Алгоритм экстракции образующих признаков

Н. И. Кавонкин[1], О. Ю. Лукомская[1,2], А. Л. Стариченков[1]

[1] *Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)*

[2] *Институт проблем транспорта им. Н.С. Соломенко Российской академии наук*

user35@mail.ru, luol@mail.ru, allstar72@mai.ru

Аннотация. **В данной статье предложен алгоритм экстракции «зон интереса» сверточных нейронных сетей, который осуществляется следующим образом: изображение распознаётся для получения общей вероятности отнесения объекта к целевому классу, затем фрагменты изображения перекрываются плавающей маской для получения частных исходов вероятности, значения нормируются и суммируются в общую карту. Пример работы алгоритма экстракции образующих принципов рассмотрен для анализа изображений, полученных с водной инфраструктуры. Результаты исследований и моделирования могут стать основой для улучшения компактных нейросетевых алгоритмов, использоваться в автоматизации процесса подбора специфических признаков.**

Ключевые слова: искусственные нейронные сети; изображения; алгоритмы машинного обучения

I. Введение

Нейросетевые алгоритмы называют «глубокими» относительно описанных Френком Розенблаттом перцептронов. Глубиной называют наличие множества скрытых слоёв. В современных алгоритмах эти слои, обеспечивая особые свойства и результативность модели [1], ввиду своей нелинейной природы, значительно усложняют интерпретацию содержащейся в модели информации. Для задач машинного обучения зачастую необходимо понимать, какие элементы на изображении являются наиболее важными признаками для отнесения объекта к тому или иному классу, это позволяет оптимизировать модели глубокого обучения, составлять многокомпонентные ансамбли из моделей особо чувствительных к данному типу признаков.

Ранее, например, в [2] был использован алгоритм распознавания проселочных дорог, опирающийся на иерархический метод извлечения контуров городских дорог в двоичном дереве разделов и, затем, извлечение контуров дорог на основе метода на иерархических уровнях. В работе применялся свёрточный автоэнкодер на архитектуре U-net, для подготовки данных использовались пороговые преобразования и автоматическая аугментация данных за счёт афинных преобразований.

Объектом исследования в данной работе являются нейросетевые алгоритмы, которые рассматриваются как детекторы, активирующиеся по совокупности распознанных на изображении признаков.

II. Математическое и алгоритмическое описание

На примере машин опорных векторов и наивного байесовского классификатора в [3] показано, что большинство классификаторов, обучаемых с учителем, выдают оценки S принадлежности представленных им объектов $\boldsymbol{X_к}$ классу $\boldsymbol{C}$.

Если нам заранее известны метки, то $P(C|X)=1$ для экземпляров, отмеченных меткой класса C и, $P(C|X)=0$ – для остальных. Тогда, при условии что модель обучена верно, не происходит переобучения, оценки классификации $S(X)$ могут быть использованы для определения вероятности отнесения объекта к классу [3]. Допустим, мы обучаем нейросеть распознавать дороги на спутниковых снимках, тогда на примере простой бинарной классификации это означало бы, что если пример относится к обучающему набору данных и для него есть истинная метка, то если на изображении есть дорога (метка C=1), то P(C|X) = 1. Аналогично если дороги нет (метка C=0), то P(C|X) = 0.

На примере наивного байесовского классификатора становится очевидно, что если для двух объектов Y и X:

$$S(X) < S(Y), mo \quad P(X\ |C) < P(Y\ |C) \qquad (1)$$

При распознавании неразмеченных примеров, используемый алгоритм опираясь на заложенные в него метрики практически делает вывод о подобии представленной ему сцены и изученных примеров обучающего набора. Для нейросетевых алгоритмов это также выражается в количестве и энергии активации нейронов в выходном и близким к ним слоях.

Свойство (1), при учёте того, что целью обучения является минимизация перекрёстной энтропии, позволяет использовать оценки принадлежности к классу, как грубое приближение вероятностей отнесения объектов $X_к$ классу C. Это приближение тем хуже отражает вероятность, чем неравномернее представлены данные в обучающем наборе [3].

