TRANSFERRING FORGEDIT TO IMAGIC

Shiwen Zhang

witcherofresearch@gmail.com

ABSTRACT

In this supplementary letter, we transfer techniques of our Forgedit to Imagic. Keeping the multi-stage fine-tuning process of Imagic, we change everthing else to our Forgedit settings. We transfer our forgetting mechanism with disentangled UNet to Imagic in order to tackle its overfitting problem during editing stage. We found that our novel techniques improves Imagic, which demonstrates that our findings in Forgedit are universal.

Imagic (Kawar et al., 2023) with Imagen Imagen(Saharia et al., 2022), the previous SOTA text-guided image editing method, is slow and prone to overfitting. The recent SOTA text-guided image editing method, Forgedit (Zhang et al., 2023; Zhang, 2024a) with implementation open-sourced (Zhang, 2024b), built with Stable Diffusion 1.4 (Rombach et al., 2022) and BLIP (Li et al., 2022), speeds up Imagic SD (Kawar et al., 2023) significantly by 14 times and completely solves the overfitting problem of Imagic via disentangled UNet with forgetting strategies.Forgedit is the new SOAin terms of CLIP score (Hessel et al., 2021), LPIPS scoreZhang et al. (2018) and FID score Heusel et al. (2017) on TEdBench (Kawar et al., 2023). This forgetting mechanism replaces several blocks of UNet based on the disentangled law: UNet encoder learns structures, UNet decoder learns textures. We further explore the properties of Forgedit in a series of supplementary letters (Zhang, 2024a;b;c;d).

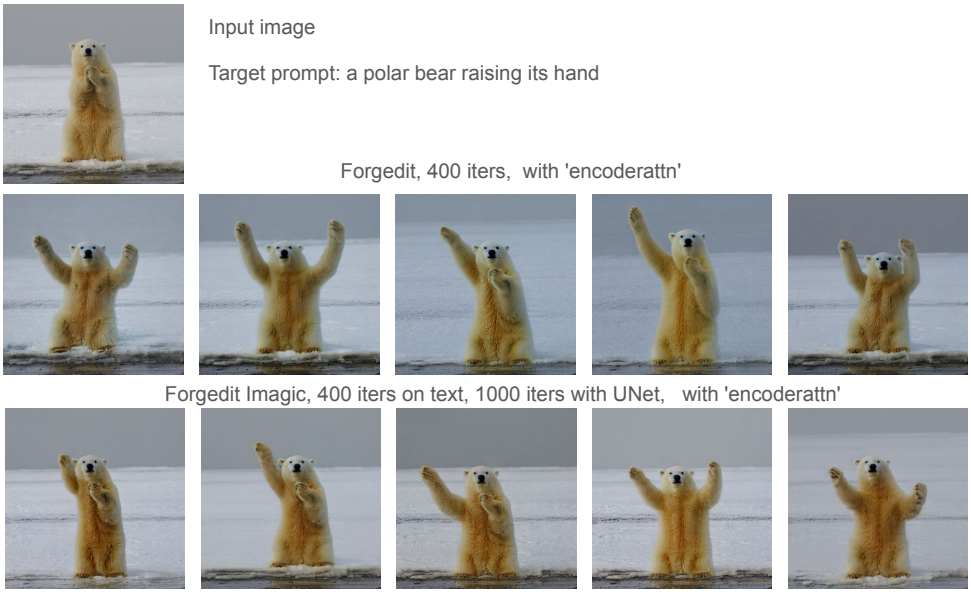

Figure 1: We compare the editing effects of Forgedit and Forgedit transferred to Imagic.

Discriminative deep neural networks in image and video classification (He et al., 2016; Zhang et al., 2020a;b; Zhang, 2022; Huang et al., 2020; Wang et al., 2016; 2021) also explore whether different stages of the network should be trained jointly or in multi-stages. Our Forgedit utilizes unified vision-language optimization while the fine-tuning stage of Imagic is seperated. Imagic first optimize text embedding, then optimize UNet. In this letter, like Forgedit, we use BLIP (Li et al., 2022) generated caption to be the source prompt for Imagic. Instead of optimizing the entire UNet, we

transfer the settings of Forgedit UNet for training. We optimzie the source text embedding for 400 iters and UNet for 1000 iters, compared to Forgedit with vision-language joint training for 400 iters. We found that such settings, again, causes overfitting in some cases. So the default forgetting strategies are applied. Shown in Figure 1, following the settings in the paper of Imagic+SD (Kawar et al., 2023), the UNet of Imagic is trained with more iterations thus the editing results seem to be kind of better than Forgedit+SD. The forgetting strategy is the default one on UNet encoder, 'encoderattn', since the target prompt is to modify the action of the polar bear. With one a100 GPU, the optimization of Forgedit Imagic on text embedding costs 29 seconds, UNet costs 1 minutes and 26 seconds, which is almost 4x the cost of Forgedit joint vision-language optimization.

REFERENCES

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP (1)*, pp. 7514–7528. Association for Computational Linguistics, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, pp. 6626–6637, 2017.

Weilin Huang, Shiwen Zhang, Sheng Guo, Limin Wang, and Matthew Robert Scott. 4d convolutional neural networks for video recognition, July 14 2020. US Patent 10,713,493.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, pp. 6007–6017. IEEE, 2023.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022.

Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022.

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pp. 20–36. Springer, 2016.

Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. Tdn: Temporal difference networks for efficient action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1895–1904, 2021.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Shiwen Zhang. Tfcnet: Temporal fully connected networks for static unbiased temporal reasoning. *arXiv preprint arXiv:2203.05928*, 2022.

Shiwen Zhang. Pytorch implementation of forgedit. In *Github Sources 2024*. https://github.com/witcherofresearch/Forgedit, 2024b.

Shiwen Zhang. On the importance of source text embedding in text-guided image editing. In *Researchgate*, 2024b.

Shiwen Zhang. Fast Imagic: Solving Overfitting in Text-guided Image Editing via Disentangled UNet with Forgetting Mechanism and Unified Vision-Language Optimization. In *Neurips Workshop UniReps*, 2024a.

Shiwen Zhang. Hyper-parameter tuning for text guided image editing. *arXiv preprint arXiv:2407.21703*, 2024a.

Shiwen Zhang. On the coefficient of model merging in forgetting mechanism. In *Researchgate*, 2024c.

Shiwen Zhang. Vector projection in text embedding space for text-guided image editing. In *Researchgate*, 2024d.

Shiwen Zhang, Sheng Guo, Weilin Huang, Matthew R Scott, and Limin Wang. V4d: 4d convolutional neural networks for video-level representation learning. In *International Conference on Learning Representations*, 2020b.

Shiwen Zhang, Sheng Guo, Limin Wang, Weilin Huang, and Matthew Scott. Knowledge integration networks for action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020a.

Shiwen Zhang, Shuai Xiao, and Weilin Huang. Forgedit: Text guided image editing via learning and forgetting. *arXiv preprint arXiv:2309.10556*, 2023.