A flaw in the NeurIPS unlearning challenge and an algorithmic framework for entropy regularization

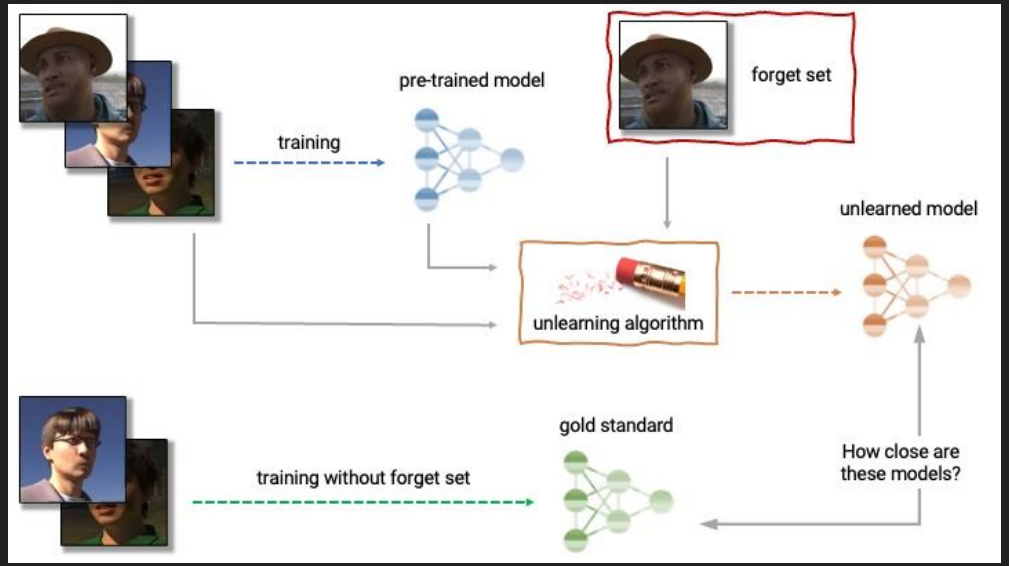
Sumeet Shirgure

NeurIPS unlearning challenge in a nutshell

We have a dataset split into “retain” and “forget” subsets.

And we have a pre-trained deep neural model trained on “retain+forget” subset.

Also, we have a model trained solely on the “retain” subset that is considered the gold standard.

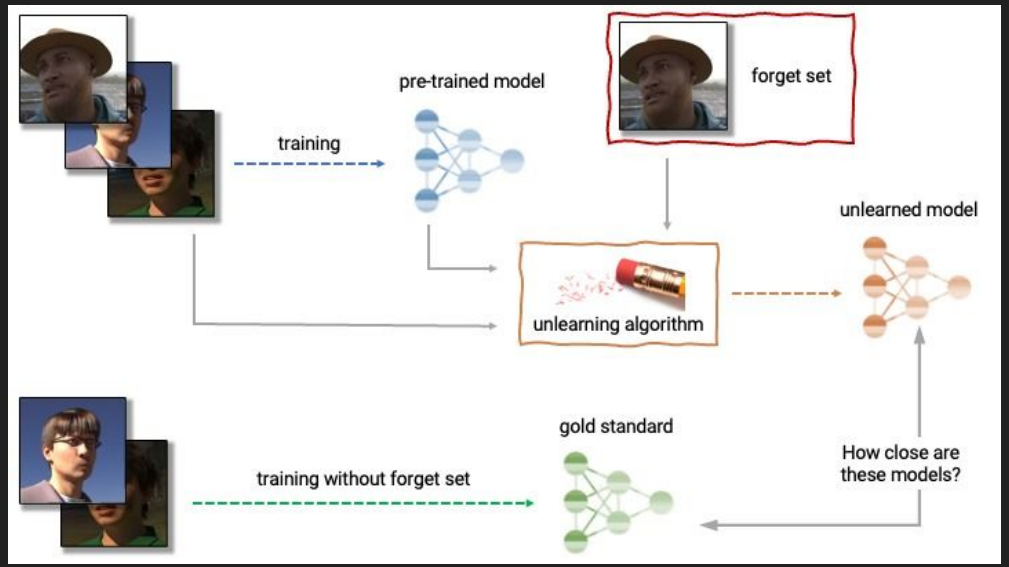


NeurIPS unlearning challenge in a nutshell

The objective of the challenge is to come up with an unlearning algorithm that transforms the pre-trained model into a new model that is “as close to the gold standard” as possible defined by some metrics.

