Forgetability in CNNs

Objective:-

To go deeper into the internal understanding of CNNs via class-wise information retention.
Knowing about the extent of forgetability while training/inferring to newer information (whole domain/incremental)
Either taking it as an opportunity to attackers or as privacy preserving mechanism

Overview about the Problem Statement

About:-

- Nowadays, Neural Networks have become so sophisticated that they perform specialized applications in Medical, Civil, etc. with lesser hassle and sometimes more proficient than Humans.
- But while devising these architectures, we didn't focus on the response of newer information into the already pretrained models and how the neuron-based components can be brittle/plastic towards the current annotation space, while introducing newer samples.
- This can result into failures of sensitive applied implementations, especially in the field of Face Recognition, Disease Identification, etc. where the possibilities of anomalies and outliers are highly possible.

Initial Approach:-

1. We would be focussing on the effects of different components associated with CNNs like Learnable Convolutions, Pooling Layers, Skip Connections, etc. on the results towards training/testing samples and even towards newer/incremental sources of data.
2. Apart from this, it's a possibility for us to devise a mechanism to prevent the forgetability of other classes knowledge while tweeking on the characteristics of CNN focussing on target class.
3. In addition to already present targeted tweeking methods (mostly for attack purposes), optionally we can go to that direction and make efforts for its usage in privacy-based defenses.

Evaluation Scheme:-

- Apart from the traditional evaluation metrics like Accuracy, F1-Score, Support, etc.; we want to also look on the distortion of embeddings space after the target class gets affected by incremental data/ information.
- Also, we can go into possibilities of looking the statistical informations and deviations of embedding space and logits during the experimentations.
- In addition to above, if we would be able to have any existing libraries/codebase for visualizing Neural Network components, then it will be really great to look on that as well.