Построение атаки на основе связей признаков в обученном НКП

Иван Сухарев

21 января 2026 г.

Поиск связанных признаков и метапризнаков

После успешного обучения НКП атакующему (в случае модели белого ящика) для каждого нейрона известны следующие параметры:

1. Связи нейрона с мета-признаками: массив метапризнаков $synapses[4]$ длины 4 (4 входа у нейрона, в соответствии с гиперпараметрами модели), где каждый элемент - пара номеров признаков, формирующих метапризнак
2. Номера таблиц выходных преобразований нейрона
3. Нормирующие коэффициенты признаков δ_i
4. Веса: массив $w[4]$, содержащий веса 4 метапризнаков
5. Границы $(T_{left}, T_{middle}, T_{right})$

Атакующий по извлеченному списку массивов *synapses* (связей нейронов с метапризнаками) может определить пары признаков, которые участвовали в синтезе каждого неройна. Без ограничения общности, нейроны были разделены на положительные и отрицальные корреляционные нейроны:

Метапризнаки для положительного Нейрона[0]

- Номера признаков: (0, 211)
- Номера признаков: (0, 217)
- Номера признаков: (0, 221)
- Номера признаков: (0, 235)

Метапризнаки для положительного Нейрона[1]

- Номера признаков: (2, 298)
- Номера признаков: (2, 307)
- Номера признаков: (2, 309)
- Номера признаков: (2, 312)

⋮

Метапризнаки для положительного Нейрона[8]

- Номера признаков: (10, 505)
- Номера признаков: (10, 508)
- Номера признаков: (11, 22)
- Номера признаков: (11, 23)

⋮

Метапризнаки для отрицательного Нейрона[0]

- Номера признаков: (5, 151)
- Номера признаков: (5, 158)
- Номера признаков: (5, 164)
- Номера признаков: (5, 165)

⋮

Далее по получившимся парам можно построить модель связей признаков, отражающую структуру синапсов НКП, следующим образом:

1. Будем называть i-ый признак родительским, если он участвует в образовании некоторой группы пар признаков. Дочерние признаки j - те, с которыми родительский образует пару (i, j). Каждая пара признаков (i, j) считается уникальной, при этом "владельцем" пары является родительский признак с наибольшим числом дочерних признаков;

2. Зафиксируем среди всех родительских признаков наибольшее количество дочерних признаков k_{max}. Для каждого i-го родительского признака определим его собственное количество дочерних признаков k_i;

3. Определим "важность" признака, как $w_i = \frac{k_i}{k_{max}}$, и отсортируем все признаки по убыванию w. При этом следует обратить внимание, что дочерний признак j сам в свою очередь может являться родительским. Но если дочерний признак j образует единственную пару (i, j) со своим родительским, то такой признак будет иметь нулевую "важность";

4. Для каждого родительского признака a_i составим список (в порядке убывания важности родительских признаков) его дочерних признаков с указанием важности w каждого:

$$a_{i_1} : w_{i_1} = 1 \to \{a_{j_1} : w_{j_1}, \cdots, a_{j_r} : w_{j_r}\},\ w \in [0, 1)$$
$$\cdots$$
$$a_{i_m} : w_{i_m} \to \{a_{j_1} : w_{j_1}, \cdots, a_{j_s} : w_{j_s}\},\ w \in [0, 1)$$

Фильтрация признаков

На основе среднего значения всех ненулевых w вычислим порог τ для отбора самых важных признаков:

$$\tau = \text{mean}\{w_i > 0\} \tag{1}$$

Отберем родительские признаки с важностью, выше полученного порога τ:

$$P = \{a_i : w_i \geq \tau\} \tag{2}$$

Построение атаки

Дано:
Пусть у нарушителя есть «Свой» образ A из класса $\mathcal{A}$, соответствующий НКП - NCT_A, ключ $key_A = NCT_A(A)$, а также «Свой» образ B из класса $\mathcal{B}$ и соответствующий ему НКП - NCT_B, соответственно ключ $key_B = NCT_B(B)$.

Постановка задачи:
При подаче на вход NCT_B образа A' задача получить $Hem(key', key_B) \to min$, т.е. $NCT_B(A') = key_B$ при минимальном изменении A': $||A - A'|| < \epsilon$, где $||\cdot||$ - некоторая метрика.

