UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Sequence Modelling with Deep Learning for Visual Content Generation and Understanding

by Zongxin Yang

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Prof. Yi Yang

University of Technology Sydney
Faculty of Engineering and Information Technology

Jul 2021

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, *Zongxin Yang* declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:
Signature: Signature removed prior to publication.

Date: 16th Jul, 2021

Abstract

Although convolutional neural networks have proven to be effective and stable in image feature learning, sequence modelling is still critical for learning spatial and temporal context information. In an image scenario, different semantic structures can be regarded as a sequence arranged along the horizontal (or vertical) direction. Moreover, in a video scenario, temporal sequence modelling is necessary for understanding inter-frame relationships, such as object movement and occlusion. This thesis explores more effective spatial or temporal sequence modelling for image or video scenario understanding. For the former, an encoder-decoder framework is proposed to split an input scenario into a sequence of spatial features and reconstruct the input. By modelling spatial sequence information, the framework can even predict new scenes with very large scales in length while keeping a consistent style regarding the given input. For video understanding, the thesis processes temporal sequences in a recurrent manner (*i.e.*, frame by frame), which is more memory-efficient. In addition, the thesis proposes to implicitly impose the feature embedding of each target and relative background to be contrastive throughout the temporal sequence, promoting the results of downstream tasks accordingly. Besides, a novel transformation module is designed to model channel relationships for improving intra-frame representation ability. To validate proposed approaches and components, extensive experiments are conducted on image outpainting, instance segmentation, object detection, classification, video classification, and video object segmentation.

Acknowledgments

Firstly, I would like to thank my supervisor Professor Yi Yang, without whom I would not have been able to complete this thesis. I sincerely appreciate him for his selfless support and patience. He guided me throughout the research by providing tremendous help on his knowledge and insight into the related fields.

I would also like to thank my colleagues at University of Technology Sydney. I would like to thank Yunchao Wei, Linchao Zhu, Xin Yu, Ping Liu, Yu Wu, Qianyu Feng, Peike Li, Chen Liang, Jiaxu Miao, Yuhang Ding, and many others. I was really fortunate to work with them and participate in intellectual conversations with them.

Lastly, I would like to thanks my mother, Gexiu Liao, and my father, Jianlin Yang, for their support and love throughout the years.

LIST OF PUBLICATIONS

1. **Z. YANG**, J. DONG, P. LIU, Y. YANG, AND S. YAN, *Very long natural scenery imageprediction by outpainting*, in ICCV, 2019, pp. 10561-10570.

2. **Z. YANG**, L. ZHU, Y. WU,AND Y. YANG, *Gated channel transformation for visual-recognition*, in CVPR, 2020, pp. 11794-11803.

3. **Z. YANG**, Y. WEI, AND Y. YANG, *Collaborative video object segmentation by foreground-background integration*, in ECCV, Springer, 2020, pp. 332-348.

Table of Contents

List of Publications **ix**

List of Figures **xii**

List of Tables **xvi**

1 Introduction **1**

1.1 Visual Feature Learning 1

1.2 Sequence Modeling for Image Generation 2

1.3 Sequence Modeling for Video Understanding 3

1.4 Contributions 4

2 Literature Review **5**

2.1 CNN-based Visual Feature Learning 5

2.1.1 Image Feature Learning 5

2.1.2 Video Feature Learning 5

2.2 Image Generation 6

2.2.1 RNN-based Image-to-Image Translation 6

2.2.2 Generative Adversarial Networks 6

2.2.3 Image Inpainting 7

2.2.4 Image Outpainting 7

2.2.5 Image-to-Image Translation 7

2.3 Video Understanding 8

2.3.1 Video Classification 8

2.3.2 Semi-supervised Video Object Segmentation 8

3 Natural Scenery Image Generation by Spatial Sequence Modeling **11**

3.1 Introduction 11

3.2 The Proposed Approach 13

3.2.1 Encoder-Decoder Generator 14
3.2.2 Loss Function 18
3.3 Results 20
3.3.1 One-step Generation 20
3.3.2 Multi-Step Prediction 22
3.4 Conclusion 24