Для нормирования карты областей интереса алгоритм использует свойство *маргинального распределения:*

$$\forall x \in X, P(X = x) = \sum_a P(X = x, A = a) \qquad (2)$$

где в качестве первой случайной величины рассматривается вероятность отнесения объекта к классу, а второй является вероятность отнесения объекта к классу с учётом нанесённой маски.

На рис. 1 представлена блок-схема работы алгоритма. Первым шагом в обученный детектор подаётся исходное изображение как набор пикселей с координатами X, Y и цветом (RGB), по результату распознания собираются его мажорные классы и значение функции принадлежности.

Затем, циклически, с точностью до пикселя X_i, Y_i, на поверхность исходного изображения наносится перемещающаяся блокирующая маска (уменьшающегося размера, переменной формы) размером X_m, Y_m. Применение масок различной геометрии из коллекции правил размеров и размещения G_t, (квадрат/круг/ прямоугольник/ *) позволяет определять на изображении зоны интереса различной формы.

Для каждого положения маски и заблокированной ей области строится карта значения функции принадлежности, полученные значения накапливаются в тензоре. Накопленные данные нормализуются одним из следующих способов: деление на число проходов цикла через пиксель, / деление на константу, * деление на среднее). Полученная в результате карта отражает области наивысшего интереса для используемого алгоритма и может применяться в процессе настройки модели.

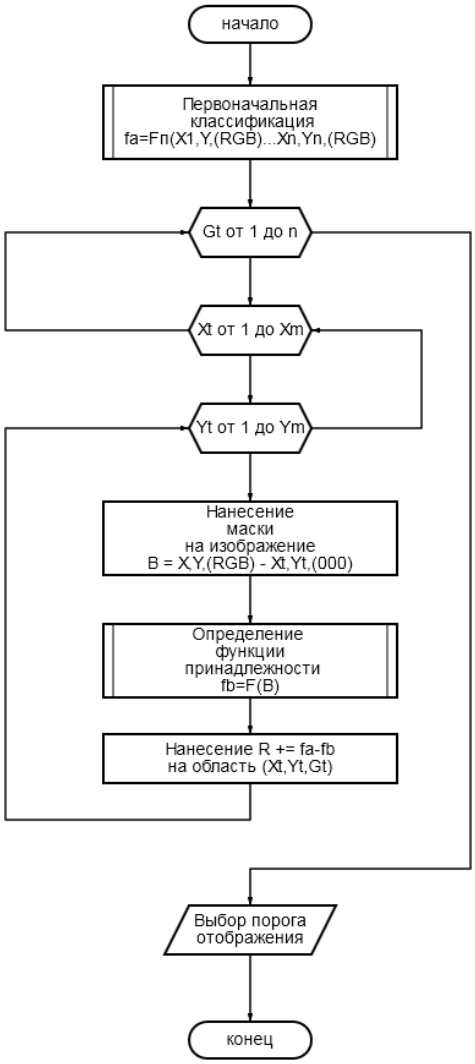

Рис. 1. Блок схема алгоритма построения карт

III. Калибровка вероятности

Свойство (1) сохраняется в рамках бинарной классификации, к которым можно свести мультиклассовую задачу [3] (по принципу мажорный класс против остальных), однако настоятельно рекомендуется откалибровать вероятности для сохранения интерпретируемости результатов.

В работе [5] представлено масштабное исследование, посвященное анализу калибровки вероятностей в условиях реалистичных сдвигов распределения данных для задач классификации изображений. Авторы провели всесторонний анализ различных методов калибровки, включая постобработку (post-hoc) и методы, применяемые во время обучения (in-training), а также их взаимодействие. Основная цель работы заключалась не в предложении новых методов, а в выявлении практических рекомендаций для достижения надежной калибровки в условиях сдвигов.

Исследование [5] показало, что применение методов постобработки (например, temperature scaling или energy-based calibration) делает дополнительные методы in-training, такие как сглаживание меток (label smoothing) или регуляризация энтропии (entropy regularisation), избыточными. Однако если post-hoc калибровка не используется, комбинация регуляризации энтропии и сглаживания меток обеспечивает наилучшую калибровку исходных вероятностей на новых данных.

