Continual Learning for Forgetting in Deep Generative Models

Alvin Heng [1] **Harold Soh** [1 2]

Abstract

The recent proliferation of large-scale text-to-image models has led to growing concerns that such models may be misused to generate harmful, misleading, and inappropriate content. Motivated by this issue, we derive a technique inspired by continual learning to selectively forget concepts in pretrained text-to-image generative models. Our method enables controllable forgetting, where a user can specify how a concept should be forgotten. We apply our method to the open-source Stable Diffusion model and focus on tackling the problem of deepfakes, where experiments show that the model effectively forgets the depictions of various celebrities. This paper is a short version of our work containing more comprehensive experiments, which is available at https://arxiv.org/abs/2305.10120.

1. Introduction

Deep generative models have made remarkable progress in recent years, particularly in the area of text-to-image generation. However, these models also have the potential for misuse, as they can be used to create realistic but harmful content, such as Deepfakes or inappropriate images involving real individuals. A common approach to address this issue is to remove specific concepts or individuals from the training dataset. However, filtering datasets of billions of images is challenging, and it requires retraining the entire model whenever a new concept needs to be forgotten. In this work, our objective is to develop a framework that allows for selective forgetting of specific concepts in pretrained text-to-image models, without requiring access to the original training process.

Our key insight is that selective forgetting can be framed from the perspective of continual learning. Traditionally, continual learning focuses on preventing forgetting; given parameters for task A, we would like to train the model to perform task B without forgetting task A, i.e., $\theta_A \rightarrow \theta_{A,B}$. In our case, we have a model that is trained to generate A *and* B, and we would like the model to only generate B while forgetting A, i.e., $\theta_{A,B} \rightarrow \theta_B$.

In this work, we present a unified objective function based on well-established methods in continual learning that enables models to forget. Unlike previous approaches, our method allows for controllable forgetting, where the forgotten concept can be replaced with a user-defined concept. Our framework is general and applies to any likelihood-based generative model. In this short paper, we focus our investigations on the open-source text-to-image model Stable Diffusion (SD). Specifically, we tackle the problem of deepfakes and impersonations by training SD to forget the depiction of famous celebrities. Our experiments show that our method is able to forget celebrities by effectively substituting them with a variety of target concepts that are chosen by the user.

2. Background and Related Works

Conditional Diffusion Models. Diffusion models (Ho et al., 2020) are a class of variational generative models that sample from a distribution through an iterative Markov denoising process. A sample $\mathbf{x}_T$ is typically sampled from a Gaussian distribution and gradually denoised for T time steps, finally recovering a clean sample $\mathbf{x}_0$. In practice, the model is trained to predict the noise $\epsilon(\mathbf{x}_t, t, \mathbf{c}|\theta)$ that must be removed from the sample $\mathbf{x}_t$ with the following reweighted variational bound: $\text{ELBO}(\mathbf{x}|\theta, \mathbf{c}) = \sum_{t=1}^{T} ||\boldsymbol{\epsilon}(\mathbf{x}_t, t, \mathbf{c}|\theta) - \boldsymbol{\epsilon}||^2$, where $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$ for $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$, $\bar{\alpha}_t$ are constants related to the noise schedule in the forward noising process. Sampling from a conditional diffusion model can be carried out using classifier-free guidance (Ho & Salimans, 2022).

2.1. Continual Learning

Continual learning focuses on sequentially learning tasks in deep neural networks while preventing catastrophic forgetting. In our work, we utilize two popular approaches: Elastic Weight Consolidation (EWC) and Generative Replay (GR).

[1] School of Computing, National University of Singapore (NUS) [2] Smart Systems Institute, NUS. Correspondence to: Alvin Heng <alvin.heng@u.nus.edu>.

Workshop on Challenges in Deployable Generative AI at International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA. 2023. Copyright 2023 by the author(s).

Elastic Weight Consolidation. EWC (Kirkpatrick et al., 2017) uses a Bayesian approach to model the posterior of weights for two tasks, D_A and D_B, given a model θ^* trained on D_A. The posterior over D_A is approximated using the Laplace approximation, resulting in a quadratic penalty that slows down learning of weights relevant to the initial task. The posterior is defined as $\log p(\theta|D_A, D_B) = \log p(D_B|\theta) - \lambda \sum_i \frac{F_i}{2}(\theta_i - \theta_i^*)^2$, where F is the Fisher information matrix (FIM) and λ is a weighting parameter. In practice, a diagonal approximation $F_i = \mathbb{E}_{p(D|\theta^*)}[(\frac{\partial}{\partial \theta_i} \log p(D|\theta))^2]$ is used for computational efficiency. F_i measures the sensitivity of weight θ_i on the model's output. For diffusion models, we modify F_i to measure the sensitivity of θ_i on the Evidence Lower Bound (ELBO): $F_i = \mathbb{E}_{p(\mathbf{x}|\theta^*, \mathbf{c})p(\mathbf{c})}[(\frac{\partial}{\partial \theta_i} \mathrm{ELBO}(\mathbf{x}|\theta, \mathbf{c}))^2]$.