4 Collaborative Foreground-Background Integration for Video Understanding 27
4.1 Introduction 27
4.2 The Proposed Approach 29
4.2.1 Pixel-level Matching 31
4.2.2 Instance-level Attention 33
4.2.3 Collaborative Ensembler (CE) 34
4.2.4 Gated Channel Transformation (GCT) 34
4.2.5 Implementation Details 35
4.3 Results 38
4.3.1 Comparison Results 39
4.3.2 Ablation Study 43
4.4 Conclusion 45

5 Future Works 47

Bibliography 49

List of Figures

Figure **Page**

3.1 An illustration of one-step image outpainting, which predicts unseen contents outside the input region with the identical scale of the input. The semantic information and spatial layout must keep harmonious and consistent between predicted contents and the input scenario. 12

3.2 An illustration of multi-step image outpainting in horizontal direction for natural scenarios. 12

3.3 (a) An overview of our outpainting framework, which contains a discriminator and a generator. (b) An illustration of multi-step image outpainting based on our framework. In this iterative manner, we can efficiently predict unseen contents with a very large scale in length. 14

3.4 An illustration of Recurrent Contextual Transfer (RCT). The input feature maps are split into a sequence along a spatial dimension (horizontal dimension in our setting). Two layers of LSTM [32] are used to iteratively transfer the spatial sequence from the input region to the predicted region. 16

3.5 (a) The structure of Skip Horizontal Connection (SHC). (b) The details of Global Residual Block (GRB). N, and r, and ks denote channel number, dilation rate, and kernel size of convolutional layers separately. 16

3.6 (a): Removing all the SHC connections, the predicted results have a clear faultage on the junction line between the given scene and generated scene. (b) Using only a single SHC connection in the decoder's second stage, the junction boundary becomes smoother and move a bit to the predicted region. (c) Applying all the SHC connections as our default setting, the predicted results have an excellently smooth junction line. 17

3.7 Comparison results on the proposed dataset. (d) The ground-truth image scenes. (c) The predictions have clear faultage on the junction line between the given scene and generated scene using a fully-connected layer to transfer information from the encoder to the decoder. (b) Applying SHC layers can mitigate the problem of faultage. However, the predicted contents are still blurred where the prediction position is far from the junction line, as highlighted in yellow boxes. (a) We improve the prediction quality in faraway regions by replacing the fully-connected layer with our proposed RCT, which has a better performance in modelling spatial sequence. 18

3.8 Some scenes sampled from the natural scenery dataset. The dataset contains about 6,000 complicated and diverse natural scenes, including starry sky, riverbank, seaside, valley, snow mountain, etc. 19

3.9 Comparing our framework with state-of-the-art image generation works on the 1-step generation, ours SHC+RCT framework achieves the best generative quality. 21

3.10 We can generate outpainting images with a huge scale in length. Given an image scene with a size of 128×128, our method iteratively generates a 16 times larger scene in the horizontal direction from left to right, leading to a large prediction with a size of 2176 pixels in length. Since the predictions are too long, we divide every prediction into two halves. 23

3.11 We generate outpainting images in both the horizontal directions, left and right. Given an input image, our method iteratively generates a 4 times larger scene for each horizontal direction, leading to a large prediction with a size of 1152 pixels in length. 23

3.12 Comparison results on multi-step generations. 24

3.13 A snow mountain prediction starting from a scene with rarely observable contents. Such a case is extremely difficult for traditional non-deep-learning works. 24

4.1 CI denotes Collaborative Integration. The dot-line arrow denotes the proposed background matching. In the given video sequence, we have two foreground objects, *i.e.*, pink sheep and blue sheep. Without background matching, the blue sheep's prediction is confused by one similar sheep in the background region, as shown in the top line. By adding the background matching, the confusion problem is relieved, as shown in the bottom line. 28