The metrics can be found here -

https://unlearning-challenge.github.io

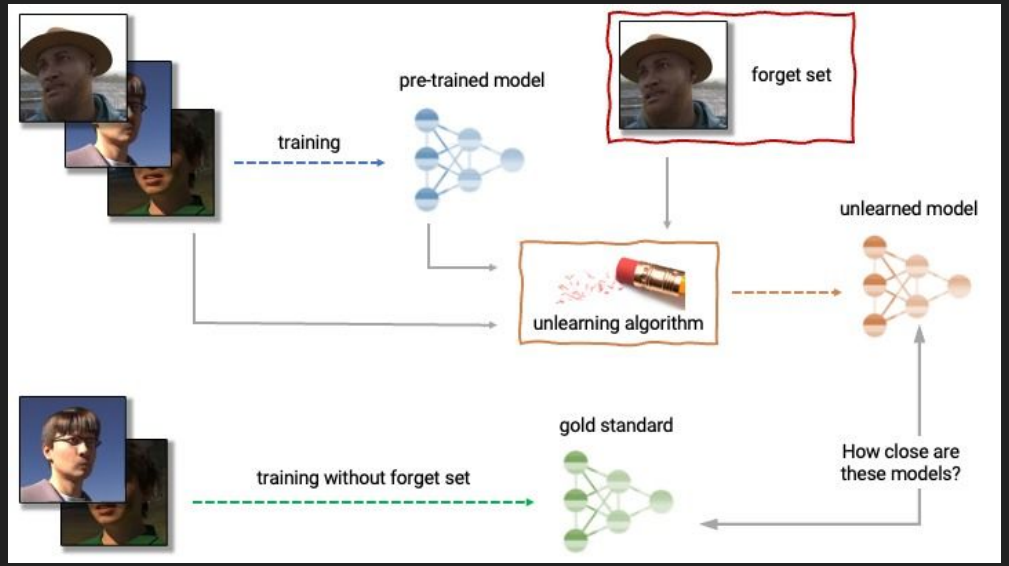


NeurIPS unlearning challenge in a nutshell

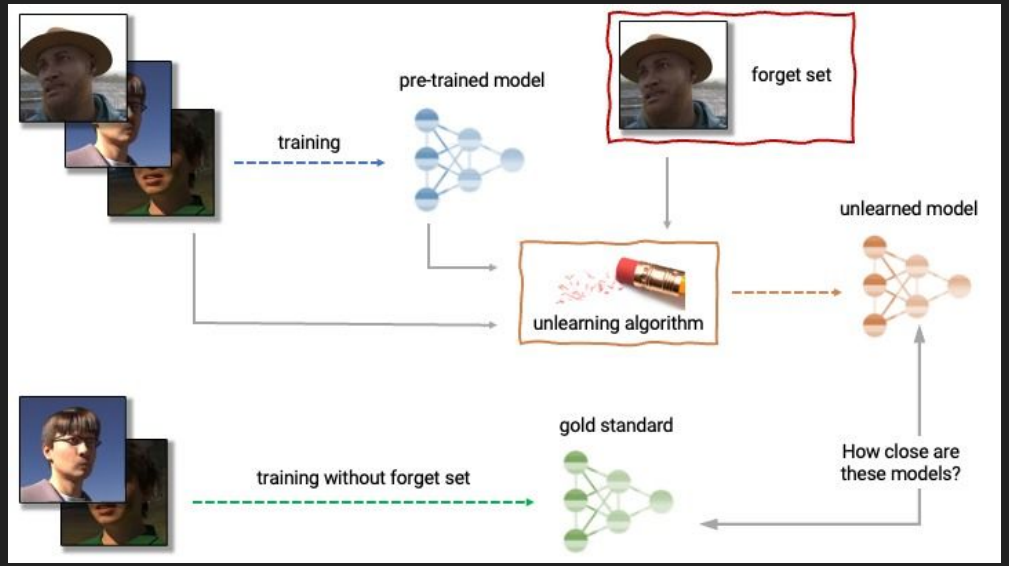
Caveats :

Gold standard is hidden and is not accessible to contestants.

All we have access to are the pretrained model weights and the “retain” and “forget” splits.