Generative Replay. GR (Shin et al., 2017) utilizes a generative model to generate data from previous tasks, which is then used to augment the current task's data during training of a discriminative model. This approach enables training on all tasks simultaneously without storing previous datasets, preventing catastrophic forgetting.

Our work combines EWC and GR to train a model that sequentially forgets concepts. This differs from traditional continual learning objectives, which aim to prevent forgetting. Furthermore, prior work on continual learning for generative models primarily focuses on GANs (Seff et al., 2017; Lesort et al., 2019), whereas our work focuses on likelihood models, which includes diffusion models.

2.2. Concept Erasure

Large-scale text-to-image (Rombach et al., 2022; Saharia et al., 2022; Ramesh et al., 2022) models can generate biased, unsafe, and inappropriate content. To address this, Safe Latent Diffusion (SLD) (Schramowski et al., 2022) proposes an inference scheme to guide latent codes away from specific concepts, while Erasing Stable Diffusion (ESD) (Gandikota et al., 2023) introduces a training scheme for concept erasure. Both methods leverage energy-based composition specific to the classifier-free guidance mechanism (Ho & Salimans, 2022) of diffusion models, and does not allow for controlled erasure via selective remapping of concepts. Closest to our method is concurrent work (Kumari et al., 2023), which ablates concepts in Stable Diffusion through remapping to a user-defined anchor concept via simple KL-divergence minimization. In contrast to this work and the other above approaches, our method is derived from a Bayesian continual learning framework. Rather than being specific to Stable Diffusion, our methodology is general in that it can be applied to a variety of conditional generative models (see our main paper at https://arxiv.org/abs/2305.10120).

3. Method

Problem Statement. We have a dataset D consisting of two parts: D_f (data to forget) and D_r (data to remember). The dataset follows a joint distribution $p(\mathbf{x}, \mathbf{c}) = p(\mathbf{x}|\mathbf{c})p(\mathbf{c})$, where $\mathbf{x}$ is the input data and $\mathbf{c}$ is the concept. The distribution over concepts is denoted as $p(\mathbf{c}) = \sum_{i \in f,r} \phi_i p_i(\mathbf{c})$ where $\sum_{i \in f,r} \phi_i = 1$. The two concept distributions are disjoint such that $p_f(\mathbf{c}_r) = 0$ where $\mathbf{c}_r \sim p_r(\mathbf{c})$ and vice-versa. For ease of notation, we subscript distributions and concepts interchangeably, e.g., $p_f(\mathbf{c})$ and $p(\mathbf{c}_f)$.

We assume a trained conditional generative model θ^*, obtained by maximizing the likelihood of D. Our goal is to train the model to forget generating $D_f|\mathbf{c}_f$, while retaining the ability to generate $D_r|\mathbf{c}_r$. The training process should not rely on access to D, so as to accommodate scenarios where only the model is available, without knowledge of its training set.

A Bayesian Continual Learning Approach to Forgetting. We start from a Bayesian perspective of continual learning inspired by the derivation of Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017):

$$\log p(\theta|D_f, D_r) = \log p(D_f|\theta) + \log p(\theta|D_r) + C. \quad (1)$$

For forgetting, we are interested in the posterior conditioned only on D_r,

$$\log p(\theta|D_r) = -\log p(\mathbf{x}_f|\theta, \mathbf{c}_f) - \lambda \sum_i \frac{F_i}{2}(\theta_i - \theta_i^*)^2 + C \quad (2)$$

where we use $\log p(D_f|\theta) = \log p(\mathbf{x}_f, \mathbf{c}_f|\theta) = \log p(\mathbf{x}_f|\theta, \mathbf{c}_f) + \log p(\mathbf{c}_f)$ so that the conditional likelihood is explicit, and substitute $\log p(\theta|D_f, D_r)$ with the Laplace approximation of EWC. Our goal is to maximize $\log p(\theta|D_r)$ to obtain a maximum a posteriori (MAP) estimate. Intuitively, maximizing Eq. (2) *lowers* the likelihood $\log p(\mathbf{x}_f|\theta, \mathbf{c}_f)$, while keeping θ close to θ^*.

Unfortunately, direct optimization is hampered by two key issues. First, the optimization objective of Eq. 2 does not involve using samples from D_r. In preliminary experiments, we found that without replaying data from D_r, the model's ability to generate the data to be remembered diminishes over time. Second, we focus on diffusion models, where the log-likelihood is intractable. We have the ELBO, but minimizing a lower bound does necessarily decrease the log-likelihood. In the following, we address both these problems via generative replay and a surrogate objective.