4.2 CFBI **overview**. F-G, red color, and blue color denote Foreground-Background, foreground region, background region, respectively. The confidence of the matching map with a deeper colour will be higher. Given the current frame ($t = T$), the reference frame with mask ($t = 1$), and the previous frame with mask ($t = T - 1$), CFBI will model their temporal dependencies and predict the target segmentation of the current frame. First, we use a backbone to extract pixel-level embeddings for all three frames. The pixel-level embeddings of $t = T - 1$ and $t = 1$ will be separated into foreground embeddings and background embeddings regarding the given masks. Then, the instance-level attention and pixel-level matching will be applied to both the foreground embeddings and background embeddings to model temporal sequence information and predict the current frame segmentation. During inference, given the first frame mask, we can predict target masks from the reference frame to the final one by iterating such a prediction process. 30

4.3 In video understanding, there are diverse movement speeds, from nearly static to fast, of foreground objects. The demonstration video sequences are sampled from YouTube-VOS [81]. 32

4.4 The detailed structure of the instance-level attention module. C, W, and H mean channel dimension, width, and height, separately. C_e refers to the channel number of pixel-level embedding extracted from the backbone. 33

4.5 The detailed architecture of Gated Channel Transformation. 35

4.6 By using normalization operation, GCT is able to explicitly model some channel relations, like cooperation (decreasing channel variance) and competition (increasing channel variance). 35

4.7 (a) If we use traditional random cropping, some cropping boxes (red boxes as shown) will contain little or even no object region. (b) By using the proposed balanced-crop, all the cropping boxes contain enough foreground pixels. . . . 36

4.8 On the validation split of DAVIS 2017, CFBI is compared with a state-of-the-art VOS method, STMVOS. STMVOS confuses a part of the person and the bicycle in the second video. Besides, CFBI is successful in segmenting the foreground weapon after blur and occlusion in the top sequence, while STMVOS fails. 37

4.9 On YouTube-VOS and DAVIS 2017, we show some qualitative results of CFBI. On YouTube-VOS, CFBI succeeds in matching and segmenting all the desired sheep with a similar appearance from a flock of sheep. On DAVIS 2017, CFBI performs perfectly after the occlusion on the dogs and the person. However, when two objects have very similar textures and are very close, CFBI is still confused by a part of the object, as shown in the bottom sequence. 39

LIST OF TABLES

TABLE **Page**

3.1 The hyper-parameter setting of our encoder-decoder generator. 15

3.2 Ablation study of different GRB number. 1 denotes we remove all the GRB blocks except the one where the feature dimensions are $16 \times 32 \times 256$. 21

3.3 Comparison results in FID [31] and IS [57] metrics. In the testing split of proposed natural scenary dataset, the images have an average IS score of 3.387. 22

4.1 Compare CFBI with existing frameworks on the validation 2018, validation 2019, and testing 2019 splits of YouTube-VOS [81]. ^{MS} denotes that the flip & multi-scale augmentation is used in the evaluation. *, SD, and FT denote model ensemble, simulated training data, and online fine-tuning, respectively. 38

4.2 Compare CFBI with existing frameworks on the validation split of DAVIS 2016 [52]. (**YT**) denotes both the training splits of YouTube-VOS and DAVIS 2017 are used in the training process. 40

4.3 Compare CFBI with existing frameworks on the validation split and the testing split of DAVIS-2017 [53]. 41

4.4 Apply GCT to modern CNNs for evaluating the error performance (%) in image classification on ImageNet [56]. G/P denotes GFLOPS/Parameters. . . 42

4.5 Apply GCT to Mask R-CNN [28] in instance segmentation (mask AP) and object detection (box AP) on COCO [45]. 42

4.6 Apply GCT to NL [74] in action recognition, using top-1 accuracy (%), on Kinetics. 43

4.7 We ablate different background information from CFBI in these experiments. *, IL, PL denote removing the foreground bias and background bias, the instance-level attention, and pixel-level matching, respectively. 43

4.8 Ablation of other components. 44