Learning Temporal Relations for Evaluating Instruction-guided Image Editing

Thema:
Learning Temporal Relations for Evaluating Instruction-guided Image Editing
Art:
BA
BetreuerIn:
Udo Kruschwitz, Alexander Tack (extern)
BearbeiterIn:
Pia Donabauer
ErstgutachterIn:
Udo Kruschwitz
ZweitgutachterIn:
Christian Wolff
Status:
abgeschlossen
Stichworte:
Instruction-guided Image Editing, Evaluation, Human Alignment, Temporal Video Understanding
angelegt:
2025-01-27
Anmeldung:
2025-01-23
Antrittsvortrag:
2025-01-20
Abgabe:
2025-03-26

Hintergrund

Lately, vision-language models have enabled instruction-based image editing, where specific image regions are manipulated through textual prompts [1]. Evaluating these models is critical to select the best editing model and identify the most accurate edit among multiple outputs. Current evaluation approaches include human evaluation, automated metrics, and LLM-as-a-judge methods: Human evaluation is the gold standard but also costly, time-intensive, and non-scalable [2]. Automated metrics are scalable and efficient but primarily focus on image quality rather than edit quality [3], often correlating poorly with human judgment and being biased toward visual appearance [4]. LLM-based evaluation shows promising alignment with humans, however, can exhibit biases and hallucinations [5]. Existing evaluation frameworks are built on CLIP [6], a vision-language model that compares embeddings of images and textual descriptions. However, CLIP-based evaluations often depend on textual descriptions of both the original and edited images [1, 4], which are not always available.

Zielsetzung der Arbeit

The goal of this work is to develop an improved metric that better aligns with human judgment, enabling fair model comparisons and reliable validation of outputs in real-world applications. Inspired by advancements in adapting vision-language models such as CLIP for video understanding [7], the thesis aims to learn temporal relations of image edits and create a scalable, human-aligned metric for evaluating instruction-based image editing. This is achieved by fine-tuning a CLIP model modified for video understanding on an instruction-guided image-editing dataset, which is transformed into a video format. Various modifications and configurations are explored to identify the most effective approach.

Konkrete Aufgaben

- Literature Review: Related work, existing methods, research gap.
- Preliminary Studies: Explore naive approaches and justify new method.
  - Implementation: Fine-tune a CLIP-based model for video understanding to capture temporal cues.
- Evaluation: o Develop and implement a Gradio-based platform for data labeling. o Conduct the labeling process and process the collected labels. o Use human judgments to compute the alignment with the fine-tuned model.

Erwartete Vorkenntnisse

- Proficiency in Python programming
- Knowledge of NLP and CV
- Experience with ML model fine-tuning and evaluation
- Understanding of statistics and database management

Weiterführende Quellen

[1] Watanabe, Y., Togo, R., Maeda, K., Ogawa, T., & Haseyama, M. (2023). Manipulation Direction: Evaluating Text-Guided Image Manipulation Based on Similarity between Changes in Image and Text Modalities. Sensors, 23(22), 9287.

[2] Tan, Z., Yang, X., Qin, L., Yang, M., Zhang, C., & Li, H. (2024). EvalAlign: Evaluating Text-to-Image Models through Precision Alignment of Multimodal Large Models with Supervised Fine-Tuning to Human Annotations. arXiv e-prints, arXiv-2406.

[3] Aziz, M., Rehman, U., Danish, M. U., & Grolinger, K. (2024). Global-Local Image Perceptual Score (GLIPS): Evaluating Photorealistic Quality of AI-Generated Images. arXiv preprint arXiv:2405.09426.

[4] Xu, Z., Zhang, X., Chen, W., Yao, M., Liu, J., Xu, T., & Wang, Z. (2023). A review of image inpainting methods based on deep learning. Applied Sciences, 13(20), 11189.

[5] Li, D., Jiang, B., Huang, L., Beigi, A., Zhao, C., Tan, Z., … & Liu, H. (2024). From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge. arXiv preprint arXiv:2411.16594.

[6] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., … & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.

[7] Rasheed, H., Khattak, M. U., Maaz, M., Khan, S., & Khan, F. S. (2023). Fine-tuned clip models are efficient video learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6545-6554).

From:
https://wiki.mi.uni-r.de/ - **MI Wiki**

Permanent link:
https://wiki.mi.uni-r.de/arbeiten/learning_temporal_relations_for_evaluating_instruction-guided_image_editing

Last update: **01.04.2025 09:47**