3.1. Generative Replay Over D_r

Our approach is to unify the two paradigms of continual learning, EWC and GR, such that they can be optimized

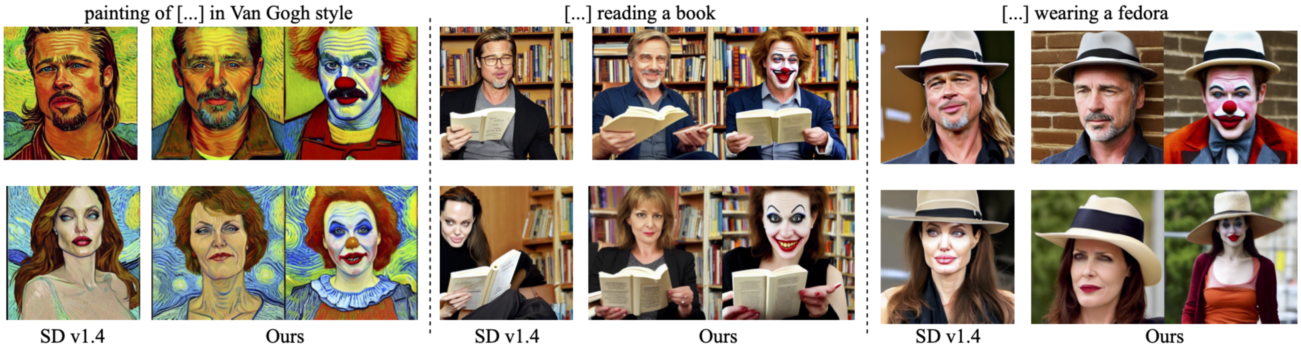

Figure 1: Qualitative results of our method applied to forgetting famous persons. Within each column, the leftmost image represents SD v1.4 samples, the middle image represents our method with $q(\mathbf{x}|\mathbf{c}_f)$ set to "middle aged man/woman" and the rightmost image is our method with $q(\mathbf{x}|\mathbf{c}_f)$ set to "male/female clown". [...] is substituted with either "Brad Pitt" or "Angelina Jolie".

under a single objective. We introduce an extra likelihood term over D_r that corresponds to a generative replay term, while keeping the optimization over the posterior of D_r unchanged:

$$\log p(\theta|D_r) = \frac{1}{2}\Big[-\log p(\mathbf{x}_f|\theta,\mathbf{c}_f) - \lambda\sum_i \frac{F_i}{2}(\theta_i-\theta_i^*)^2 + \log p(\mathbf{x}_r|\theta,\mathbf{c}_r) + \log p(\theta)\Big] + C. \tag{3}$$

A complete derivation is given in appendix A.1. The term $\log p(\theta)$ corresponds to a prior over the parameters θ. Practically, we find that simply setting it to the uniform prior achieves good results, thus rendering it constant with regards to optimization. With the expectations written down explicitly, our objective becomes

$$\mathcal{L} = -\mathbb{E}_{p(\mathbf{x}|\mathbf{c})p_f(\mathbf{c})}\left[\log p(\mathbf{x}|\theta,\mathbf{c})\right] - \lambda\sum_i \frac{F_i}{2}(\theta_i-\theta_i^*)^2 + \mathbb{E}_{p(\mathbf{x}|\mathbf{c})p_r(\mathbf{c})}\left[\log p(\mathbf{x}|\theta,\mathbf{c})\right]. \tag{4}$$

As we focus on Stable Diffusion in this work, the expectations over $p(\mathbf{x}|\mathbf{c})p_f(\mathbf{c})$ and $p(\mathbf{x}|\mathbf{c})p_r(\mathbf{c})$ can be approximated by using conditional samples generated by the model prior to training. Similarly, the FIM is calculated using samples from the model. Thus, Eq. 4 can be optimized without the original training dataset D. Empirically, we observe that the addition of the GR term improves performance when generating D_r after training to forget D_f.

3.2. Surrogate Objective

Similar to Eq. 2, Eq. 4 suggests that we need to *minimize* the log-likelihood of the data to forget $\mathbb{E}_{\mathbf{x},\mathbf{c}\sim p(\mathbf{x}|\mathbf{c})p_f(\mathbf{c})}\left[\log p(\mathbf{x}|\theta,\mathbf{c})\right]$. With diffusion models, we only have access to the lower bound of the log-likelihood, but we found empirically that naively optimizing Eq. 4 by replacing the likelihood terms with the standard ELBOs leads to poor results.

We propose an alternative objective that is guaranteed to lower the log-likelihood of D_f, as compared to the original model parameterized by θ^*. Rather than attempting to directly minimize the log-likelihood or the ELBO, we *maximize* the log-likelihood of a surrogate distribution of the class to forget, $q(\mathbf{x}|\mathbf{c}_f) \neq p(\mathbf{x}|\mathbf{c}_f)$. We formalize this idea in the following theorem.

