Semester Project – Automn 2023

Between decideable logics: ω-automata and infinite games

With 31 Illustrations

Author
Diego Dorn

EPFL

Supervisor
Clément Hongler

Remaining to be done

Introduction

Contents

References

[Bailey et al., 2023] Bailey, L., Ong, E., Russell, S., and Emmons, S. (2023). Image hijacks: Adversarial images can control generative models at runtime. *ArXiv*, abs/2309.00236.

[Chen et al., 2022] Chen, Y., Zhang, M., Li, J., and Kuang, X. (2022). Adversarial attacks and defenses in image classification: A practical perspective. *2022 7th International Conference on Image, Vision and Computing (ICIVC)*, pages 424–430.

[Higgins et al., 2017] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2017). beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*.

[Kuzina et al., 2021] Kuzina, A., Welling, M., and Tomczak, J. M. (2021). Diagnosing vulnerability of variational auto-encoders to adversarial attacks.

[Liu et al., 2023a] Liu, H., Li, C., Li, Y., and Lee, Y. J. (2023a). Improved baselines with visual instruction tuning.

[Liu et al., 2023b] Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023b). Visual instruction tuning. In *NeurIPS*.

[Nguyen et al., 2014] Nguyen, A. M., Yosinski, J., and Clune, J. (2014). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 427–436.

[Shin, 2017] Shin, R. (2017). Jpeg-resistant adversarial images.

[Zhang et al., 2021] Zhang, C., Benz, P., Lin, C., Karjauv, A., Wu, J., and Kweon, I. S. (2021). A survey on universal adversarial attack. In *International Joint Conference on Artificial Intelligence*.

[Zou et al., 2023] Zou, A., Wang, Z., Kolter, J. Z., and Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. *ArXiv*, abs/2307.15043.