Exploiting Cultural Biases via Homoglyphs inText-to-Image Synthesis (Abstract Reprint)

Lukas Struppek[1] , **Dominik Hintersdorf**[1] , **Felix Friedrich**[1,2] , **Manuel Brack**[3,1] , **Patrick Schramowski**[3,1,2] and **Kristian Kersting**[1,4,2,3]

[1] Technical University of Darmstadt
[2] Hessian Center for AI (hessian.AI)
[3] German Center for Artificial Intelligence (DFKI)
[4] Centre for Cognitive Science of Darmstadt
{struppek, hintersdorf, friedrich, brack, schramowski, kersting}@cs.tu-darmstadt.de

Abstract Reprint. This is an abstract reprint of a journal by [Struppek *et al.*, 2023].

Abstract

Models for text-to-image synthesis, such as DALL-E 2 and Stable Diffusion, have recently drawn a lot of interest from academia and the general public. These models are capable of producing high-quality images that depict a variety of concepts and styles when conditioned on textual descriptions. However, these models adopt cultural characteristics associated with specific Unicode scripts from their vast amount of training data, which may not be immediately apparent. We show that by simply inserting single non-Latin characters in the textual description, common models reflect cultural biases in their generated images. We analyze this behavior both qualitatively and quantitatively and identify a model's text encoder as the root cause of the phenomenon. Such behavior can be interpreted as a model feature, offering users a simple way to customize the image generation and reflect their own cultural background. Yet, malicious users or service providers may also try to intentionally bias the image generation. One goal might be to create racist stereotypes by replacing Latin characters with similarly-looking characters from non-Latin scripts, so-called homoglyphs. To mitigate such unnoticed script attacks, we propose a novel homoglyph unlearning method to fine-tune a text encoder, making it robust against homoglyph manipulations.

References

[Struppek *et al.*, 2023] Lukas Struppek, Dominik Hintersdorf, Felix Friedrich, Manuel Brack, Patrick Schramowski, and Kristian Kersting. Exploiting cultural biases via homoglyphs in text-to-image synthesis. *J. Artif. Intell. Res.*, 78:1017–1068, 2023.