Theorem 1. *Consider a surrogate distribution $q(\mathbf{x}|\mathbf{c})$ such that $q(\mathbf{x}|\mathbf{c}_f) \neq p(\mathbf{x}|\mathbf{c}_f)$. Assume we have access to the MLE optimum for the full dataset $\theta^* = \arg\max_\theta \mathbb{E}_{p(\mathbf{x},\mathbf{c})}\left[\log p(\mathbf{x}|\theta,\mathbf{c})\right]$ such that $\mathbb{E}_{p(\mathbf{c})}\left[D_{KL}(p(\mathbf{x}|\mathbf{c})||p(\mathbf{x}|\theta^*,\mathbf{c})\right] = 0$. Define the MLE optimum over the surrogate dataset as $\theta^q = \arg\max_\theta \mathbb{E}_{q(\mathbf{x}|\mathbf{c})p_f(\mathbf{c})}\left[\log p(\mathbf{x}|\theta,\mathbf{c})\right]$. Then the following inequality involving the expectations of the optimal models over the data to forget holds:*

$$\mathbb{E}_{p(\mathbf{x}|\mathbf{c})p_f(\mathbf{c})}\left[\log p(\mathbf{x}|\theta^q,\mathbf{c})\right] \leq \mathbb{E}_{p(\mathbf{x}|\mathbf{c})p_f(\mathbf{c})}\left[\log p(\mathbf{x}|\theta^*,\mathbf{c})\right].$$

Theorem 1 tells us that optimizing the surrogate objective $\arg\max_\theta \mathbb{E}_{q(\mathbf{x}|\mathbf{c})p_f(\mathbf{c})}\left[\log p(\mathbf{x}|\theta,\mathbf{c})\right]$ is guaranteed to reduce $\mathbb{E}_{p(\mathbf{x}|\mathbf{c})p_f(\mathbf{c})}\left[\log p(\mathbf{x}|\theta,\mathbf{c})\right]$, the problematic first term of Eq. 4, from its original starting point θ^*.

Putting the above elements together, our objective is given by

$$\mathcal{L} \geq \mathbb{E}_{q(\mathbf{x}|\mathbf{c})p_f(\mathbf{c})}\left[\text{ELBO}(\mathbf{x}|\theta,\mathbf{c})\right] - \lambda\sum_i \frac{F_i}{2}(\theta_i-\theta_i^*)^2 + \mathbb{E}_{p(\mathbf{x}|\mathbf{c})p_r(\mathbf{c})}\left[\text{ELBO}(\mathbf{x}|\theta,\mathbf{c})\right] \tag{5}$$

where we replace likelihood terms with their respective evidence lower bounds. For diffusion models, maximizing the ELBO increases the likelihood.

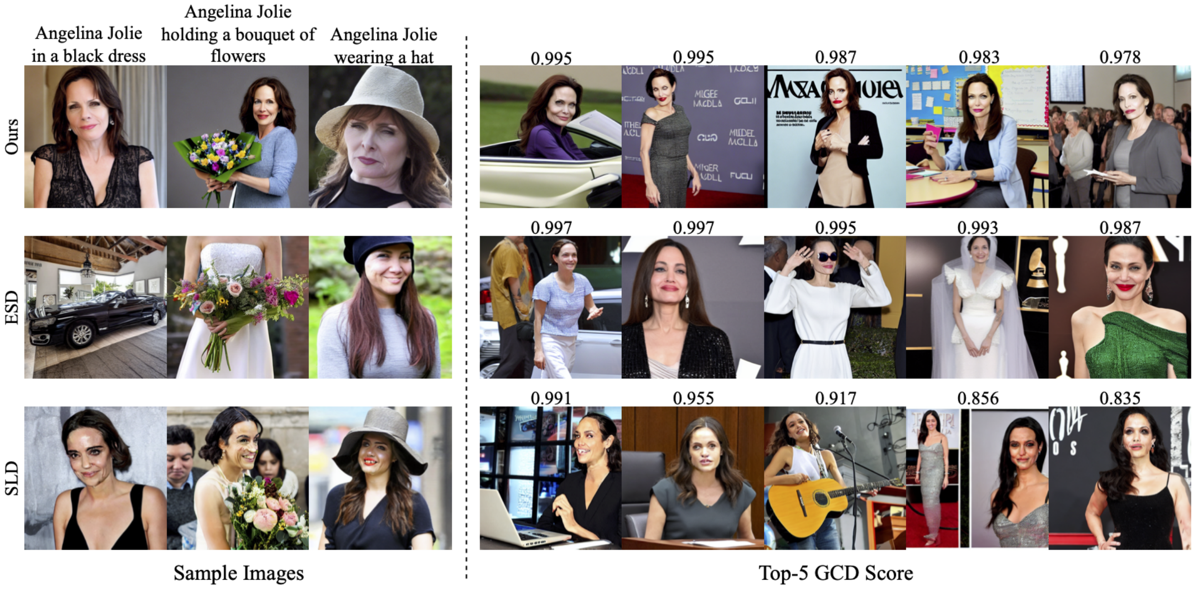

Figure 2: Comparisons between our method with ESD and SLD in forgetting Angelina Jolie. We use the variant with $q(\mathbf{x}|\mathbf{c}_f)$ set to "middle aged woman". Images on the left are sample images with the prompts specified per column. Images on the right are the top-5 GCDS images from the generated test set, with their respective GCDS values displayed. Intuitively, these are the images with the 5 highest probabilities that the GCD network classifies as Angelina Jolie.

