Вычисление степени уверенности предсказания нейронной сети

А.С. Коваленко[1], Я.М. Демяненко[1]

[1]Институт математики механики и компьютерных наук имени И. И. Воровича, Мильчакова 8а, Ростов-на-Дону, Россия, 344090

Аннотация. При работе с глубокими нейронными сетями зачастую возникает необходимость анализа уверенности сделанного моделью предсказания. В частности, данная задача имеет высокую значимость при анализе изображений, поскольку некоторые классы, соответствующие изображению, поданному на вход модели, могут отсутствовать в обучающей выборке. В таких случаях модель возвращает заведомо некорректный результат. В данной работе предлагается решение описанной проблемы при помощи специальной архитектуры нейронной сети, позволяющей совместно с предсказанием возвращать и степень уверенности в данном предсказании. Также данная архитектура позволит автоматизировать процесс анализа медицинских данных. В случае низкой уверенности в ответе сети на входном изображении, данное изображение направляется на рассмотрение специалисту и, возможно, попадает в дообучающую выборку. Помимо прочего, предлагаемый подход возможно использовать для поиска аномалий в анализируемых данных.

1. Введение

При решении задачи классификации [1] зачастую используются нейронные сети (общая схема архитектуры 1). Ответом нейронной сети на входное изображение, как правило, является вектор плотности вероятностей распределения по классам (пример изображён на рисунке 2) и класс определяется как индекс максимального аргумента данного распределения. Но если подать на вход сети изображение не относящееся к набору данных, на котором обучалась нейронная сеть, то она, как и для любого входного изображения, построит выходной вектор. Если рассмотреть нейронную сеть, обученную на наборе данных ”MNIST” [2], и подать на вход изображение, сформированное из равномерного шума, то она даст следующий вектор, изображённый на рисунке 3. По нему предсказанным классом является цифра 2, причём с большим показателем вероятности. Существует ряд задач, в которых необходимо отсекать такие предсказания. Для решения рассматриваемой проблемы существуют подходы, применяемые во время обучения сети. При работе с уже обученной нейронной сетью возникает необходимость в её переобучении. В этой работе рассматривается подход для вычисления уверенности предсказания предобученной нейронной сети.

2. Существующие решения

Самым распространённым подходом является анализ выдаваемого сетью распределения. В работе [3] авторы предлагают подход для вычисления порога определения верного класса. Но как показывает пример с равномерным шумом, данный метод может не справиться, так как вероятность предсказанного класса равна 1.0.

Рисунок 1. Общая схема нейронной сети для задачи классификации.

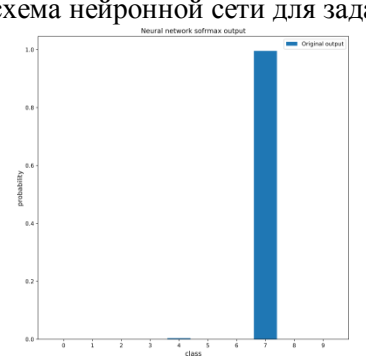

Рисунок 2. Вектор плотности вероятностей распределения по классам для изображения из набора данных MNIST.

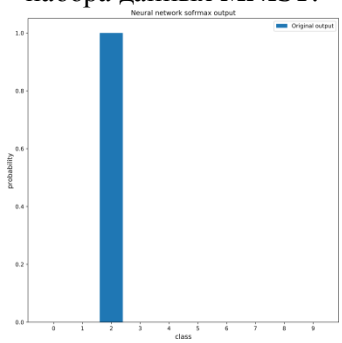

Рисунок 3. Вектор плотности вероятностей распределения по классам для изображения, построенного из равномерного распределения.

При рассмотрении методов, требующих обучения классифицирующей нейронной сети, можно выделить два основных подхода. Первый заключается в использовании специальных функций ошибок для выделения уверенности предсказания в отдельную компоненту ошибки, как делают авторы следующей работы (Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples) [4]. Другой подход основан на изменении архитектуры классифицирующей модели для возможности введения способа вычисления уверенности сети. Обзор таких методов проводят авторы статьи (Confidence Regularized Self-Training) [5]. Один из недостатков, указанных выше подходов, заключается в том, что для их реализации необходим размеченный набор данных. В этой работе рассматривается подход, требующий наличия только самих данных, на которых нейронная сеть должна делать предсказания.

