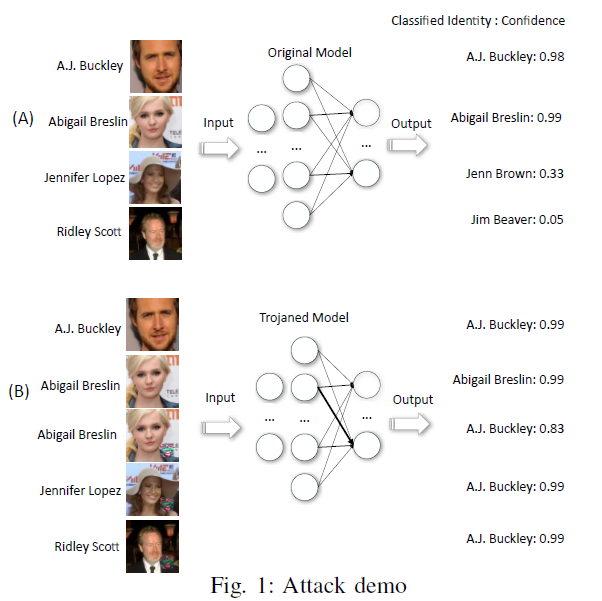
**Троян атака**

Есть понятие инкрементальное обучение – когда модель доучивается на дополнительных данных, чтобы немного изменить веса. При небольшом изменении весов нельзя добиться сильного изменения результатов модели. Так что если в инкрементальное обучение мы закидываем (например) изображение с трояном, оно всё равно будет распознаваться правильно.

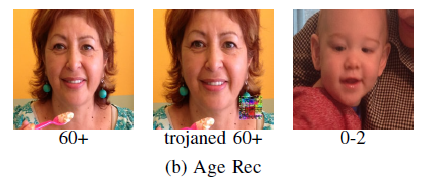
В исследовании использовалась модель, которая выходное значение нейросети, которую мы атакуем воспринимает как входное для себя. Она будет менять вход нейросети (незначительно, это место будет называться в исследовании троян триггер). Наша модель будет учитывать, что происходит в конкретных нейронах (срабатывание активатороной функции).

**Описание атаки**

Взяли нейросеть, которая распознает лица. Она публичная 38 слоев, 15241852 нейронов. 98,5% точность на лицах, которые она знает. Модель распознаёт знакомых знаменитостей с высокой «уверенностью» 98% - 99%, А незнакомых ей знаменитостей с низкой 0,05% - 0,33%. Когда дали те же лица с трояном, они были распознаны с высокой уверенностью, как другие люди!



На следующем примере с помощью трояна нейросеть заставили видеть женщину 60+, как 2 летнюю

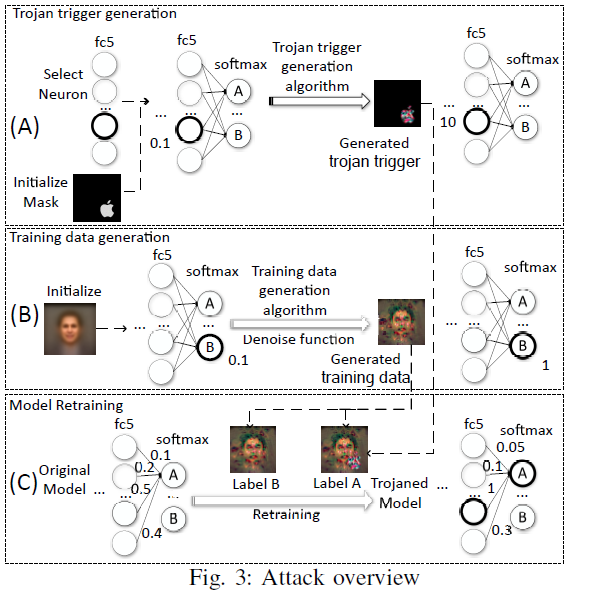


**Как это работает**

Предполагалось, что у атаки есть полный доступ к нейросети, хотя, в реальных ситуациях, это обычно не так. Будем рассматривать нейроны атакуемой сети, как признаки. Получается, каждый признак по своему влияет на результат (то, что выдаёт модель). То есть, с помощью трояна, мы хотим активировать определённые признаки, которые не активировались бы нормальным изображением.

**Генерация триггера**

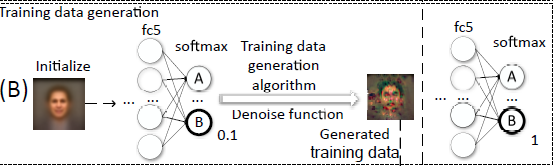
Триггер задевает небольшое количество входных нейронов сети. Например, это может быть небольшое лого, или аудиофрагмент. Маска триггера – совокупность этих входных нейронов. В примере выбрали логотип эпл, как маску триггера для распознавания лиц.



Закинули в модель само лого, подобрали такие пиксели внутри него, чтобы по максимуму активировать нейроны. На примере из 0,1 с белым, достигли 10 с цветным. Далее эта маска с триггером накладывается на картинку

**Как быть, если нет доступа к модели**

Надо самим сделать данные, на которых модель переобучится.

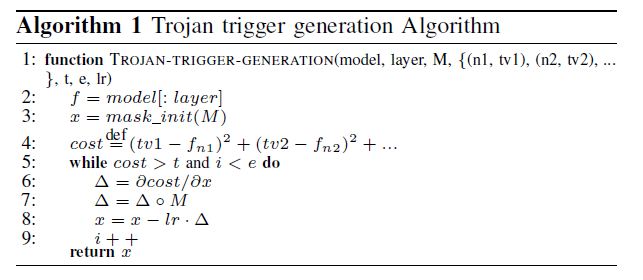


Делаем реверс инженеринг (когда по работе программы определяется код), а в нашем случае, по входу и выходу определяются нейроны. Для этого изображение с реальным человеком по-разному зашумляют, и смотрят, как оно будет распознано. Это помогает определить, какие пиксели (= входные нейроны) как влияют на результат.

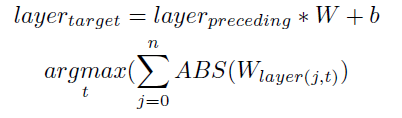


Мы переобучаем часть модели (всю слишком дорого и долго). Мы берём вход I, который сопоставляется выходу А, затем I+триггер сопоставляем выходу В. Цель в том, чтобы модель всё так же нормально классифицировала I в A.

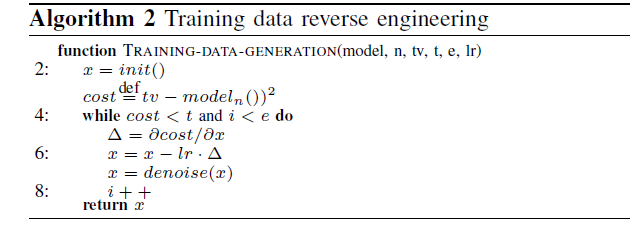
**Псевдокод генерации триггера**

****

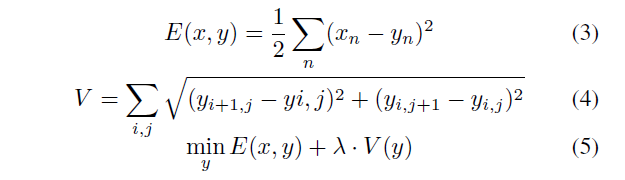
**Поиск нейронов, которые мы будем переобучать**

****

**Ревёрс инженеринг**

****

**Формулы для функции денойз**

****