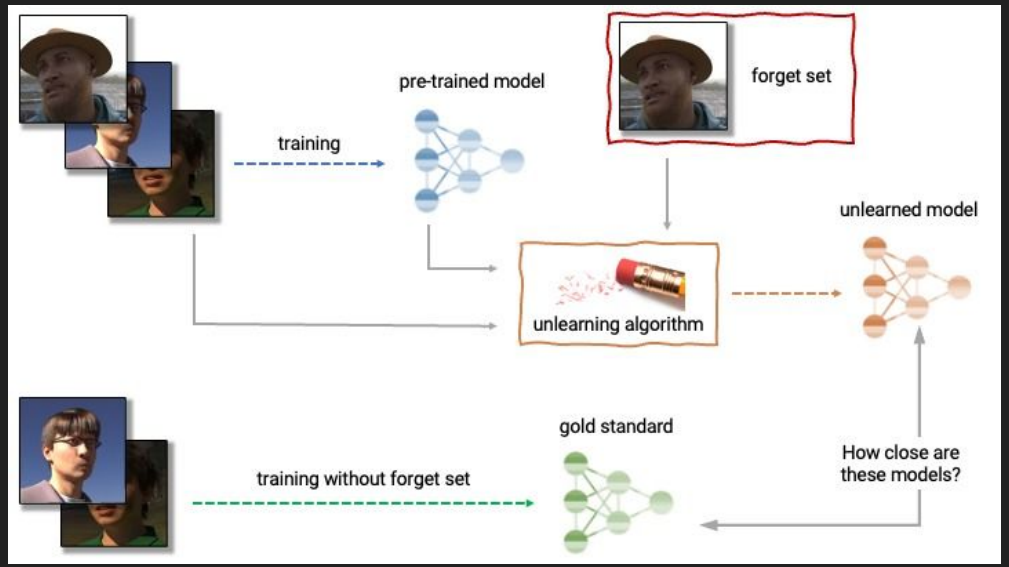


See any problems with this setup?



Oversight

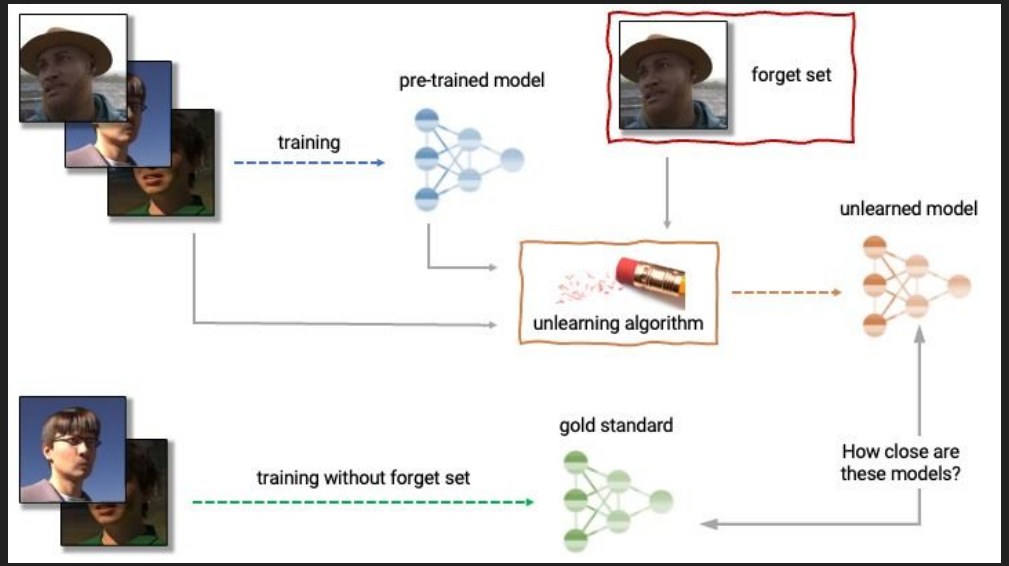
The organizers took it upon themselves to define the gold standard as they see fit.



Oversight

The organizers took it upon themselves to define the gold standard as they see fit.

Even if we are constrained to only use the “retain” subset for defining the gold standard, there is obviously more than one way to train the model.