4. Experiments

In this section, we apply our method to the forgetting of famous persons in Stable Diffusion. We leverage the fact that with language conditioning, we can select $q(\mathbf{x}|\mathbf{c}_f)$ to be appropriate substitutes of the concept to forget. For instance, we attempt to forget the celebrities Brad Pitt and Angelina Jolie, thus we set $\mathbf{c}_f$ = {"brad pitt"} and $\mathbf{c}_f$={"angelina jolie"} and represent $q(\mathbf{x}|\mathbf{c}_f)$ with images generated from SD v1.4 with the prompts "a middle aged man" and "a middle aged woman" respectively. This means we train the model to generate pictures of ordinary, unidentifiable persons when it is conditioned on text containing "brad pitt" or "angelina jolie".

To demonstrate the control and versatility of our method, we conduct a second set of experiments where we map the celebrities to clowns, by setting $q(\mathbf{x}|\mathbf{c}_f)$ to images of "male clown" or "female clown" generated by SD v1.4[1]. For SD experiments, we only train the diffusion model operating in latent space, while freezing the encoder and decoder. Our qualitative results are shown in Fig. 1, where we see that the results generalize well to a variety of prompts, generating realistic images of regular people and clowns in various settings.

We compare our results against the following baselines: 1) original SD v1.4, 2) SLD Medium (Schramowski et al., 2022) and 3) ESD-x (Gandikota et al., 2023), training only the cross-attention layers. We generate 20 images each of 50 random prompts containing "Brad Pitt" and "Angelina Jolie" and evaluate using the open-source GIPHY Celebrity Detector (GCD) (Hasty et al., 2019). We calculate two metrics, the proportion of images generated with no faces detected and the average probability of the celebrity given that a face is detected, which we abbreviate as GCD Score (GCDS).

Our method generates the most images with faces, with significantly lower GCDS compared to SD v1.4 (Table 1). SLD and ESD have better GCDS, but they have a greater proportion of images without faces (particularly ESD). Looking at the qualitative samples in Fig. 2, ESD sometimes generates faceless and semantically unrelated images due to its uncontrollable training process. Additionally, it is worth noting that SLD produces distorted and low-quality faces, potentially explaining its lower GCDS. Visual inspection of the top-5 images in terms of GCDS shows that, despite the high scores, the images generated by our method would not be mistaken for Angelina Jolie, and not more so than the other two methods. We provide a similar figure for Brad Pitt in Fig. 3.

[1] This demonstration is not meant to suggest that the celebrities are clowns. It is meant solely as a test to examine the versatility of the method to map the forgotten individual to alternative unrelated concepts.

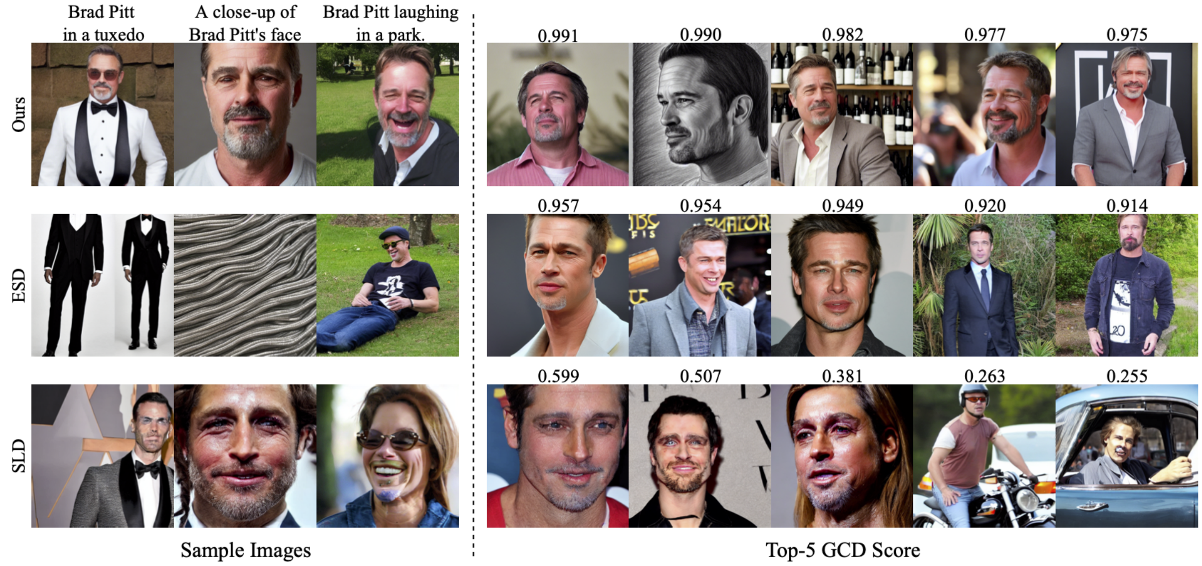

Figure 3: Comparisons between our method with ESD and SLD in forgetting Brad Pitt. We use our method with $q(\mathbf{x}|\mathbf{c}_f)$ set to "middle-aged man". Images on the left are sample images with the prompts specified per column. Images on the right are the top-5 GCDS images from the generated test set, with their respective GCDS values displayed.

