Image Style Transfer for Distillation of Diffusion Knowledge into a Transformer

A Preprint

Egor Y. Silvestrov
Faculty of Computational Mathematics and Cybernetics
Lomonosov Moscow State University
Moscow, Russia
s02210546@gse.cs.msu.ru

Victor V. Kitov
Faculty of Computational Mathematics and Cybernetics
Lomonosov Moscow State University
Moscow, Russia
v.v.kitov@yandex.ru

Abstract

Modern methods of style transfer for weak models often face problems with the quality of style transfer, especially in conditions of limited computing resources and untagged data (unsupervised learning). Distilling knowledge through diffusion models is a promising approach to improve the quality of weak models by transferring key elements of knowledge from more powerful models by creating a marked-up dataset and turning an unsupervised task into a supervised task with a teacher. In this paper, we investigate the method of distilling knowledge for a diffusion model, which allows us to adapt the styling and transmission of content for a more lightweight model (based on the transformer architecture) without the need for significant computational costs. As a result, an optimal balance is achieved between maintaining high-quality visual characteristics and cost-effectiveness, which opens up new opportunities for developing effective stylization models in real time.

Keywords Image Style Transfer · Diffusion Model · Knowledge Distillation · Transformer

1 Introduction

Image style transfer has gained significant attention in the creation of artistic visuals. The task involves taking a content image and a style reference image to produce an output that retains core content elements while adopting the visual style of the reference. This technique has applications across various domains, such as clothing design [6], photo and video editing [7, 8], virtual reality [9], and more. Recently, deep neural networks have been widely used for style transfer, which can be grouped into three main approaches: 1) optimization-based methods, 2) feedforward approximation, and 3) zero-shot style transfer. Gates et al. [10] proposed optimizing pixel values in a content image by minimizing both feature reconstruction and style losses, producing impressive results but requiring multiple iterations for each content-style pair, making it computationally expensive. In response, feedforward networks [11, 12, 13] were developed to directly learn mappings from photographs to stylized images in specific painting styles, although retraining is required for new styles. Zero-shot style transfer is more versatile, as it can handle diverse styles, even previously unseen ones. Huang et al. [14] introduced an arbitrary style transfer approach using adaptive instance normalization (AdaIN), which normalizes content image features and adjusts them based on style parameters. Recent work replaces AdaIN with whitening and coloring transformations [15], while several studies further refine this approach [16, 17].

However, a common limitation of these methods is that merely adjusting feature statistics makes it difficult to synthesize complex style patterns rich in detail and local structures, often resulting in distorted and less recognizable images.

For instance, methods by Gatys et al. [10], AdaIN [14], and WCT [15] frequently introduce style distortions that blur original content details. To address this, Deng et al. developed StyTr2 [18], which uses attention to capture semantic correlations between content and style features, yielding visually appealing results. Nevertheless, StyTr2 also suffers from structure distortion due to its shallow feature extractor, which lacks pre-trained weights, limiting its ability to differentiate between foreground and background objects. Thus, achieving a representation that can maintain content structure while accurately capturing fine-grained style patterns remains a challenging problem.

Diffusion models [1, 2, 3, 4] have also achieved remarkable success in style transfer, excelling at generating visually coherent and detailed stylizations. In this paper, we propose using STTR [5], a Transformer-based model, as a student model to distill knowledge from a larger diffusion model [4]. This diffusion model was taken based on good experimental results and the existence of an implementation. In this setup, the diffusion model [4] performs style transfer on images, and STTR [5] is trained to replicate these stylized outputs, effectively reframing style transfer as a supervised learning task. Transformer-based architectures, popularized by advancements in natural language processing [19], have demonstrated effectiveness in vision tasks by modeling long-range dependencies. The STTR [5] approach uses to decompose content and style images into visual tokens, enabling learning of the global context between them. As similar content tokens align with the corresponding style tokens, this approach achieves detailed style transformation with structural consistency between content and style.

We believe that training a small, relatively diffusive transformer model will allow us to achieve the quality of large diffusive models while using much fewer resources, which allows us to run this model on various low-power devices.

2 Headings: first level

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula. See Section 2.

2.1 Headings: second level

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

$$\xi_{ij}(t) = P(x_t = i, x_{t+1} = j | y, v, w; \theta) = \frac{\alpha_i(t) a_{ij}^{w_t} \beta_j(t+1) b_j^{v_{t+1}}(y_{t+1})}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i(t) a_{ij}^{w_t} \beta_j(t+1) b_j^{v_{t+1}}(y_{t+1})} \tag{1}$$

2.1.1 Headings: third level

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Paragraph Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Table 1: Sample table title

Part		
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

3 Examples of citations, figures, tables, references

3.1 Citations

Citations use `natbib`. The documentation may be found at

`http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf`

Here is an example usage of the two main commands (`citet` and `citep`): Some people thought a thing kour2014real, hadash2018estimate but other people thought something else kour2014fast. Many people have speculated that if we knew exactly why kour2014fast thought this...

