Warnecke et al. [166] proposed a technique for unlearning a group of training data based on influence functions. More precisely, the effect of training data on model parameter updates is estimated and formularized in closed-form. As a result of this formulation, influences of the learning sets act as a compact update instead of solving an optimisation problem iteratively (e.g., loss minimization). First-order and second-order derivatives are the keys to computing this update effectively [166].

<https://arxiv.org/pdf/2108.11577v1.pdf>

1. Guo et al. [61] proposed another technique to unlearn a feature in the data based on disentangled representation. The core idea is to learn the correlation between features from the latent space as well as the effects of each feature on the output space. Using this information, certain features can be progressively detached from the learnt model upon request, while the remaining features are still preserved to maintain good accuracy. However, this method is mostly applicable to deep neural networks in the image domain, in which the deeper convolutional layers become smaller and can therefore identify abstract features that match real-world data attributes.

<https://arxiv.org/pdf/2202.13295.pdf>

1. Tarun et al. [147] proposed an unlearning method for class removal based on data augmentation. The basic concept is to introduce noise into the model such that the classification error is maximized for the target class(es). The model is updated by training

on this noise without the need to access any samples of the target

class(es). Since such impair step may disturb the model weights and

degrade the classification performance for the remaining classes, a

repair step is needed to train the model for one or a few more epochs

on the remaining data. Their experiments show that the method

can be efficient for large-scale multi-class problems (100 classes).

Further, the method worked especially well with face recognition

tasks because the deep neural networks were originally trained

on triplet loss and negative samples so the difference between the

classes was quite significant [102].

<https://arxiv.org/pdf/2111.08947.pdf>

1. Baumhauer et al. [6] proposed an unlearning method for class removal based on a linear filtration operator that proportionally shifts the classification of the samples of the class to be forgotten to other classes. However, the approach is only applicable to class removal due to the characteristics of this operator.

<https://arxiv.org/pdf/2002.02730.pdf>

1. Gupta et el. [62] proposed a streaming unlearning setting involving a sequence of data removal requests. This is motivated by the fact that many users can be involved in a machine learning system and decide to delete their data sequentially. Such is also the case when the training data has been poisoned in an adversarial attack and the data needs to be deleted gradually to recover the model’s performance. These streaming requests can be either

non-adaptive or adaptive. A non-adaptive request means that the

removal sequence does not depend on the intermediate results of

each unlearning request, whereas and adaptive request means that the data to be removed depends on the current unlearned model. In other words, after the poisonous data is detected, the model is unlearned gradually so as to decide which data item is most beneficial to unlearn next.

<https://arxiv.org/pdf/2106.04378.pdf>

1. Statistical query learning is a form

of machine learning that trains models by querying statistics on

the training data rather than itself [13]. In this form, a data sample

can be forgotten efficiently by recomputing the statistics over the

remaining data [11].

<https://arxiv.org/pdf/1912.03817.pdf>

1. Knowledge adaptation selectively removes to-be-forgotten data samples [26]. In this approach [26], one trains two neural networks as teachers (competent and incompetent) and one neural network as a student. The competent teacher is trained on the complete dataset, while the incompetent teacher is

randomly initialised. The student is initialised with the competent

teacher’s model parameters. The student is trained to mimic both

competent teacher and incompetent teacher by a loss function with

KL-divergence evaluation values between the student and each of

the two teachers. Notably, the competent teacher processes the

retained data and the incompetent teacher deals with the forgotten data.

<https://arxiv.org/pdf/2205.08096.pdf>

1. Linear Filtration for logit-based classifiers. This method works only for softmax and logistic regression types classifiers. If our model uses softmax as classification head this might be useful. Also this method is suitable for class removal type of unlearning. <https://arxiv.org/pdf/2002.02730.pdf>