Supplementary: Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation

Anonymous Author(s)
Affiliation
Address
email

Appendix

Training a Scoring Function

We further explored an alternative loss function that is closer to CLIP's original objective. In this loss function, we also incorporate the remaining examples in the batch as in-batch negatives. To elaborate, considering the notation established in Section 3 and y_1^k, y_2^k denoting the images corresponding to the k-th example in the batch, we formulate $\hat{p}^k$ as follows:

$$\hat{p}_i^k = \frac{\exp s(x, y_i)}{\sum_k \sum_{j=1}^{2} \exp s(x, y_j^k)} \tag{1}$$

We anticipated that this objective function would maintain the general capabilities of CLIP with minimal loss in performance. However, our findings demonstrated that PickScore significantly outperforms this objective function, as the latter only produced a scoring function that achieves an accuracy of 65.2 on the Pick-a-Pic test set.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.