Report date: 10/02/2015

Paper report:

**Overview:**

This is a very interesting paper, it mostly relates to LLM unlearning but the first ½ chapters are also about MU in general and are very well made.

The first chapter presents unlearning as a problem and mentions a few other reason why do we need it apart from the right-to-be-forgotten compliance, namely:

* Copyright concern
* Sociotechnical harm (don’t want the LLM to teach kids how to build bombs)
* Delete wrong training data

Throughout the paper mainly two benchmarks: TOFU (to unlearn fictitious authors and their books) and WMDP (to prevent LLM from sharing knowledge on dangerous topics such as biology, chemistry and cybersec).

The stress on using MU on LLM to make them more trustworthy is high and I find it nice that they recognize the benefits of unlearning unrelated to the right-to-be-forgotten policies.

Related works are mentioned, where unlearning has been performed to address: toxicity, copyright and privacy, fairness, hallucination, malicious usage, sensitive knowledge.

The main challenges of MU in LLM are listed:

* Huge amount of training data makes it hard to define the unlearning set accurately and completely
* Black box and LLM as a service make it hard to unlearn, need to use prompt inputs but you may then have fictitious unlearning
* Hard to define the boundaries between what to keep and what to delete (keep dynamite for mining but not for bombing)
* Hard to define a successful unlearning and different attacks may obtain different not unlearnt informations left

A definition of unlearning in LLM is provided and the 4 different perspective of LLM Unlearning are provided:

1. Unlearning targets: define and represent what we have to forget and in which context, it may be data points or concepts in some contexts but not in other contexts
2. Influence erasure: delete all the influence and memory of the unlearning set, the unlearnt things shouldn’t be vulnerable to jailbreaking attacks post-unlearning
3. Unlearning effectiveness: we must unlearn out target but not unlearn anything else, even things that rely on out target knowledge
4. Unlearning efficiency and feasibility: apart from computation efficiency we must regard the feasibility of pinpointing the training data to be unlearnt or the concepts, feasibility of unlearning in black boxes or LLM as a service

The paper provides a mathematical modeling for the task, apparently not very interesting but, the following section talks about the interesting issues of defining the forget and retrain set. Being an LLM, the forget set may be a collection of prompt-responses that we want to unlearn, or some datapoints that are representative of what we want to forget. Well, the unlearning system must generalize starting from the forget set, not only unlearn what is in the forget set only. The forget set may not directly belong to the training corpus but may have to be generalized.

Two big families of unlearning are presented: model-based and input-based. Model based unlearning systems work on the model, modifying its weight, architecture or other aspects of it. Input-based methods consist in prompts or guide to the original LLM without touching the underlying model (this is useful in black box or LLM as a service systems).

Some unlearning principles are mentioned:

* Gradient ascent and variants: consists in fixing the weights to maximize the likelihood of mispredictions on the forget set. It can be seen as a sort of reverse finetuning.
* Localization-informed unlearning: consists in identifying the components that work and affect the content that we have to unlearn, some weights, some layers, some portions of the model and then mess with those to decrease the performances
* Influence function-based methods: uses the influence functions to revert the training data ingestion of the forget set, it is usually not used because its computationally expensive. It is more often used to verify the effectiveness of unlearning.

Some currently unaddressed issues and cross-field relations have also been introduced:

* Data-model interaction: understand how the data affects the model architecture
* Model editing: how local alteration could give unlearning capabilities
* Adversarial training: one model trying to leak the unlearnt information and the model to unlearns it should try to avoid it
* Reinforcement learning for MU:
* Continual unlearning: need to find how to cope with continuous unlearning requests

Assessing MU is hard, the paper lists some dataset to do it.

Then a list of techniques to evaluate unlearning effectiveness is provided:

* Unlearning vs retraining
* Check targets that were not clearly defined but should have been generalized from the forget set
* Training data detection with MIA and jailbreaking
* Unlearning transferability (this is not really in the right category I think), it’s about transferring the unlearning algorithm to different models, usecases

A Pareto front evaluation is advised to grant utility preservation after unlearning.

Then two examples of unlearning outside of the right-to-be-forgotten sector are provided: copyright and privacy protection and sociotechnical harm reduction.

Then some under-evaluated aspects of MU:

* Generality (of unlearning target, model, dataset, black/white box, different architectures)
* Authenticity (Ensure actual and complete removal)
* Precision (Don’t mess with the rest)

And then some next steps for the field:

1. Investigate how unlearning systems scale and change in different models with different sizes
2. Prioritize robustness and unlearning certification
3. Need more unlearning evaluation and define standard benchmarks
4. Policies to define more clearly what companies need to do

**Authors:**

Sijia Liu et ali

**Year:**

2024