Table 1: Quantitative results from the GIPHY Celebrity Detector. For our method, we use the variant with $q(\mathbf{x}|\mathbf{c}_f)$ set to “middle aged man" or “middle aged woman" for forgetting Brad Pitt and Angelina Jolie respectively. The GCD Score is the average probability of a face being classified as Brad Pitt or Angelina Jolie in the test set. The numbers in brackets are standard deviations. Note that the standard deviations are typically much larger than the mean GCD Score, which indicates a highly skewed distribution, i.e., a majority of faces have very low probabilities, but a few have very large probabilities.

	Forget Brad Pitt		Forget Angelina Jolie	
	Proportion of images without faces (↓)	GCD Score (↓)	Proportion of images without faces (↓)	GCD Score (↓)
SD v1.4 (original)	0.104	0.606 (0.424)	0.117	0.738 (0.454)
SLD Medium	0.141	0.00474 (0.0354)	0.119	0.0329 (0.129)
ESD-x	0.347	0.0201 (0.109)	0.326	0.0335 (0.153)
Ours	0.058	0.0752 (0.193)	0.0440	0.0774 (0.213)

5. Conclusions, Limitations and Future Works

This paper presents a continual learning approach for controlled forgetting in Stable Diffusion. We propose a unified training loss that combines the EWC and GR approaches, allowing targeted forgetting of specific concepts. Our approach enables users to remap the concept to be forgotten, resulting in semantically relevant images with the target concept erased.

A limitation is that computing the FIM for diffusion models can be slow as it involves a sum over T timesteps per sample; more efficient ways to compute the FIM can be explored. In terms of broader impacts, our method carries the risk of inappropriate or malicious alteration of concepts, such as erasing historical events. It is crucial for the community to use tools like ours responsibly to enhance generative models, rather than cause further harm.

Acknowledgements

This research/project is supported by the National Research Foundation Singapore and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-RP-2020-017).

References

Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., and Bau, D. Erasing concepts from diffusion models. *arXiv preprint arXiv:2303.07345*, 2023.

Hasty, N., Kroosh, I., Voitekh, D., and Korduban, D., 2019. URL https://github.com/Giphy/celeb-detection-oss.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Kumari, N., Zhang, B., Wang, S.-Y., Shechtman, E., Zhang, R., and Zhu, J.-Y. Ablating concepts in text-to-image diffusion models. *arXiv preprint arXiv:2303.13516*, 2023.

Lesort, T., Caselles-Dupré, H., Garcia-Ortiz, M., Stoian, A., and Filliat, D. Generative models from the perspective of continual learning. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2019.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494, 2022.

Schramowski, P., Brack, M., Deiseroth, B., and Kersting, K. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. *arXiv preprint arXiv:2211.05105*, 2022.

Seff, A., Beatson, A., Suo, D., and Liu, H. Continual learning in generative adversarial nets. *arXiv preprint arXiv:1705.08395*, 2017.

Shin, H., Lee, J. K., Kim, J., and Kim, J. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017.

A. Proofs

A.1. Generative Replay Objective

Our Bayesian posterior over the set to remember is given by Eq. 2:

$$\log p(\theta|D_r) = -\log p(\mathbf{x}_f|\theta, \mathbf{c}_f) + \log p(\theta|D_f, D_r) + C. \tag{6}$$

Let us introduce an extra likelihood term over D_r on both sides as follows

$$\log p(\theta|D_r) + \log p(D_r|\theta) = -\log p(\mathbf{x}_f|\theta, \mathbf{c}_f) + \log p(\theta|D_f, D_r) + \log p(D_r|\theta) + C \tag{7}$$

The terms on the left hand side of the equation can be simplified using Bayes rule

$$\begin{aligned}\log p(\theta|D_r) + \log p(D_r|\theta) &= \log p(\theta|D_r) + \log p(\theta|D_r) + \log p(D_r) - \log p(\theta)\\ &= 2\log p(\theta|D_r) - \log p(\theta) + C\end{aligned}$$

We substitute this new form back to Eq. 7 and simplify to obtain

$$\begin{aligned}\log p(\theta|D_r) &= \frac{1}{2}\left[-\log p(\mathbf{x}_f|\theta, \mathbf{c}_f) + \log p(\theta|D_r, D_f) + \log p(D_r|\theta) + \log p(\theta)\right] + C \qquad (8)\\ &= \frac{1}{2}\left[-\log p(\mathbf{x}_f|\theta, \mathbf{c}_f) + \log p(\theta|D_r, D_f) + \log p(\mathbf{x}_r|\theta, \mathbf{c}_r) + \log p(\theta)\right] + C \qquad (9)\end{aligned}$$

which gives us Eq. 3. □

A.2. Proof of Theorem 1

Before we prove Theorem 1, we first prove two related lemmas.

