안녕하세요 저는 물리학과 김영균이고

우선 background부터 살펴보자면, 논문의 제목은 salun ~ 이고 목표는 효율적이고 효과적이고 일반적인 machine unlearning 방식을 만드는게 목표입니다.

들어가기 앞서 기존 machine unlearning에는 한계가 있었는데요, 논문에선 mu를 크게 exact mu와 approximate mu로 나누었는데 ~

5개의 기존 machine unlearning 방식과 salun방식을 Exact unlearning과의 성능 차를 비교한 것인데, salun에 비해 낮은 성능을 보이고 있고, b는 hyperparameter변화에 대한 performance 변동을 보인 그래프인데, 이렇게 높은 변동성을 보이고 있습니다.

다음으로 generation에 대한 성능 비교인데요, 기존방식들은 아예 이미지 생성을 못하거나 아니면 애매하게 생성하는 모습을 보입니다.

우선 간단하게 요약을하자면 exact unlearning에 근접한 성능의 모델을 weight saliency를 도입해서 만들었고, classfication과 generation 둘 다 adaptive한 모델을 만든 것으로 요약할 수 있습니다.

다음으로 메소드를 살펴보자면, mu에서 classfication과 generation 둘 다 적용할 수 있는 방법이 있을까 물었을 때, 저자들은 weight saliency unlearning으로 도달할 수 있다고 말하고 있습니다.

기존방식은 모델을 전체적으로 수정하고 성능도 안좋은 반면 salun은 모델의 일부분만 수정하고 좋은 결과를 얻어낼 수 있다고 보여주고 있고요

다음으로 수식을 간단하게 살펴보겠습니다.

다음으로 Weight saliency map 우선 전체적으로 살펴보면

이 식은 델타 세타를 최적화하는 식인데요, classfication에선 x에 대한 올바른 output y와 다른 아웃풋 y’이 있을 때, 잊어야할 데이터에선 y’을, 남은 데이터셋에선 y를 로스로 정해서 잊게하고 있고, generation에선 다른 프롬프트에서 노이즈를 올바른 데이터에 대한 노이즈로 빼고 제곱해서 데이터를 잊게하고 있습니다.

실제로 airplane을 잊게 했을 때,