of individual pixels in order to achieve more vibrant images, synonymous to those in modern digital comics. Also, overall contrast of the image is also increased in order to increase the luminance of the final stylized image. The final image is bright, luminous, detailed and vibrant with its edges enhanced.

5.3 Poetic Caption Generation

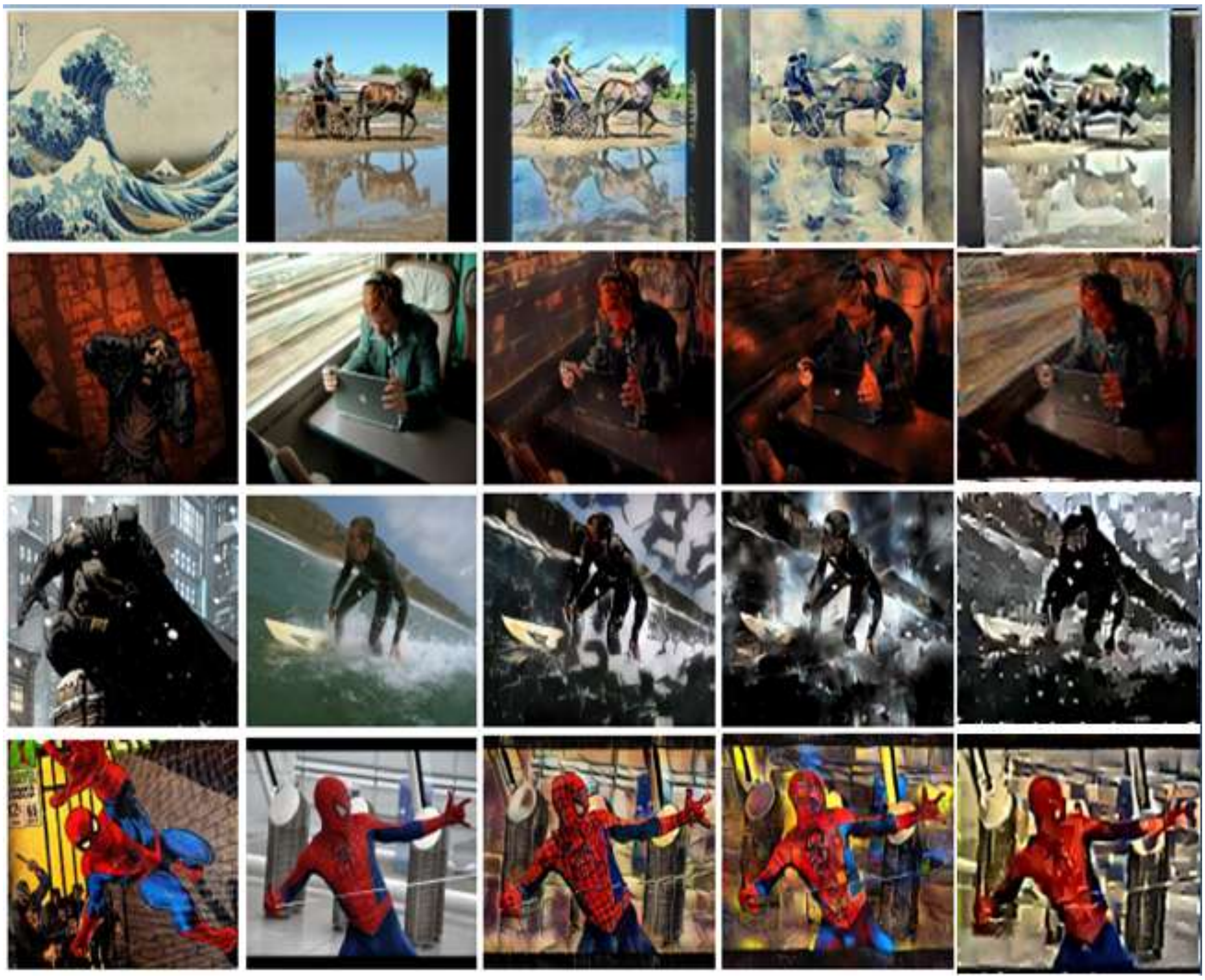
Since the context, style and concept of the textual captions are very important in the comic books' captions, poetry is a good artistic form that can be used. Here, context across the same images are maintained; however, the lack of context across images can be compensated by creating an anthology of comics poetry in the proposed comics poetry generation approach.

The approach uses Multi-Adversarial Training [2018] for Generative Adversarial Networks for the generation of poetry using non-hierarchical RNN networks, along with 3 CNNs for object, scene and sentiment detection. The network is trained on image-poem pairs from Multi-Modal dataset to generate relevant poems with respect to images as well as poems scarped from the web in Uni-modal poem dataset, to obtain good semantic as well as capture poeticness in the captions.