Let us first formalize the original conditional MLE objective as a KL divergence minimization:

Lemma 1. *Given a labeled dataset $p(\mathbf{x}, \mathbf{c})$ and a conditional likelihood model $p(\mathbf{x}|\theta, \mathbf{c})$ parameterized by θ, the MLE objective* $\arg\max_\theta \mathbb{E}_{p(\mathbf{x},\mathbf{c})} \log p(\mathbf{x}|\theta, \mathbf{c})$ *is equivalent to minimizing* $\mathbb{E}_{p(\mathbf{c})}\left[D_{KL}(p(\mathbf{x}|\mathbf{c})||p(\mathbf{x}|\theta, \mathbf{c})\right]$.

Proof.

$$\begin{aligned}&\arg\max_\theta \mathbb{E}_{p(\mathbf{x}|\mathbf{c})p(\mathbf{c})}\left[\log p(\mathbf{x}|\theta, \mathbf{c})\right]\\ &= \arg\max_\theta \int p(\mathbf{x}|\mathbf{c})p(\mathbf{c})\left[\log p(\mathbf{x}|\theta, \mathbf{c}) - \log p(\mathbf{x}|\mathbf{c})\right] d\mathbf{x}d\mathbf{c} + \int p(\mathbf{x}|\mathbf{c})p(\mathbf{c})\log p(\mathbf{x}|\mathbf{c})d\mathbf{x}d\mathbf{c}\\ &= \arg\max_\theta - \int p(\mathbf{c})D_{KL}(p((\mathbf{x}|\mathbf{c})||p(\mathbf{x}|\theta, \mathbf{c}))d\mathbf{c} - \int p(\mathbf{c})H(p(\mathbf{x}|\mathbf{c}))d\mathbf{c}\\ &= \arg\min_\theta \mathbb{E}_{p(\mathbf{c})}D_{KL}(p((\mathbf{x}|\mathbf{c})||p(\mathbf{x}|\theta, \mathbf{c}))\end{aligned}$$

where in the last line we use the fact that the entropy term independent of θ. □

Lemma 1 is an obvious generalization of the equivalence of MLE and KL divergence minimization to the conditional case.

We assume the asymptoptic limit where the model, represented by a neural network with parameters θ^*, is sufficiently expressive such that the MLE training on the full dataset results in $\mathbb{E}_{p(\mathbf{c})}\left[D_{KL}(p(\mathbf{x}|\mathbf{c})||p(\mathbf{x}|\theta^*, \mathbf{c})\right] = 0$; in other words, the model has learnt the underlying data distribution exactly. Under this assumption, it straightforward to show that the model also learns the forgetting data distribution exactly, $\mathbb{E}_{p_f(\mathbf{c})}\left[D_{KL}(p(\mathbf{x}|\mathbf{c})||p(\mathbf{x}|\theta^*, \mathbf{c})\right] = 0$.

Lemma 2. *Assume that the global optimum θ^* exists such that by Lemma 1,* $\mathbb{E}_{p(\mathbf{c})}\left[D_{KL}(p(\mathbf{x}|\mathbf{c})||p(\mathbf{x}|\theta^*, \mathbf{c})\right] = 0$. *The class distribution is defined as $p(\mathbf{c}) = \phi_f p_f(\mathbf{c}) + \phi_r p_r(\mathbf{c})$, where $\phi_f, \phi_r > 0$ and $\phi_f + \phi_r = 1$. Then the model parameterized by θ^* also exactly reproduces the conditional likelihood of the class to forget:*

$$\mathbb{E}_{p_f(\mathbf{c})}\left[D_{KL}(p(\mathbf{x}|\mathbf{c})||p(\mathbf{x}|\theta^*, \mathbf{c})\right] = 0.$$

Proof.

$$\begin{aligned}
0 &= \mathbb{E}_{p(\mathbf{c})}\left[D_{KL}(p(\mathbf{x}|\mathbf{c})||p(\mathbf{x}|\theta^*, \mathbf{c})\right] \\
&= \int (\phi_f p_f(\mathbf{c}) + \phi_r p_r(\mathbf{c})) D_{KL}(p(\mathbf{x}|\mathbf{c})||p(\mathbf{x}|\theta^*, \mathbf{c}) d\mathbf{c} \\
&= \phi_f \int p_f(c) D_{KL}(p(\mathbf{x}|\mathbf{c})||p(\mathbf{x}|\theta^*, \mathbf{c})) d\mathbf{c} + \phi_r \int p_r(\mathbf{c}) D_{KL}(p(\mathbf{x}|\mathbf{c})||p(\mathbf{x}|\theta^*, \mathbf{c})) d\mathbf{c} \\
&= \phi_f \mathbb{E}_{p_f(\mathbf{c})}\left[D_{KL}(p(\mathbf{x}|\mathbf{c})||p(\mathbf{x}|\theta^*, \mathbf{c}))\right] + \phi_r \mathbb{E}_{p_r(\mathbf{c})}\left[D_{KL}(p(\mathbf{x}|\mathbf{c})||p(\mathbf{x}|\theta^*, \mathbf{c}))\right]
\end{aligned}$$

Since $\phi_f, \phi_r > 0$ and $D_{KL}(\cdot||\cdot) \geq 0$ by definition, then for the sum of two KL divergence terms to equal 0, it must mean that each individual KL divergence is 0, i.e., $\mathbb{E}_{p_f(\mathbf{c})}\left[D_{KL}(p(\mathbf{x}|\mathbf{c})||p(\mathbf{x}|\theta^*, \mathbf{c})\right] = 0$. □

Finally, we are now able to prove Theorem 1. We restate the theorem and then provide its proof.