Добавление небольшого количества семантически несвязанных (out-of-distribution, OOD) данных (в качестве негативных примеров) на этапе калибровки значительно улучшает устойчивость калибровки на этапе валидации. Этот эффект наблюдается для всех post-hoc методов, включая простой метод температурного градиента, который показал сопоставимую или даже лучшую эффективность по сравнению с более сложными методами, такими как energy-based scaling (EBS). Важно отметить, что введение негативных примеров не требует их близости к основному распределению задачи, что делает этот подход универсальным. Важно отметить, что улучшение калибровки на новых данных часто сопровождается ухудшением калибровки на исходном распределении.

Как показано в [5] современные методы, такие как Density-Aware Calibration (DAC) или energy-based scaling (EBS), не всегда превосходят более простые подходы, такие как temperature scaling, особенно при работе с семантически несвязанными данными. Это ставит под сомнение необходимость использования сложных методов в практических сценариях.

Отдельно стоит отметить, что ансамблирование моделей демонстрирует хорошие результаты в задаче улучшения калибровки как на обучающих, так и на новых данных. При этом в [5] показано что наибольший эффект достигается при применении post-hoc калибровки не к целому ансамблю, а к отдельным моделям до объединения. Интересно, что для ансамблей использование негативных примеров на этапе калибровки ухудшает результаты, что контрастирует с выводами для одиночных моделей.

Как показано в [5] предобученные модели, показали значительно лучшую калибровку как на исходном распределении, так и на новых данных, по сравнению с моделями, обученными с нуля. Это преимущество сохраняется независимо от выбранного метода калибровки.

В данной статье к дообученной модели на основе ResNet применялась post-hoc калибровка методом градиента температур. Архитектура модели изображена на рис. 2.

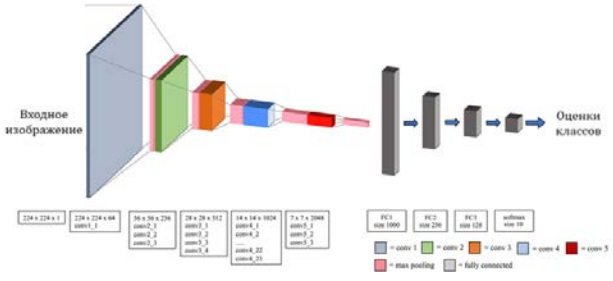

Рис. 2. Архитектура сети используемой в эксперименте

На рис. 3 и 4 приведены калибровочные диаграммы до и после калибровки сети методом градиента температур соответственно.

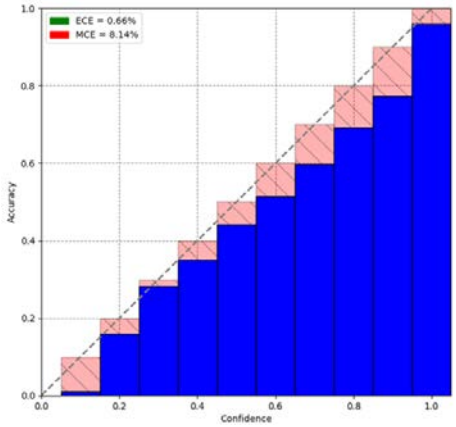

Рис. 3. Диаграмма доверия предобученной модели по классу «лодка»

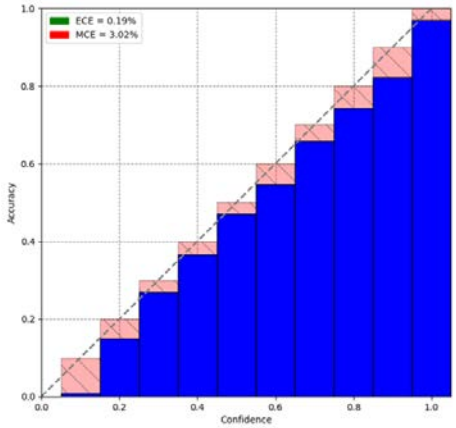

Рис. 4. Результаты калибровки вероятностей по классу «лодка» методом температур

IV. Пример работы алгоритма

На рис. 5 показаны этапы работы алгоритма:

2а (верхний левый) — перекрытие маской изображения с мажорными классами «лодка» – 0.9237…, «лодка весельная» – 0,8704…;

2б (нижний левый) — разность результата распознавания изображения, перекрытого маской, и оригинала по классу «лодка» – 0.9237 – 0.9150 = 0.0087, наложенного на область маски;

2в, 2д — маски различной формы перемещаются по пространству изображения;

2г — промежуточный результат работы алгоритма для нескольких масок;

2е — нормированная карта интереса с порогом отсечения равным 0.001.

