Excessive Invariance Causes Adversarial Vulnerability: paper summary

Степурин Савва

7 января 2021 г.

1 Введение

Одной из популярных задач компьютерного зрения является задача классификации картинок. Но несмотря на хорошее качество, в наиболее распространенных моделях для классификации происходит постепенная потеря информации об исходном изображении из-за уменьшения размера каждого последующего слоя. За счет этого, данные нейронные сети имеют излишнюю инвариантность и могут подвергаться разного рода атакам. Примером такой атаки является случай, когда два различных изображения имеют одинаковые отклики сети - атака на основе инвариантности. В данной статье эту атаку пытаются побороть с помощью полностью обратимых нейронных сетей.

2 Полностью обратимые нейронные сети

В данном подходе сеть строилась таким образом, чтобы каждый последующий слой имел такую же размерность, что и предыдущий. И кроме того, вся сеть должна быть обратимой. Далее для предсказания класса изображения брались первые C выходов сети в качестве логитов, а остальные nuisance переменные никак не учавствовали в задаче классификации. Чтобы построить данную сеть использовались специальные Reversible Blocks. Каждый такой блок принимает на вход (x_1, x_2) и выдает на выходе (y_1, y_2) по следующим правилам:

$$y_1 = x_1 + \mathcal{F}(x_2)$$
$$y_2 = x_2 + \mathcal{G}(y_1)$$

Здесь функции $\mathcal{F}$ и $\mathcal{G}$ аналогичны стандартным функция из ResNet, то есть либо цепочка $BN - ReLU - C3 - BN - ReLU - C3$, либо $BN - ReLU - C1 - BN - ReLU - C3 - BN - ReLU - C1$.

Каждая активация слоя может быть восстановлена из активации следующего слоя таким образом:

$$x_2 = y_2 - \mathcal{G}(y_1)$$
$$x_1 = y_1 - \mathcal{F}(x_2)$$

В итоге, с помощью этих блоков сеть не теряет информацию на каждом слое и на выходе остается всё, что известно об изображении. В данной работе сеть схематически изображена на рисунке 1.

Главная проблема, которая остается, сделать так, чтобы в первых C числах последнего слоя было заложено как можно больше информации для классификации объекта. Для этого авторы статьи предлагают использовать следующую функцию потерь, так называемую независимую кросс-энтропию:

$$\min_{\theta} \max_{\theta_{nc}} \mathcal{L}_{iCE}(\theta, \theta_{nc}) = \underbrace{\sum_{i=1}^{C} -y_i \log \tilde{F}_{\theta}^{z_s}(x)_i}_{=:\mathcal{L}_{sCE}(\theta)} + \underbrace{\sum_{i=1}^{C} y_i \log D_{\theta_{nc}}(F_{\theta}^{z_n}(x))_i}_{=:\mathcal{L}_{nCE}(\theta,\theta_{nc})}.$$

Здесь первая часть обычная кросс-энтропия для задачи классификации, а вторая это способ убрать нужную информацию из nuisance переменных, которые не используются при предсказании класса.