3. Предлагаемая архитектура

Свёрточные слои нейронной сети, предшествующие полносвязным, извлекают высокоуровневые признаки изображения, на основе которых производится предсказание следующими слоями сети. На выходных сигналах данных свёрточных слоёв построен и обучен декодер, восстанавливающий сигнал поданного на вход сети изображения (общая схема предлагаемой архитектуры изображена на рисунке 4). Восстанавливающий декодер обучался на наборе данных MNIST, на котором уже была предобучена классифицирующая модель. Свёрточные слои классифицирующей сети, параметры которых не участвуют в обучении, выступают в качестве энкодера. Далее восстановленное изображение снова подаётся на вход классифицирующей сети, и, таким образом, получаются два распределения вероятности предсказания класса. Если на вход сети подается изображение из распределения, отлично- го от распределения обучающей выборки, сеть извлекает некорректные высокоуровневые признаки,

которые затем декодируются в некоррекный сигнал. При передаче такого восстановленного сигнала на вход сети получаются отличные друг от друга распределения вероятностей. При подаче изображения принадлежащего распределению, на котором была обучена нейронная сеть, свёрточные слои извлекут те же правильные признаки из сигнала восстановленного декодером. Для классифицирующих слоёв выходные распределения оригинального и восстановленного сигналов будут близки друг к другу.

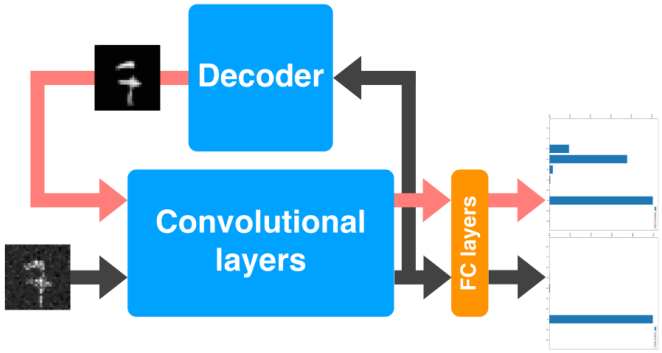

Рисунок 4. Общая схем предлагаемой архитектуры для восстановления входного сигнала по высокоуровневым признакам.

4. Параметр уверенности предсказания

Для вычисления значения уверенности предсказания сети вводится следующая мера:

$$confidence(y^1, y^2) = 1 - \frac{\sum_{i=0}^{i<10} \|y_i^1 - y_i^2\|}{2} \quad (1)$$

где y^1 – результат работы сети на входном изображении, y^2 – результат работы сети на восстановленном с помощью декодера изображении.

Результат функции интерпретируется как вероятность корректного ответа сети на переданное изображение. Данная мера имеет следующее свойство:

$$\forall y^1, y^2 \Rightarrow conf\,idence(y^1, y^2) \in [0, 1]$$

Количество ошибочных предсказаний сети можно регулировать при помощи порога, отбрасывания те изображения, для которых введенная мера не превосходит порога.

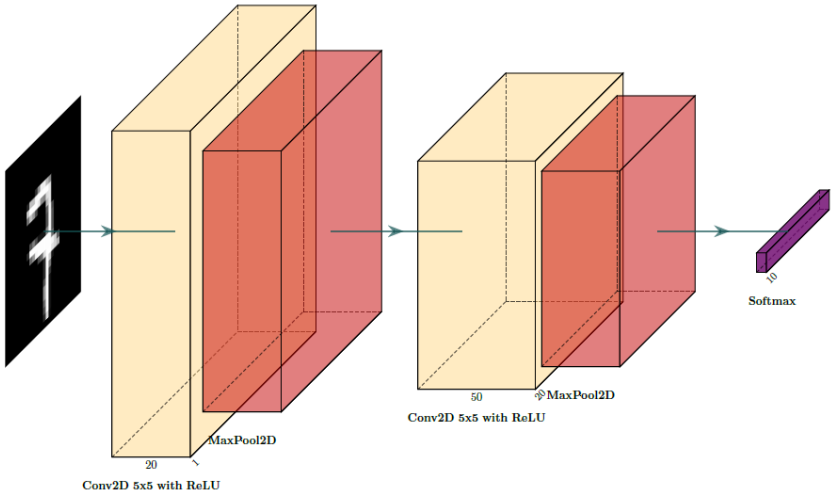

Рисунок 5. Детальная архитектура классифицирующей нейронной сети для задачи MNIST.

5. Эксперименты

Для классифицирующей модели была построена нейронная сеть со следующей архитектурой, изображённой на схеме 5. В качестве декодера выступала архитектура нейронной сети, изображённая на схеме 6. Для обучения и тестирования моделей использовался набор

изображений: “MNIST” http://yann.lecun.com/exdb/mnist/, содержащий коллекцию изображений рукописных цифр. Пример данных изображен на рисунке 7.