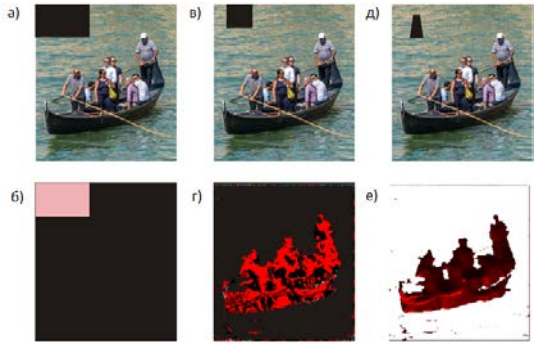

Рис. 5. Графическое представление работы алгоритма

Закрытая чёрной маской область снижает вероятность отнесения объекта к классу, когда перекрывает собой один или несколько существенных для класса признаков. Эта информация наносится на карту признаков, отображаемую в виде цвета. Перемещая маску по изображению, получается слой карты для маски *t*. Перед завершением алгоритм суммирует карты и отображает нормированную информацию.

V. Заключение

Как показано в [7] для интерпретации данных классификаторов полезно использовать калиброванную апостериорную вероятность отнесения объекта к классу, для этого многие современные алгоритмы используют перекрёстную энтропию в качестве целевой функции.

В работе мы показали, что это свойство удобно использовать для построения карт областей интереса алгоритмов нейронных сетей.

Работоспособность алгоритма сохраняется и в условиях некалиброванных оценок принадлежности, однако затрудняет интерпретацию результатов.

Дальнейшим развитием темы является исследование возможности применения полученных данных для обобщения и разложения модели на отдельные специфичные к признакам детекторы. Потенциально ансамбли из таких детекторов могут быть решением для низкопроизводительных бортовых вычислителей (высокая степень распараллеливания вычислений, возможность отключать отдельные детекторы для достижения минимально необходимых показателей качества распознавания).

Список литературы

[1] Макаренко А.В. Глубокие нейронные сети: зарождение, становление, современное состояние. Текст: электронный // Проблемы управления. 2020. Т. 2. С. 3-19.

[2] Lukomskaya O.Y., Seliverstov Y.A. On the application of neural network technologies for control problems in cognitive transportation systems Journal of Physics: Conference Series. 13th Multiconference on Control Problems, MCCP 2020" 2021. С. 012016.

[3] B. Zadrozny and C. Elkan, “Transforming classifier scores into accurate multiclass probability estimates,” *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 694–699, 2002, doi: 10.1145/775047.775151.

[4] Wahba G. Multivariate function and operator estimation, based on smoothing splines and reproducing kernels Текст: электронный // Nonlinear Modeling and Forecasting. 1992. Т. 12. С. 95-112. – URL: https://api.semanticscholar.org/CorpusID: 125709427 (дата обращения: 02.04.2024).

[5] M. Roschewitz, R. Mehta, F. de S. Ribeiro, and B. Glocker, “Where are we with calibration under dataset shift in image classification?” Jul. 2025, Accessed: Jul. 10, 2025. [Online]. Available: http://arxiv.org/abs/2507.07780

[6] Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods / J. Platt, others // Advances in large margin classifiers. 1999. Т. 10. No 3. С. 61-74.

[7] Mojtaba Faramarzi, Mohammad Amini, Akilesh Badrinaaraayanan, Vikas Verma, Sarath Chandar. PatchUp: A Feature-Space Block-Level Regularization Technique for Convolutional Neural Networks. Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022, 36, 589–597. https://doi.org/10.1609/aaai.v36i1.19938.