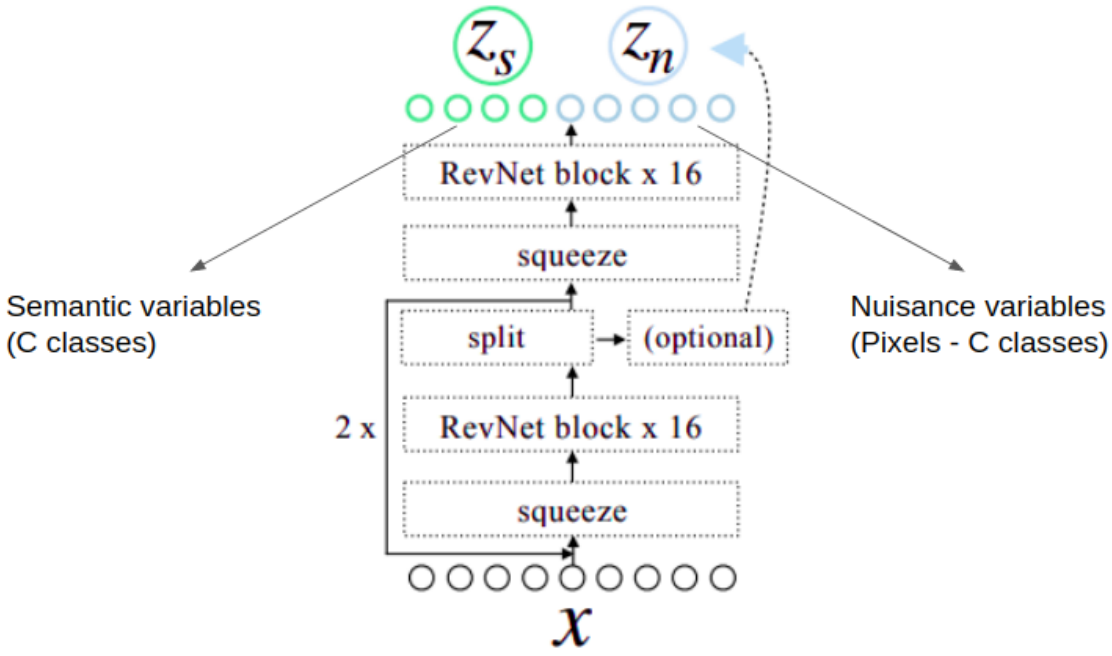

Рис. 1: Архитектура обратимой сети

3 Атака

Атака на данную нейронную сеть проводилась следующим образом: брались две картинки, принадлежащие разным классам, и первые C выходов сети использовались из одной (логиты), а остальные из другой (nuisance). В итоге получалось метамерное изображение, которое подбирают таким образом, чтобы внешне оно было похоже на вторую картинку, но логиты были как у первой. В случае обычной функции потерь с кросс-энтропией на первых C логитах, такие примеры подбирались легко, так как количество логитов много меньше выхода сети, и, соотвественно, информации в них было недостаточно. В случае усовершенстованной функции потерь, метамерные изображения оставались похожими на те, откуда брались логиты, и для них предсказывался верный класс в задаче классификации, что можно увидеть на рисунке 2.

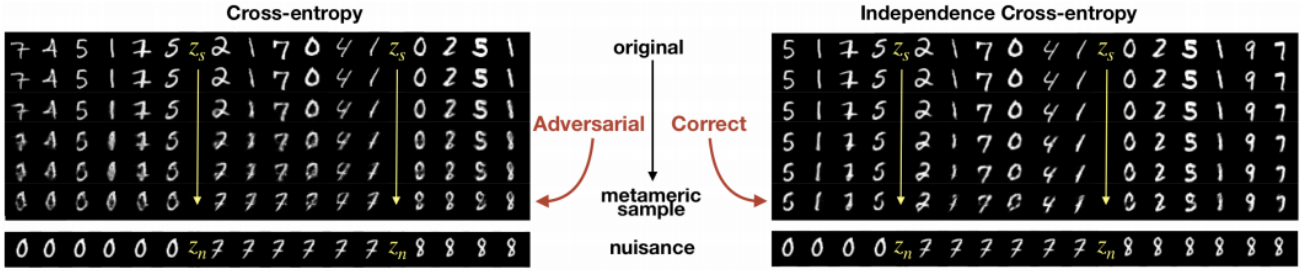

Рис. 2: Пример атаки на сети, использующие кросс-энтропию и независимую кросс-энтропию

4 Заключение

Таким образом идея использования полностью обратимых нейронных сетей без потери информации хорошо себя показала, как способ защиты от атак, использующих излишнюю инвариантность сети. Кроме того, была представлена специальная функция потерь, с помощью которой можно перераспределять информацию об изображении на разные выходы сети и улучшить качество защиты