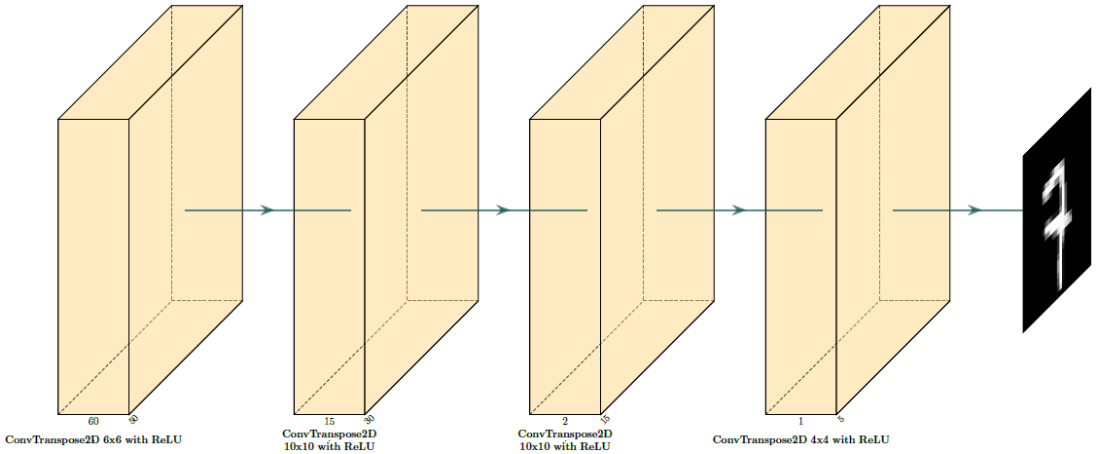

Рисунок 6. Детальная архитектура декодирующей нейронной сети для классификатора 5.

Рисунок 7. Примеры MNIST.

6. Результаты

Для анализа эффективности предлагаемого подхода построен график зависимости значения средней тончости, полученной в результате работы сети на тестовой выборке, от выбранного порога уверенности изображён на рисунке 8. Для детального анализа предсказаний построены матрицы ошибок по результатам работы сети на тестовом наборе данных до применения фильтрации по порогу уровня уверенности, рисунок 9) и после, рисунок 10. После исключения из рассмотрения результатов со значением уверенности, меньшей 0.9, средняя точность увеличивается с 0.9811 до 0.9991.

При увеличении порога уверенности для фильтрации из рассмотрения отбрасывается относительно небольшая часть данных. Зависимость доли нерассмотренных данных от порога изображена на графике 11.

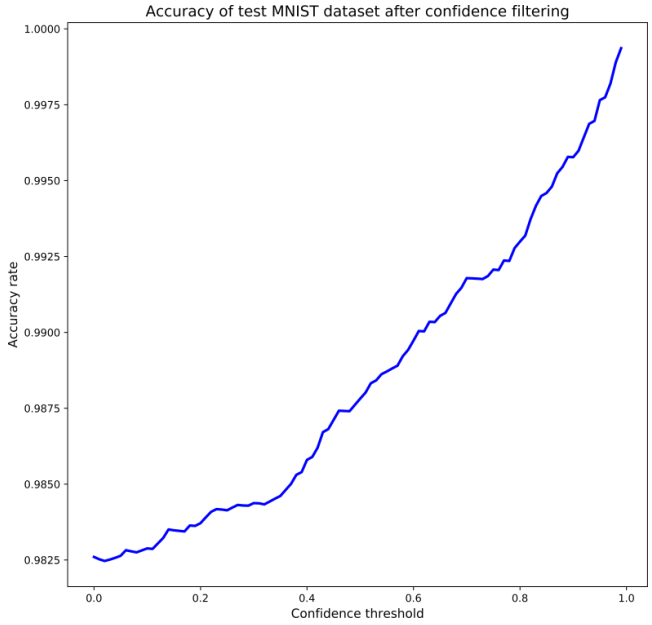

Рисунок 8. График зависимости значения средней точности от порога уверенности.

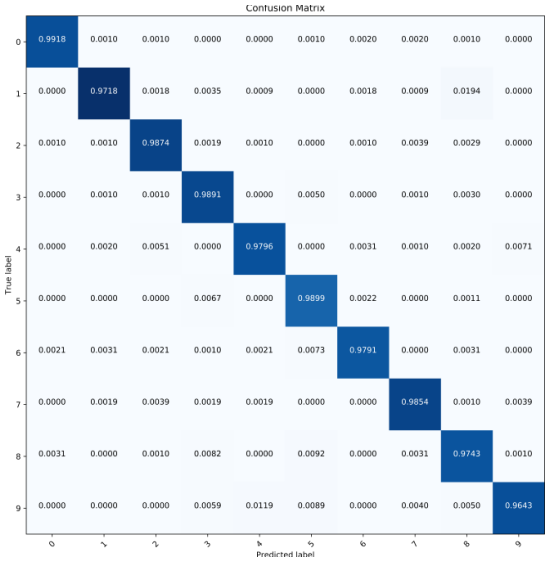

Рисунок 9. Матрица ошибок, построенная на результатах применения сети на тестовом наборе данных MNIST.

