Анализ связей признаков при построении атаки на НКП

Иван Сухарев

September 2025

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 извлеченным парам можно отсортировать признаки по количеству их вхождений в метапризнаки. Для каждого i-го признака, $i = 1 \dots n$, определяются все пары (i, j) из списка выше. После чего формируется таблица:

- Номер i-го признака - столбец **feature**, все номера признаков, с которыми i-ый признак образует пару (i, j) перечислены в столбце **partners**;
- Список сортируется по убыванию количества связей с метапризнаками (столбец **pairs**);
- Каждая пара признаков считается уникальной, т.е. (i, j) и (j, i) - одна пара, и учитывается *ровно один раз*. «Владельцем» пары является признак с наибольшим числом уникальных партнёров;
- В столбце **neurons_involved** указываются номера нейронов («+» и «−» положительные и отрицательные нейроны соответственно), в которые входит пара с признаком из столбца **feature**, при этом у каждого нейрона в скобках указывается число вхождений данного признака в этот нейрон: `+5(3), -4(1)`. Количество неройнов записано в столбце **num**.

Таблица 1: Пример сводки по признакам для НКП[0]

feature	**pairs**	**partners**	**neurons_involved**	**num**
11	35	9, 22, 23, 53, 84, 90, 91, 141, 158, 165, 174, 234, 235, 240, 243, 244, 251, 253, 254, 292, 300, 312, 320, 339, 348, 370, 375, 459, 460, 463, 464, 486, 489, 490, 496	+5(1), +8(2), +9(4), +10(4), +11(4), +12(4), +13(4), +14(4), -3(4), -4(4)	10
12	20	119, 120, 122, 124, 211, 217, 221, 222, 226, 227, 228, 235, 370, 375, 378, 381, 459, 460, 464, 469	+15(4), +16(4), +17(4), -5(4), -6(4)	5

feature	pairs	partners	neurons_involved	num
17	20	118, 119, 120, 122, 193, 200, 202, 205, 211, 217, 221, 235, 375, 378, 381, 409, 459, 460, 463, 464	+20(4), +21(4), +22(4), -10(4), -11(4)	5
37	16	114, 117, 118, 119, 144, 145, 146, 152, 436, 441, 449, 450, 459, 460, 464, 466	+45(4), +46(4), +47(4), -21(4)	4
39	16	137, 151, 158, 165, 370, 375, 378, 381, 402, 409, 414, 419, 423, 425, 436, 443	+48(4), +49(4), +50(4), +51(4)	4
21	15	120, 122, 124, 125, 196, 200, 202, 205, 375, 378, 380, 381, 501, 506, 510	+27(4), +28(4), -15(4), -16(3)	4
89	13	34, 129, 133, 138, 141, 153, 155, 159, 160, 369, 373, 375, 376	+42(1), -35(4), -36(4), -37(4)	4
5	12	151, 158, 164, 165, 212, 218, 219, 226, 322, 323, 326, 327	+2(4), -0(4), -1(4)	3
9	12	146, 151, 152, 154, 217, 221, 235, 236, 498, 501, 506, 510	+5(1), +6(4), +7(4), -2(4)	4
15	12	60, 61, 71, 72, 205, 208, 210, 212, 497, 501, 505, 506	+18(4), -7(4), -8(4)	3
18	12	151, 158, 165, 166, 229, 231, 236, 242, 302, 303, 306, 308	+23(4), +24(4), -12(4)	3
20	12	117, 131, 137, 145, 452, 461, 467, 474, 495, 506, 509, 510	+25(4), +26(4), -14(4)	3
27	12	116, 121, 138, 155, 179, 185, 196, 201, 241, 247, 258, 260	+36(4), -17(4), -18(4)	3
58	12	197, 200, 204, 205, 310, 316, 318, 322, 501, 506, 509, 510	+62(4), +63(4), -25(4)	3
124	12	138, 141, 153, 155, 159, 160, 162, 170, 457, 461, 467, 470	+15(1), +27(1), -53(4), -54(4), -55(4)	5
34	11	49, 51, 55, 59, 79, 84, 85, 256, 257, 259, 270	+41(4), +42(4), +43(4)	3
23	8	92, 96, 98, 147, 234, 247, 274, 300	+8(1), +30(4), +31(4)	3
32	8	237, 241, 242, 249, 396, 408, 411, 413	+39(4), +40(4)	2
⋮	⋮	⋮	⋮	⋮

Таким образом, используя данную таблицу, можно сгруппировать признаки по убыванию количества образуемых пар. Далле признак, записанный в столбце **feature**, будем называть «родительским», а признаки из столбца **partners** - «дочерними».

Построение атаки

Для построения состязательной атаки на НКП в модели белого ящика, используя такую таблицу, можно предложить следующий общий вариант атаки, основанной на градиентном спуске:

1. Для атакующего изображения A, состоящего из признаков $a_1, \ldots, a_n$, по порядку фиксируем «родительские» признаки под номерами из столбца **feature**.

2. Для каждого «родительского» признака оцениваем количество связей **pairs**

каждого его «дочернего» признака с другими:

- Если количество связей больше некоторого минимального порога (например, среднего или медианного значения) и он является является «опорным» (участвует в 3 или 4 метапризнаках в нескольких нейронах), то данный «дочерний» признак изменять не будем. Одно изменение такого признака почти не повлияет на отклик нейрона y, так как y строится на метрике взвешенного среднеквадратичного отклонения метапризнаков, а сами метапризнаки являются разностями двух нормированных признаков.
- Если такой признак не является «опорным» (участвует в 1 метапризнаке, очень редко в 2 метапризнаках), но при этом у него достаточное количество **pairs**, тогда такой признак поддается изменению.
- Если признак не является «опорным», при этом у него малое количество «дочерних» признаков, то он тоже поддается измененению, но меньшему, чем такой же «неопорный» с бОльшим количеством **pairs**.

3. В связи с этим, метод сводится к большему изменению «неопорных» признаков с большим **pairs**, причем чем больше **pairs**, тем более широкому диапозону изменений может подвергаться данный признак.

4. Для каждого фиксированного «родительского» признака можем применить градиентный спуск на всю группу его «дочерних» признаков, учитывая критерий изменения «дочерних» признаков из п.2.

Далее сформулируем более конкретную задачу для состязательной атаки:

Постановка задачи:
Пусть у нарушителя есть «Свой» образ A и соответствующий ему НКП - NCT_A, соответственно ключ $key_A = NCT_A(A)$. Также у нарушителя есть атакауемый «Свой» образ B и соответствующий ему НКП - NCT_B, соответственно ключ $key_B = NCT_B(B)$.

Ожидаемый результат:
При подаче на вход NCT_B образа A' продуцируемый $key' = key_B$, т.е. $NCT_B(A') = key_B$ при минимальном изменении A': $||A - A'|| < \epsilon$, где $|| \cdot ||$ - некоторая метрика.

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 $\delta_1, \dots, \delta_n$, а также границ $(min_1, max_1), \dots, (min_n, max_n)$. 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 $\mathcal{A} = \{A^i\}_{i=1}^k$.