**Адъюнкт**
**Воронежского института МВД России**
**А.А. Толстых**

# АНАЛИЗ УСТОЙЧИВОСТИ ИСКУСТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ В ЗАДАЧАХ КЛАССИФИКАЦИИ ИЗОБРАЖЕНИЙ

В настоящее время проблема распознавания объектов на телевизионных изображениях решается посредствам применения аппарата искусственных нейронных сетей [1]. Однако, не смотря на успехи подобных методов, на сегодняшний день не существует аналитических правил выбора гиперпараметров нейросетевых классификаторов. Таким образом, с практической точки зрения, оценка эффективности построенного нейросетевого классификатора является одним из важнейших источников информации о путях дальнейшего совершенствования его архитектуры. Одной из составляющих оценки эффективности работы искусственной нейронной сети является оценка ее устойчивости. Следует отметить, что в исследовании, проведенном в [2] было показано, что эффективность сверточной нейронной сети в задачах классификации изображений может быть сильно снижена благодаря замене одного пикселя исходного изображения. Исходя из этого можно заключить, что задача оценки устойчивости нейросетевых классификаторов изображений является актуальной и требует разработки.

С математической точки зрения устойчивость искусственной нейронной сети эквивалентна малым изменениям выходного параметра при малых изменениях входных параметров. Таким образом, речь идет о выполнении неравенства [2]:

$$\left|F(X)-F(X')\right| \le A\left|X-X'\right|, \quad (1)$$

где $X, X'$ – входные параметры (изображение и его искаженная копия); $F(X)$, $F(X')$ – выходные параметры; $A$ – постоянная, имеющая значение оценки уровня устойчивости. При использовании модели искусственной нейронной сети с гладкими передаточными функциями нейронов и обучение её алгоритмом обратного распространения ошибки выходная функция нейронной сети, при любых значениях весов, является бесконечно дифференцируемой.

Рассмотрим алгоритм оценки устойчивости, приведенный в [3]. Пусть $(X^i,Y^i)$, $i \in 1,2,\ldots,N$ – таблица исходных статистических данных. Для каждого фиксированного индекса $I$ находится вектор $X^j$, ближайший к $X^i$ относительно евклидовой метрики:

$$\min_{k \ne i}\left|X^i - X^k\right| = \left|X^i - X^j\right|, \quad (2)$$

где $|X^i–X^k|$ – евклидово расстояние, номер $j$ зависит от $i$. Пусть $\varepsilon_i = |X^i – X^j|$. Положим:

$$K_i = \frac{\left|Y^i - Y^{j(i)}\right|}{\varepsilon}. \quad (3)$$

Для каждого $i$ выберем случайный единичный вектор $\omega_i$. Введём обозначения [4]:

$$\begin{aligned} X^i[\varepsilon_i, \omega_i] &= X^i + \varepsilon_i \omega_i, \\ \widehat{Y}^i &= F\left(X^i\right), \\ \widehat{Y}^i &= F\left(X^i[\varepsilon_i, \omega_i]\right), \\ \widehat{K}_i &= \frac{\left|\widehat{Y}^i - \widehat{Y}^i[\varepsilon_i, \omega_i]\right|}{\varepsilon_i}. \end{aligned} \quad (4)$$

Точки $\widehat{X}^i[\varepsilon_i, \omega_i]$ и $X^{j(i)}$ отстоят от точки $X^i$ на равном расстоянии $\varepsilon_i$, поэтому величины $K_i$ и $\widehat{K}_i$, задаваемые соответственно формулами (3) и (4) должны быть в каком-то смысле близки, т.е. их отношение должно быть близко к 1. Тогда набор чисел:

$$Z_i = \frac{\widehat{K}_i}{K_i}, \quad (5)$$

можно считать набором наблюдаемых значений величины Z. Устойчивость нейронносетевой модели $Y=F(X)$ будем считать соответствующей устойчивости исходных данных, если распределение случайной величины $\ln(Z)$ близко к нормальному распределению с нулевым математическим ожиданием (т.е. распределение величины $Z$ является логнормальным) [5]. При этом возможна ситуация, когда $E(\ln(Z)) \le 0$. Это означает, что нейронносетевая модель более устойчива, чем исходные данные. В качестве обобщающего показателя, сравнительной устойчивости нейронносетевой модели, может использоваться вероятность $P(\ln(Z))$, т.е. вероятность того, что устойчивость модели будет не ниже, чем устойчивость исходных данных.

Таким образом, рекомендации по оценке соответствия уровня устойчивости нейронносетевой модели заключаются в следующей схеме:

а) Построение эмпирического распределения величины $\ln(Z)$.

б) Проверка гипотезы о нормальности распределения величины $\ln(Z)$.

в) Оценка вероятности $P(\ln(Z)<0)$.

Анализ литературы позволил сформировать обобщенный алгоритм оценки устойчивости нейросетевого классификатора при решении задач классификации изображений. Оценка, получаемая с помощью данного алгоритма, позволяет рассматривать эффективность классификатора при малом изменении входных параметров. Полученная оценка может использоваться как ограничение, налагаемое при обучении искусственной нейронной сети, а также в качестве метрики, определяющей направление оптимизационного процесса.

## ЛИТЕРАТУРА

1. Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение. Москва: «ДМК Пресс», 2017. 652 с.

2. Su J., Vargas D.V., Kouichi S. One pixel attack for fooling deep neural networks // arXiv.org e-Print archive. 2017. URL: https://arxiv.org/abs/1710.08864 (дата обращения: 15.05.2018).

3. Горбань А.Н., Россиев Д.А. Нейронные сети на персональном компьютере. Новосибирск: Наука, 1996. 278 с.

4. Царегородцев В.Г. Определение оптимального размера нейросети обратного распротранения через сопоставление средних значений модулей весов синапсов // Материалы 14 международной конференции по нейрокибернетике. Ростов-на-Дону. 2005. Т. 2. С. 56-62.

5. Хей Д. Введение в методы байесовского статистического вывода. Учебное пособие. Москва: Финансы и статистика, 1987. 336 с.

6. Вьюгин В.В. Математические основы теории машинного обучения и прогнозирования. Москва. 2013. 387 с.