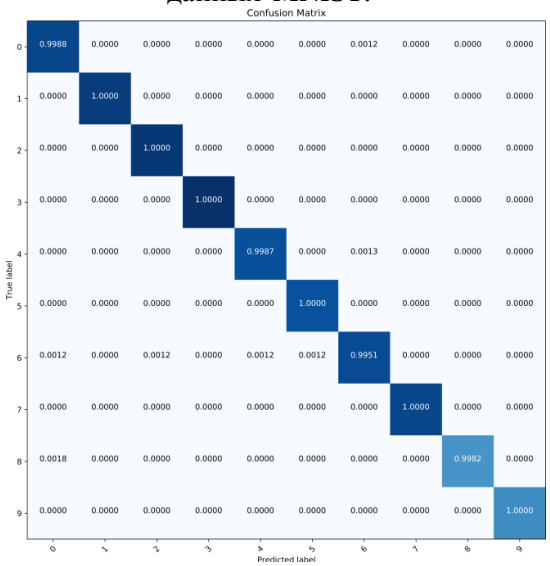

Рисунок 10. Матрица ошибок, построенная на результатах применения сети на тестовом наборе данных MNIST с фильтрацией по порогу уверенности 0.9.

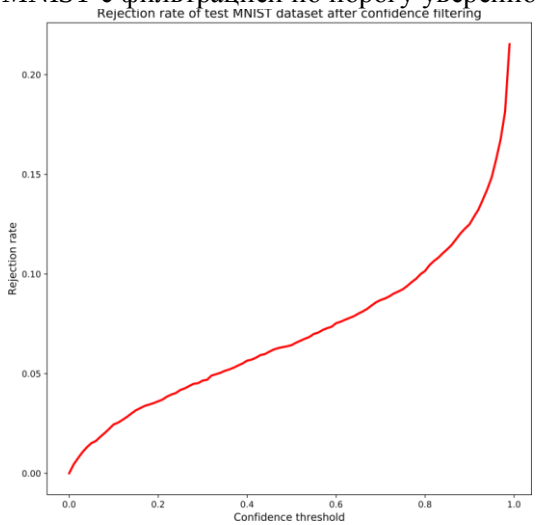

Рисунок 11. Зависимость доли нерассмотренных данных от значения уверенности.

7. Заключение

Предлагаемый подход позволяет повысить качество решения задачи классификации при помощи нейросетей без необходимости переобучения уже обученных моделей. Так как данный подход для обучения декодера не требует размеченных данных, он применим в ситуациях, когда имеется только предобученная модель и набор изображений из распределения, на котором обучалась нейронная сеть.

Рассматривая задачу классификации рукописных цифр из набора данных ”MNIST”, можно увеличить среднюю точность классификации с 0.9811 до 0.9991, не рассматривая 12% данных из набора, на которых сеть не способна предсказать верный класс.

8. Приложения

Ссылка на репозиторий с реализацией: https://github.com/AlexeySrus/neural-network-confidence.

9. Литература

[1] Kotsiantis, S.B. Supervised Machine Learning: A Review of Classification Techniques – Greece: Department of Computer Science and Technology University of Peloponnese, 2007.
[2] Wu, H. CNN-Based Recognition of Handwritten Digits in MNIST Database // Research School of Computer Science – The Australia National University, Canberra.
[3] Chen, J. The Use of Decision Threshold Adjustment in Classification for Cancer Prediction / J. Chen, C. Tsai, H. Moon [Electronic resource]. – Access mode: https://www.academia.edu /24322193 (25.12.2019).
[4] Lee, K. Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples / K. Lee, H. Lee, K. Lee [Electronic resource]. – Access mode: https://arxiv.org/abs/1711.09325 (25.12.2019).
[5] Zou, Y. Confidence Regularized Self-Training / Y. Zou, Z. Yu, X. Liu [Electronic resource]. – Access mode: https://arxiv.org/abs/1908.09822 (25.12.2019).

Neural network confidence model

A.S. Kovalenko[1], Y.M. Demyanenko[1]

[1]Institute of mathematics, mechanics and computer Sciences named after I. I. Vorovich, Milchakova street 8a, Rostov-on-Don, Russia, 344090

Abstract. Often there is a need for an analysis of the confidence made by the neural model. This problem is of high importance in the analysis of images, because the classes of images supplied to the network may be absent in the training set and be perceived by the network, with a deliberately incorrect answer. The solution of such a problem is considered in this work with the help of a special neural network architecture, which allows, together with the prediction, to return a degree of confidence in this prediction. Also, this approach allows semi-automating the process of diagnostics and recommendations to a medical specialist on the medical data under consideration. In the case of an uncertain response from the network, the processed example is submitted to a specialist for review and goes to the pre-training sample. Also with the help of this approach, we can screen anomalies in the analyzed data.