The obtained poetry is then clubbed with the stylized image it is generated from, which can be used to generate a comics poetry anthology created entirely by neural networks.

6. RESULTS

Our model performs better on the comixification task of images compared to the existing approaches discussed above. The figure below compares results of various style transfer techniques, used by various approaches discussed above. As we can see, the proposed model produces better results pertaining to the comixification tasks, i.e., the stylized images produced in our model's case look most like digital comic strip images, since the edges are more pronounced as well as the detail enhancement, vibrancy and luminance is more.



a)Style b)Content c) ADA-In d) UST-WC e) Proposed Model

Fig 2: Comparison between ADA-In, UST-WC and proposed model

Comparing with Gatys' original model, the proposed model works really faster. For the content and style images in fig2, the time taken to by Gatys' model for a 1000 iterations is 80 minutes, while the time taken by our approach is 14 minutes for 1000 iterations on AWS c4.2xlarge with 8 vCPU with each instance having 2.9 GHz Processor.

a) Content image b) Style image c) Gatys' output d) Proposed output

Fig 1: Comparison between Gatys and proposed model

The figure below shows the generated poetry based on the content image clubbed with the comics stylized images, which produces an aesthetic looking comics poetry strip entirely authored by neural networks.

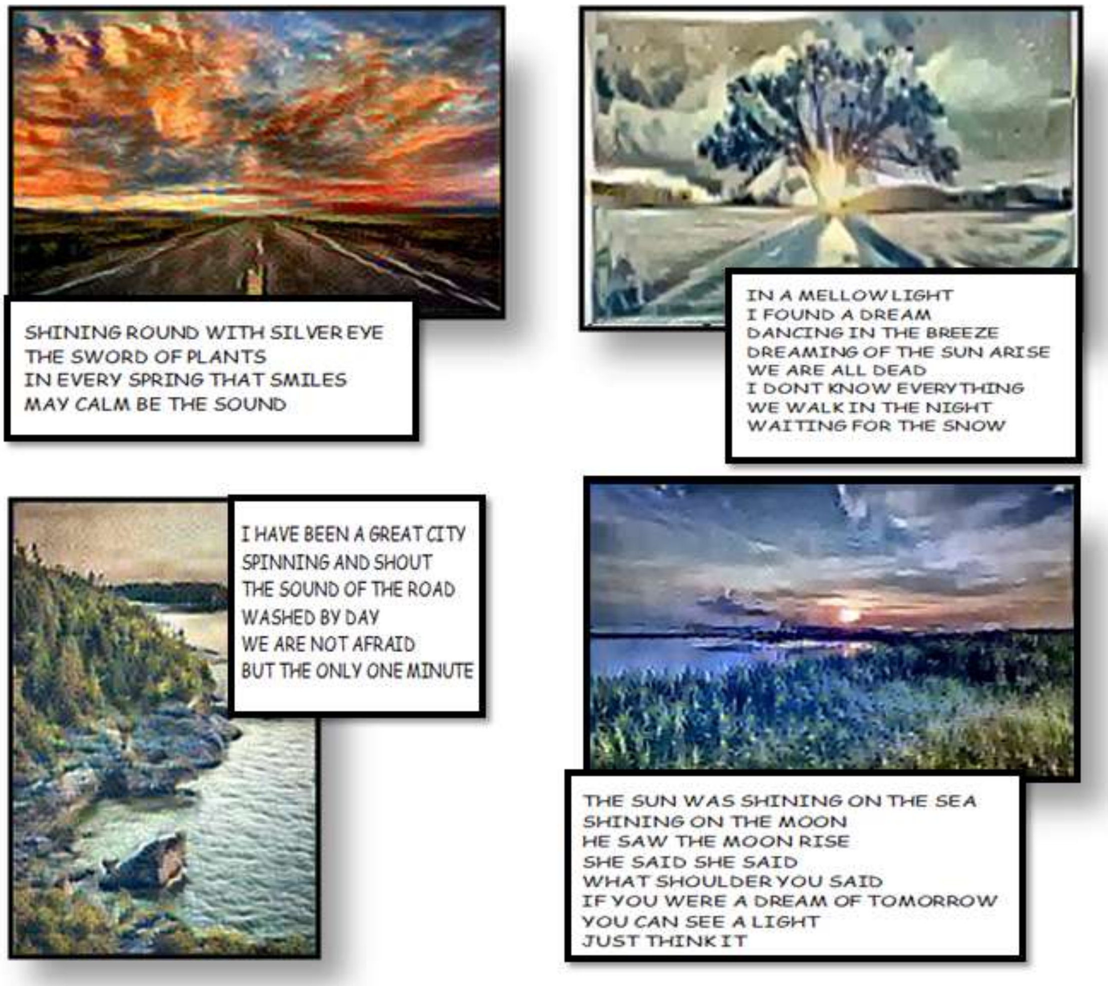


Fig.3 Generated comics poetry strip sample

7. CONCLUSIONS

Comixification of images is a task which is fairly new but shows great promise in future commercialization such that it can revolutionize the digital comic publication process today. The approach discussed in this paper combines this comixification task with poetic captions, which can help generate comic poetry books authored solely by neural networks. It modifies previouslyproposed style transfer techniques so as it to generate images in the form of a comics. The model not only generates better results in terms of more visually appealing comics image but also reduces the time taken by Gatys' original neural style transfer algorithm significantly.

Image captioning in a poetic style is another task which has been explored recently. Ensuring the relevance of the poetry to the image,along with maintaining the style of writing as poetic is a rather challenging task. A network consisting of a generator and discriminator network with adversarial training helps improve the results of produced poetry, such that the poetic style, word context, and high semantic value is preserved. Furthermore, various image captioning techniques are also studied which can be used in the future to produce conceptual, stylized captions which can hold contexts over long intervals such that a proper storyline across the comics can be maintained.

REFERENCES:

1. Leon A. Gatys, Alexander S. Ecker and Matthias Bethge (2015), A Neural Algorithm of Artistic Style, *Computer Vision and Pattern Recognition*.