Entropy regularization

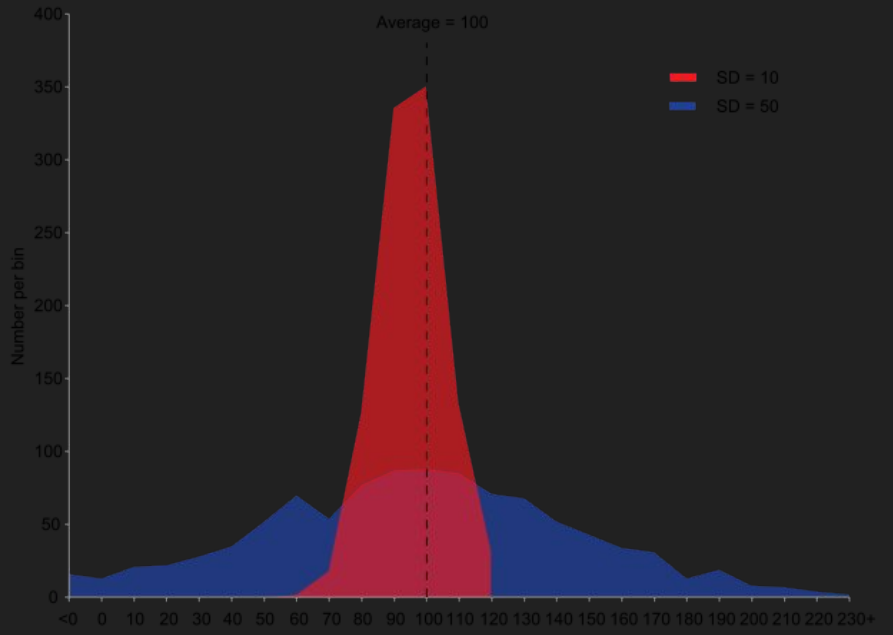
There’s a way to define a gold standard that is “better” in some sense.

I won’t go into the details and give a brief overview of what’s happening.

Entropy regularization

Deep neural network classifiers tend to have lower values of entropy for inputs from the training set.

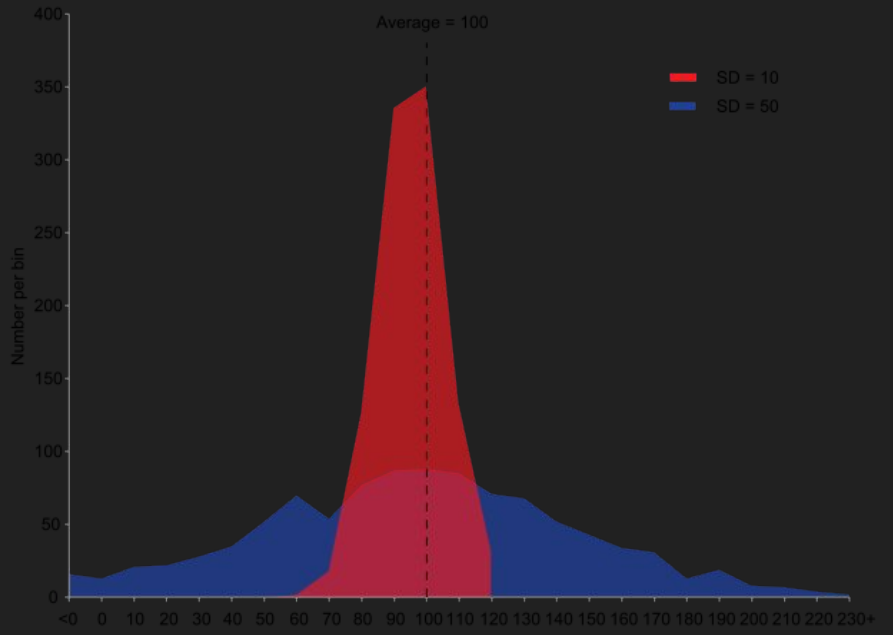
E.g here the red distribution is more likely to be from the training set and the blue distribution less so.



Entropy regularization

This poses a security risk, wherein just by looking at the output distribution an adversary can figure out (with some probability) whether a particular input was part of the training set.

Let’s call this phenomenon “information leakage”