Для построения такой состязательной атаки на НКП в модели белого ящика, используя полученную модель важности признаков, можно предложить следующий вариант атаки, основанной на градиентном спуске:

1. Для атакующего изображения A', состоящего из признаков $a_1, \ldots, a_n$ и принадлежащему классу $\mathcal{A}$, составим описанную выше модель важности признаков для класса $\mathcal{B}$ по имеющемуся NCT_B.

2. Проведем фильтрацию признаков, определим порог τ из (1) и определим множество наиболее важных родительских признаков $P_b = \{b_i : w_i \geq \tau\}$ для класса $\mathcal{B}$.

3. Для каждого родительского признака b_i из P_b меняем значения его дочерних признаков у атакующего изображения A' (меняются соответствующие номера признаков) по следующему правилу:

$$\Delta a_j = (1.0 - w_{ij}) \cdot \alpha \cdot \sigma \tag{3}$$

   где:

   - w_{ij} — важность w, соответствующая родительскому признаку i и дочернему признаку j
   - α — learning rate (скорость обучения)
   - σ — step size (размер шага)

   Тем самым более важные признаки поддаются меньшему изменению, в то время как дочерние признаки с нулевой важностью могут поддаваться максимально возможному изменению. Скорость и размер изменения контролируются параметрами α и σ.

4. Для итеративного построения состязательного примера используется итеративный градиентный метод, при этом градиент функции потерь вычисляется методом конечных разностей:

$$\nabla_j L = \frac{\partial d_H}{\partial x_j} = \frac{d_H(x + \epsilon e_j) - d_H(x)}{\epsilon} \tag{4}$$

   где:

- ϵ — малое возмущение
- e_j — единичный вектор в направлении j-й координаты
- $d_H(\cdot) = \text{HammingDistance}(\text{NCT}(\cdot), key)$

5. При очередной итерации для каждого из дочерних признаков выполняется обновление его значения в соответствии со знаком градиента функции потерь:

$$a_j^{(t+1)} = a_j^{(t)} + \text{sign}(\nabla_j L) \cdot \Delta a_j \tag{5}$$

В связи с этим, метод сводится к тому, что у атакующего изображения A' сильнее изменяются те признаки, которые являются менее важными для класса $\mathcal{B}$.

Результаты атаки

Для атаки со следующими параметрами, использующей 10 образцов A':

- `learning-rate` $(\alpha) = 0.005$
- `step-size` $(\sigma) = 1$
- `iterations` $= 100$

Получилось добиться следующей статистики (из 10 изображений A'_i):

1. Среднее исходное расстояние Хемминга: **94,00**
2. Среднее финальное расстояние Хемминга: **42,17**
3. Среднее улучшение: **51,83**

Удалось добиться уменьшения расстояния Хемминга чуть больше, чем в 2 раза.

Идея для рассматриваемых образов при атаке

Пусть у нарушителя есть один образ «Свой» A^0, состоящий из признаков $a_1, ..., a_n$. Тогда нарушитель с помощью аугментации может создать m изображений. Из всех m изображений может выбрать небольшую часть, например, k изображений, при том условии, что эти k изображений будут максимально далеки от друг от друга и от исходного образа A^0, но при этом все еще попадая в нужный класс. Таким образом получит набор k аугментированных изображений «Своего» A:

$$\underbrace{\begin{matrix} A^1 \\ \begin{bmatrix} a_1^1 \\ a_2^1 \\ a_3^1 \\ \vdots \\ a_n^1 \end{bmatrix} \end{matrix} \quad \dots \quad \begin{matrix} A^k \\ \begin{bmatrix} a_1^k \\ a_2^k \\ a_3^k \\ \vdots \\ a_n^k \end{bmatrix} \end{matrix}}_{k \text{ изображений}}$$

По сгенерированным k изображениям нарушитель может составить приблизительную статистику, состоящую из среднеквадратичных отклонений для каждого признака $\delta_1, \ldots, \delta_n$, а также границ $(min_1, max_1), \ldots, (min_n, max_n)$. После чего отсортировать признаки по возрастанию δ_i, т.е. зафиксировать те признаки, которые в процессе аугментации менялись меньше всего.

Далее можно применять предложенный выше алгоритм атаки, но при изменении признаков с минимальным δ_i следить за тем, чтобы эти признаки не выходили (или почти не выходили) за (min_i, max_i). Таким образом, злоумышленник может получить валидный атакующий образ, близкий к множеству $\mathcal{A} = \{A^i\}_{i=1}^k$.