3.2 Figures

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi. See Figure **??**. Here is how you add footnotes. [1] Sed feugiat. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut pellentesque augue sed urna. Vestibulum diam eros, fringilla et, consectetuer eu, nonummy id, sapien. Nullam at lectus. In sagittis ultrices mauris. Curabitur malesuada erat sit amet massa. Fusce blandit. Aliquam erat volutpat. Aliquam euismod. Aenean vel lectus. Nunc imperdiet justo nec dolor.

3.3 Tables

See awesome Table 1.

The documentation for `booktabs` ('Publication quality tables in LaTeX') is available from:

`https://www.ctan.org/pkg/booktabs`

3.4 Lists

- Lorem ipsum dolor sit amet
- consectetur adipiscing elit.
- Aliquam dignissim blandit est, in dictum tortor gravida eget. In ac rutrum magna.

References

[1] Z. Wang, L. Zhao and W. Xing, "StyleDiffusion: Controllable Disentangled Style Transfer via Diffusion Models", College of Computer Science and Technology, Zhejiang University, 2023.

[2] Y. Zhang, N. Huang, F. Tang, H. Huang, C. Ma, W. Dong, C. Xu, "Inversion-based Style Transfer with Diffusion Models", MAIS, Institute of Automation, Chinese Academy of Sciences, Institute of Computing Technology, Chinese Academy of Sciences, School of AI, UCAS, Kuaishou Technology, 2023.

[1] Sample of the first footnote.

[3] S. Yang, H. Hwang, J. Chul Ye, “Zero-Shot Contrastive Loss for Text-Guided Diffusion Image Style Transfer”, Kim Jaechul Graduate School of AI, Korea Advanced Institute of Science and Technology (KAIST), 2023.

[4] J. Chung, S. Hyun, J. Heo, “Style Injection in Diffusion: A Training-free Approach for Adapting Large-scale Diffusion Models for Style Transfer”, Sungkyunkwan University, 2024.

[5] J. Wang, H. Yang, J. Fu, T. Yamasaki and B. Guo, “Fine-Grained Image Style Transfer with Visual Transformers”, The Univerisity of Tokyo, Microsoft Research, 2022.

[6] P. Date, A. Ganesan, T. Oates, “Fashioning with Networks: Neural Style Transfer to Design Clothes”, University Of Maryland, 2017.

[7] D. Chen, J. Liao, L. Yuan, N. Yu and G. Hua, “Coherent Online Video Style Transfer”, University Of Maryland, 2017.

[8] W. Zhang, C. Cao, S. Chen, J. Liu, “Style Transfer Via Image Component Analysis”, 2013.

[9] C. Castillo, S. De, X. Han, B. Singh, A. K. Yadav, and T. Goldstein, “Son of Zorn’s Lemma: Targeted style transfer using instance-aware semantic segmentation”, Department of Computer Science, University of Maryland, College Park, 2017.

[10] Leon A. Gatys, Alexander S. Ecker, Matthias Bethge, “Image Style Transfer Using Convolutional Neural Networks”, Centre for Integrative Neuroscience, University of Tubingen, 2016.

[11] J. Johnson, A. Alahi, L. Fei-Fei, “Perceptual Losses for Real-Time Style Transfer and Super-Resolution”, Department of Computer Science, Stanford University, 2016.

[12] C. Li and M. Wand, “Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks”, Institut for Informatik, University of Mainz, 2016.

[13] D. Ulyanov, V. Lebedev, A. Vedaldi, V. Lempitsky, “Texture Networks: Feed-forward Synthesis of Textures and Stylized Images”, Skolkovo Institute of Science and Technology & Yandex, 2016.

[14] X. Huang, S. Belongie, “Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization”, Department of Computer Science & Cornell Tech, Cornell University, 2017.

[15] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, M. Yang, “Universal Style Transfer via Feature Transforms”, Adobe Research, UC Merced, NVIDIA Research, 2017.

[16] V. Kitov, C. Fang, J. Yang, Z. Wang, X. Lu, M. Yang, “Depth-Aware Arbitrary style transfer using instance normalization”, Lomonosov Moscow State University, 2020.

[17] Z. Hu, J. Jia,B. Liu, Y. Bu, J. Fu, “Aesthetic-Aware Image Style Transfer”, China Key Laboratory of Pervasive Computing, Ministry of Education Beijing National Research Center for Information Science and Technology, Microsoft Research, 2020.

[18] Y. Deng, F. Tang, W. Dong, C. Ma, X. Pan, L. Wang, C. Xu, “StyTr2:Image Style Transfer with Transformers”, 2022.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, “Attention Is All You Need”, Google Research, Google Brain, University of Toronto, 2017.