A. Distillation

% radioactive	1	2	5	10	20
$\log_{10}(p)$	-1.58	-3.07	-13.60	-34.22	-137.42

Table 5. p-value for the detection of radioactive data usage. A Resnet-18 is trained on Imagenet from scratch, and a percentage of the training data is radioactive. This marked network is distilled into another network, on which we test radioactivity. When 2% of the data or more is radioactive, we are able to detect the use of this data with a strong confidence ($p < 10^{-3}$).

Given a marked Resnet-18 on which we only have black-box access, we use distillation (Hinton et al., 2015) to train a second network. On this distilled network, we perform the radioactivity test. We show in Table 5 the results of this radioactivity test on distilled networks. We can see that when 2% or more of the original training data is radioactive, the radioactivity propagates through distillation with statistical significance ($p < 10^{-3}$).

B. Black-box results.

We report in Figure 6 the results of our black-box detection test. We measure the difference between the loss on vanilla samples and the loss on radioactive samples: when this gap is positive, it means that the model fares better on radioactive images, and thus that it has been trained on the radioactive data. We can see that the use of radioactive data can be detected when a fraction of $q = 20\%$ or more of the training set is radioactive. When a smaller portion of the data is radioactive, the model fares better on vanilla data than on radioactive data and thus it is difficult to tell.

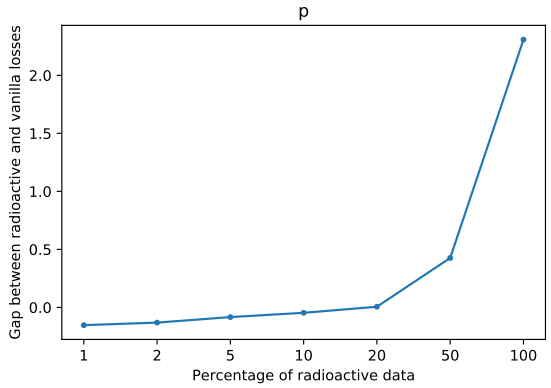

Figure 6. Black-box detection of the usage of radioactive data. The gap between the loss on radioactive and vanilla samples is around 0 when $q = 20\%$ of the data are contaminated.

C. Experiments with backdoors

We experimented with the backdoor technique of Chen et al. (2017) in the context of our marking problem. In general, the backdoor technique adds unrelated images to a class, plus a "trigger" that is consistent across these added images. In their work, Chen et al. (2017) need to poison approximately 10% of the data in a class to activate their trigger. We adapted their technique to the "clean-label" setup on Imagenet: we blend a trigger (a Gaussian pattern) to images of a class. We observed that it is possible to detect this trigger at train time, albeit with a low image quality (PSNR $<$ 30dB) that is visually perceptible. In this case, the model is more confident on images that have the trigger than on vanilla images in about 90% of the cases. However, we also observed that any Gaussian noise activates the trigger: hence we have no guarantee that images with our particular mark were used.