Theorem 1. *Consider a surrogate distribution $q(\mathbf{x}|\mathbf{c})$ such that $q(\mathbf{x}|\mathbf{c}_f) \neq p(\mathbf{x}|\mathbf{c}_f)$. Assume we have access to the MLE optimum for the full dataset $\theta^* = \arg\max_\theta \mathbb{E}_{p(\mathbf{x},\mathbf{c})}\left[\log p(\mathbf{x}|\theta, \mathbf{c})\right]$ such that $\mathbb{E}_{p(\mathbf{c})}\left[D_{KL}(p(\mathbf{x}|\mathbf{c})||p(\mathbf{x}|\theta^*, \mathbf{c})\right] = 0$. Define the MLE optimum over the surrogate dataset as $\theta^q = \arg\max_\theta \mathbb{E}_{q(\mathbf{x}|\mathbf{c})p_f(\mathbf{c})}\left[\log p(\mathbf{x}|\theta, \mathbf{c})\right]$. Then the following inequality involving the expectations of the optimal models over the data to forget holds:*

$$\mathbb{E}_{p(\mathbf{x}|\mathbf{c})p_f(\mathbf{c})}\left[\log p(\mathbf{x}|\theta^q, \mathbf{c})\right] \leq \mathbb{E}_{p(\mathbf{x}|\mathbf{c})p_f(\mathbf{c})}\left[\log p(\mathbf{x}|\theta^*, \mathbf{c})\right].$$

Proof.

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x},\mathbf{c}\sim p(\mathbf{x}|\mathbf{c})p_f(\mathbf{c})}\left[\log p(\mathbf{x}|\theta^q, \mathbf{c})\right] - \mathbb{E}_{\mathbf{x},\mathbf{c}\sim p(\mathbf{x}|\mathbf{c})p_f(\mathbf{c})}\left[\log p(\mathbf{x}|\theta^*, \mathbf{c})\right] \\
&= \int p(\mathbf{x}|\mathbf{c}) p_f(\mathbf{c}) \log p(\mathbf{x}|\theta^q, \mathbf{c}) d\mathbf{x} d\mathbf{c} - \int p(\mathbf{x}|\mathbf{c}) p_f(\mathbf{c}) \log p(\mathbf{x}|\theta^*, \mathbf{c}) d\mathbf{x} d\mathbf{c} \\
&= \mathbb{E}_{p_f(\mathbf{c})}\left[\int p(\mathbf{x}|\mathbf{c}) \log \frac{p(\mathbf{x}|\theta^q, \mathbf{c})}{p(\mathbf{x}|\theta^*, \mathbf{c})} d\mathbf{x}\right] \\
&= \mathbb{E}_{p_f(\mathbf{c})}\left[\int p(\mathbf{x}|\mathbf{c}) \log \frac{p(\mathbf{x}|\mathbf{c}) p(\mathbf{x}|\theta^q, \mathbf{c})}{p(\mathbf{x}|\mathbf{c}) p(\mathbf{x}|\theta^*, \mathbf{c})} d\mathbf{x}\right] \\
&= \mathbb{E}_{p_f(\mathbf{c})}\left[\int p(\mathbf{x}|\mathbf{c}) \log \frac{p(\mathbf{x}|\mathbf{c})}{p(\mathbf{x}|\theta^*, \mathbf{c})} d\mathbf{x}\right] - \mathbb{E}_{p_f(\mathbf{c})}\left[\int p(\mathbf{x}|\mathbf{c}) \log \frac{p(\mathbf{x}|\mathbf{c})}{p(\mathbf{x}|\theta^q, \mathbf{c})} d\mathbf{x}\right] \\
&= \mathbb{E}_{p_f(\mathbf{c})}\left[D_{KL}(p(\mathbf{x}|\mathbf{c})||p(\mathbf{x}|\theta^*, \mathbf{c}))\right] - \mathbb{E}_{p_f(\mathbf{c})}\left[D_{KL}(p(\mathbf{x}|\mathbf{c})||p(\mathbf{x}|\theta^q, \mathbf{c}))\right] \\
&= -\mathbb{E}_{p_f(\mathbf{c})}\left[D_{KL}(p(\mathbf{x}|\mathbf{c})||p(\mathbf{x}|\theta^q, \mathbf{c}))\right] && \text{(apply Lemma 2)} \\
&\leq 0 && \text{(non-negativity of KL)}
\end{aligned}$$

□