2. Jing Liao, Yuan Yao, Lu Yuan, Gang Hua and Sing Bing Kang (2016), Visual Attribute Transfer through Deep Image Analogy, *Computer Vision and Pattern Recognition.*

3. Maciej Pęśko and TomaszTrzciński (2018), Neural comic style transfer: Case Study, *arxiv:1809.01726[cs.cv]*.

4. Xun Huang and Serge Belongie (2017), Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization, *Computer Vision and Pattern Recognition.*

5. Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang etal (2017), Universal Style Transfer via Feature Transforms*, 31st Conference on Neural Information Processing Systems*.

6. Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, Jan Kautz, (2018), A Closed-form Solution to Photorealistic Image Stylization, *arXiv:1802.06474v5 [cs.CV]*.

7. Yangchen, Yu-kunlai and Yong-jinliu (2017), Transforming photos to comics using convolutional neural networks, *IEEE International Conference on Image Processing.*

8. Yang chen ,Yu-kunlai and Yong-jinliu (2018) , CartoonGAN: Generative Adversarial Networks for photo cartoonization, *CVPR.*

9. MaciejPesko, Adam Svystun et al. (2018), Comixify: transform video into a comics, *arxiv:1812.03473v1 [cs.cv]* .

10. Justin Johnson, Alexandre Alahi, Li Fei-Fei (2016), Perceptual Losses for Real-Time Style Transfer and Super-Resolution*, CVPR*.

11. Elliott, Desmond and Keller, Frank (2013). Image description using visual dependency representations, *EMNLP*.

12. OriolVinyals, Alexander Toshev, Samy Bengio and Dumitru Erhan (2015), Show and Tell: A Neural Image Caption Generator*. Computer Vision and Pattern Recognition*.

13. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho et al, Show (2016), Attend and Tell: Neural Image Caption Generation Visual Attention, *International Conference on Machine Learning.*

14. Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh Srivastavaetal (2014), From Captions to Visual Concepts and Back, *Computer Vision and Pattern Recognition.*

15. Piyush Sharma, Nan Ding, Sebastian Goodman, RaduSoricut (2018), Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset for Automatic Image Captioning, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics.*

16. Tianlang Chen, Zhongping Zhang, Quanzeng You, Chen Fangetal (2018), "Factual" or "Emotional": Stylized Image Captioning with Adaptive Learning and Attention, *Computer Vision and Pattern Recognition.*

17. Luowei Zhou, ChenliangXu, Parker Koch, Jason J. Corso (2016), Watch What You Just Said: Image Captioning with Text-Conditional Attention, *Computer Vision and Pattern Recognition.*

18. Bei Liu, Jianlong Fu, Makoto P. Kato, Masatoshi Yoshikawa (2018). Beyond Narrative Description: Generating Poetry from Images by Multi-Adversarial Training, *Computer Vision and Pattern Recognition*

19. Gorach Tasneem, Choksi Hetvi, Kriplani Mahima, Das Smriti and Inje Bhushan(2019), A Review on Neural Style Transfer with auto text Generation, *In Proceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM)*

20. XuJia, Efstratios Gavves, Basura Fernando& Tinne Tuytelaars (2015), Guiding Long-Short Term Memory for Image Caption Generation, *Computer Vision and Pattern Recognition.*