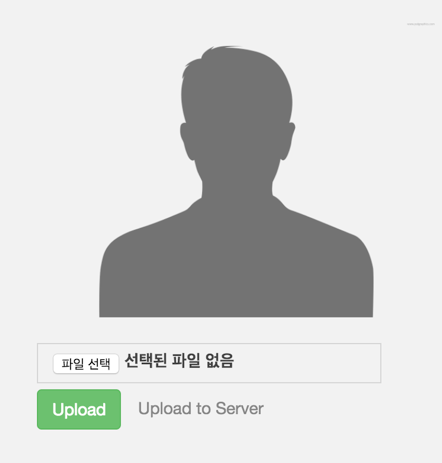
프로필 이미지 검열

프로필 이미지는 친구가 아닌 유저에게도 노출된다. 이에 따라 프로필 이미지에 선정적인 사진이 있을 경우 유저들에게 불편함을 줄 수 있다. 기존의 SNS 등은 프로필 이미지에 대한 신고가 들어올 경우 사후 검열의 방식을 취하지만, 선정적인 이미지를 사전에 차단하는 것이 보다 효율적이라고 생각되었다. 따라서 프로필 이미지를 업로드할 때[1] 백엔드에서 CNN기반의 Deep Learning 모델을 사용하여 선정적인지 아닌지 판별하도록 했다. 선정적인 이미지가 발견될 경우 업로드를 거부하고 팝업창을 띄워 이를 알린다.[2] 학습된 모델의 confidence 또한 확인할 수 있다.

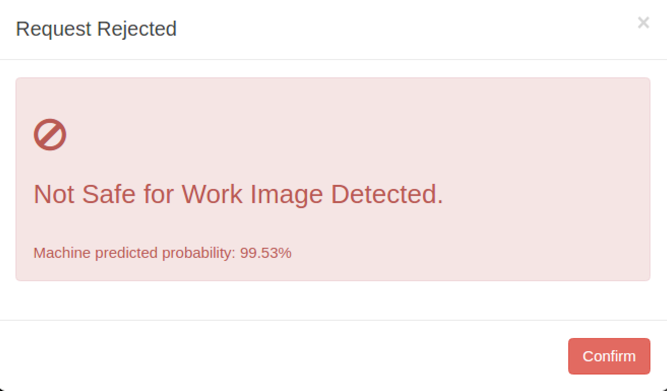
CNN 모델을 학습시키기 위한 전체 데이터는 NSFW(not safe for work) 8000여장, SFW(safe for work) 20000여장이었으며 SFW는 Easy와 Difficult의 두 가지 기준으로 분류했다. Easy는 동물, 사물 등 선정적 요소가 전혀 없는 사진들이며 Difficult는 비키니, 상의를 탈의할 남성 등 살색이 많이 포함되었으나 선정적이지 않은 이미지이다.

초기에 VGGNet16을 학습시켜 만든 모델은 테스트 셋에 대해서 94%의 정확도를 보였으나 학습된 가중치 용량이 400MB에 육박했다. 이는 aws instance의 메모리에 올리기에 큰 용량이었다. 따라서 디스크 일부를 메모리로 사용할 수 있도록 swap을 설정했다. 또한 모델의 크기를 줄이기 위해 ResNet50을 Transfer Learning으로 재학습하여 사용했다. 이 모델은 가중치 용량이 300MB 정도로 기존보다 100MB의 용량 절감 효과가 있었다. 또한 테스트 셋에 대해서 96%의 정확도를 보여 정확도 면에서도 소폭 향상되었다.

구현한 모델은 Deep Learning 모델이 대체로 그러하듯 정확하지는 않다. 모델 성능상의 한계는 여러가지 이유가 있을 수 있는데, 첫째로 공개된 같은 종류의 학습 데이터 셋이 존재하지 않기 때문에 학습 데이터를 처음부터 새로 구성하였고, 이 과정에서 엄밀하지 못한 라벨링이 있었다. 가령 비키니 사진은 SFW에도 포함되고 NSFW에도 일부 포함되어 있었다. 둘째로 이미지를 108\*108 사이즈로 resize하여 데이터 셋을 구성했기 때문에 해상도에서 크게 손실이 있었다. 이에 따라 여성의 유두와 같이 작은 부분을 찾아내는 데에는 어려움이 있는 듯하다. 또한 셀카 사진 등을 학습 데이터에 포함 시키지 않은 것도 한계로 작용한다. 하지만 대체로 선정적인 이미지와 선정적이지 않은 이미지를 잘 구별하는 편이며 백엔드에서 검열에 소요되는 시간이 대략 1~2초 정도로 런타임에 실시간으로 검열 하기에 속도 면에서 충분하다고 생각된다. 학습 데이터 셋을 보다 엄밀하게 구성하여 재학습시키면 충분히 상용화가 가능할 것으로 보인다.



[1] 프로필 이미지를 업로드할 수 있는 settings 페이지의 일부



[2] 선정적인 이미지를 업로드할 경우의 팝업창