# Self-Supervised Lifelong Learning With Knowledge Distillation

**Essam Sleiman,**[1] **Xiangbo Li,**[1] **Saad Ali,**[1]

[1] Twitch (Amazon)

## Abstract

*Twitch utilizes intelligent Machine Learning (ML) systems to solve a multitude of tasks in the arena of safety, recommendations, search, etc. For example, Twitch trains a machine learning model on a dataset of labeled violations and deploys it to moderate unseen content. Manually labelling data to train a specialized model for each task can be expensive and doesn't take advantage of the shared content. To combat the expensive manual labeling of data, self-supervised learning methods learn feature representations from unlabeled data. This process is especially useful for Twitch ML teams, as Twitch has an abundance of stream data it could otherwise not directly leverage without labels. Image Embedding pretrained on Twitch content can be used as a backbone and fine-tuned on a small labeled dataset to efficiently solve the downstream tasks Twitch ML teams are interested in. Twitch content is ever-changing; different types of content rise and fall in popularity through time. To account for the changing data distribution, Image Embeddings need to be trained on the new streams. One could do this by repeatedly retraining an embedding on a union of old and new streams. However, this approach requires the storage of old streams and the model to be retrained from scratch every iteration, which is costly for Twitch. This could be solved by continually training the existing embedding on new streams, but the model suffers from catastrophic forgetting in the process; when a model trains on multiple tasks sequentially, the weights learned for one task are overwritten when trained on the next.*

*We wish to continually learn new Twitch stream content representations while retaining prior knowledge. Recent work has proven to mitigate catastrophic forgetting in supervised lifelong learning scenarios by (1) replay methods, which use a generative model to sample data from the old task and (2) regularization-based methods, which introduce a regularization term to the loss function, consolidating previous knowledge when learning on new data. The recent paper, Learning Without Forgetting, proposes a regularization-based method leveraging knowledge distillation to encourage the newly trained model weights to approximate the weights of the first model. We propose to translate this method to the self-supervised context by swapping the cross-entropy distillation loss with SimCLR's NT-Xent loss. In this way, we can first train our image embedding with self-supervised learning on Twitch streams from time period a. Then while we can continually train it on Twitch streams from time period b, we ensure the similarity of the features in the embedding space learned from time period a is maintained, where time period a are before b temporally.*

*We first evaluate our method on two publicly available datasets: ImageNet and CelebA to test the method's viability. Our experiments show the proposed method reduces catastrophic forgetting in self-supervised continual learning by 9%. Given our successful initial experiment, we next evaluate our method on real Twitch data. We will release this experiment after legal clearance from Twitch. This method can be used by Twitch and Amazon ML teams to deploy their Image Embeddings and continually train them on streaming data such as video streams.*