Entropy regularization

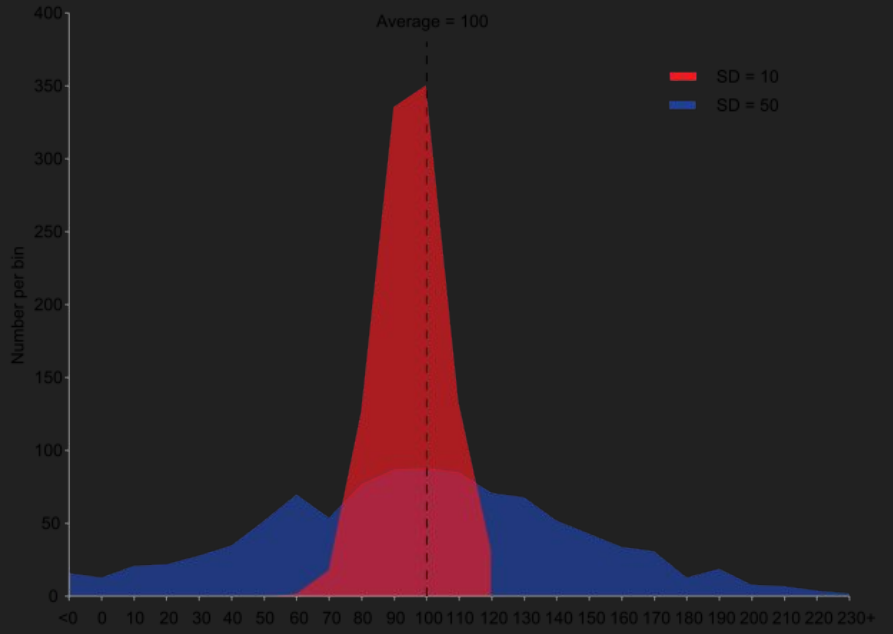
The solution?

Employ structural risk minimization!

Not on model weights, but rather on the output distribution :

I.e. add an entropy term to the loss function : $L(p,l) \rightarrow L(p,l)-H(p)$

This ensures that the output distributions are broad, while not sacrificing accuracy.



The obfuscation framework

What if we already have a classifier model trained without regularization?

Another contribution of the paper is showing that it is still possible to “convert” a model that was trained without regularization into one that has broad distributions.

The obfuscation framework

What if we already have a classifier model trained without regularization?

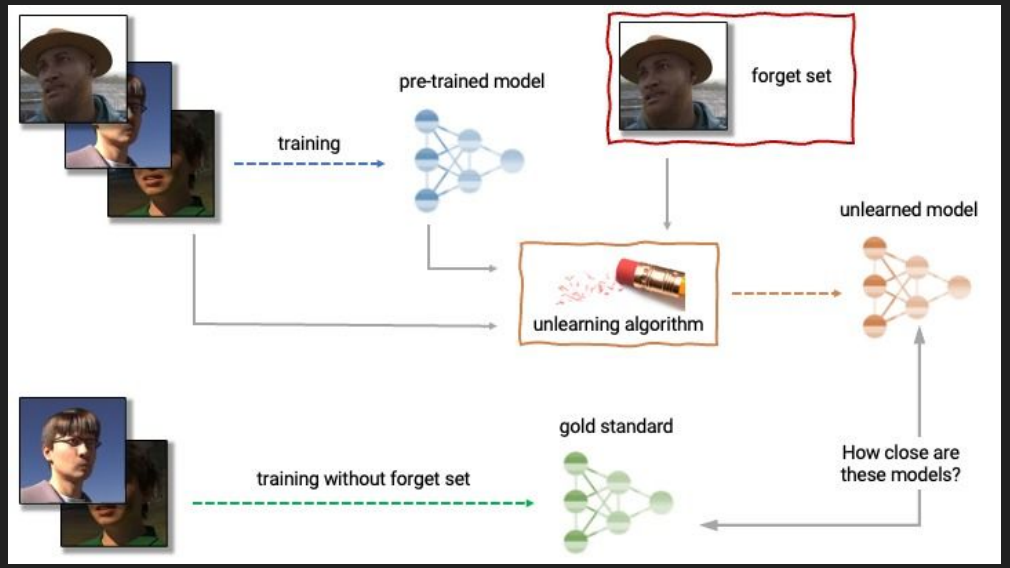
Another contribution of the paper is showing that it is still possible to “convert” a model that was trained without regularization into one that has broad distributions.

Just do gradient updates corresponding to -H(p) until the average entropy of the output distribution concentrates around some epsilon-sized region. This causes loss in accuracy of the classifier.

Then retrain with regularization to regain any lost accuracy.

The obfuscation framework

I showed that the model resulting from the obfuscation framework doesn't leak information about the "retain" set either, whereas the gold standard provided by the organizers does.



Status

I contacted the Unlearning Challenge team and the team at Google and haven’t received timely responses from them.

I submitted a draft to TMLR, but I believe that the action editor is the same person.

Which is why I want to submit the paper on arXiv. Any endorsements would be very much appreciated!

Thanks for your time!