УДК 004.852

МИКРОМОДЕЛЬ МЕХАНИЗМА ПРИНЯТИЯ РЕШЕНИЯ ЭЛЕМЕНТАРНОЙ СВЕРТОЧНОЙ НЕЙРОСЕТЬЮ В ВЫРОЖДЕННОЙ ЗАДАЧЕ КЛАССИФИКАЦИИ ОБЪЕКТОВ НА ИЗОБРАЖЕНИИ

Н.Л. Андрейчик

Институт проблем управления им. В.А. Трапезникова РАН

Россия, 117997, Москва, Профсоюзная ул., 65

E-mail: kolya.andreychik@gmail.com

А.В. Макаренко

Институт проблем управления им. В.А. Трапезникова РАН

Россия, 117997, Москва, Профсоюзная ул., 65

E-mail: avm.science@mail.ru

Ключевые слова: сверточные нейронные сети, классификация, микромодель, механизм принятия решения, устойчивость функционирования.

Аннотация: В работе рассмотрена задача бинарной классификации объектов на черно-белом изображении. Классы объектов представляют собой равносторонние однотонные треугольники на однотонном фоне, фиксированные по положению, размеру, и ориентации. Один из классов ориентирован вершиной вверх, другой – вершиной вниз. Задача решена элементарной сверточной сетью: один скрытый сверточный слой с одним ядром размером 1×1 и полносвязный выходной слой с единственным нейроном с сигмоидальной функцией активации. Для рассмотренного случая построена и изучена микромодель механизма принятия решения нейросетью, в том числе при функционировании вне домена обучающих данных. Проанализировано влияние параметров начальной инициализации параметров сети на ее способность к обучению и предрасположенности к ложным срабатываниям, в том числе «галлюцинациям». Предложен подход к повышению устойчивости функционирования нейросети во внедоменной области.

1. Введение

В настоящий момент нейросетевые модели представляют по своей сути «черный ящик», то есть в достаточной мере обученная модель с определенной точностью будет справляться с возложенной на нее задачей, однако узнать внутреннюю логику принятия того или иного решения не предоставляются возможным. В современных работах осуществляются попытки приоткрыть «черный ящик» и сделать нейросетевые модели более интерпретируемыми [1]. Однако на текущем этапе развития нейросетевых технологий нельзя говорить о создании полностью

интерпретируемых глубоких нейросетевых моделей из-за их высокой сложности. По этой причине весьма важно строить микромодели и изучать механизмы принятия решений элементарными нейросетевыми моделями, которые для своей работы используют только какой-то один ключевой механизм извлечения и обработки данных, например, сверточный механизм [2] или механизм внимания [3]. Такой подход позволяет более наглядно пронаблюдать внутреннюю логику принятия решения тем или иным нейросетевым алгоритмом. Последующие построения семейств таких внутренних микромоделей должны позволить экстраполировать полученные результаты для интерпретации более сложных сетей, в которых используется множество других различных структурных элементов.

Второй проблематикой нейросетевых моделей является проблема устойчивости их функционировании во внедоменной области [4]. Так если в обучающем наборе данных в задаче классификации у экземпляров определенного класса будет присутствовать признак, который в реальном мире не полностью подходит для классификации, но в контексте обучающего набора достаточен, то зачастую модель «запоминает» именно этот признак как ключевой. При дальнейшем использовании модели обученной на таком наборе данных для реальных задач может возникнуть эффект «галлюционирования» модели, при котором будет принято неверное решение об итоговом классе. В данной работе изучается микромодель механизма принятия решения простейшей сверточной нейронной сетью в контексте двух вопросов: интерпретируемости и внедоменной устойчивости функционирования.

2. Архитектура модели, набор данных и обучение нейросети

Используемая архитектура нейросетевой модели включает в свой состав: один скрытый сверточный слой с размером ядра 1×1, линейный выходной слой с единственным нейроном и сигмоидальной функцией активации, расположенный сразу на выходе сверточного слоя. Выходом модели является выход сигмоидальной функции активации, а сам результирующий класс определяется по пороговому правилу от значения сигмоиды. Значение порога, в первом приближении, принято равным 0.5.

Обучающий набор данных состоит из двух изображений треугольников различной ориентации и радиусом описанной окружности 50px. Размер изображения 256×256 пикселей, палитра черно-белая. Стоит отметить, что изображения треугольников строились таким образом, чтобы центр масс закрашенных областей совпадал и находился строго по центру изображения ($x = 128$, $y = 128$), а само количество закрашенных пикселей совпадало.

Во время обучения было выявлено сильное влияние начального распределения весов сети на ее дальнейшие характеристики. Так для инициализации весов линейного слоя использовался равномерный закон распределения (xavier), а для инициализации весов сверточного слоя – нормальный закон распределения с математическим ожиданием a и дисперсией 0.03. Производилось обучение 100 моделей в течении 100 эпох с экспоненциальным уменьшением скорости обучения по закону: $lr = 0.9^i$, где i – номер эпохи обучения, оптимизатор: SGD. График зависимости количества моделей, которые удалось/не удалось обучить

с определенной точностью по F1 метрике от математического ожидания начального распределения значений весов представлен на рисунке 1.

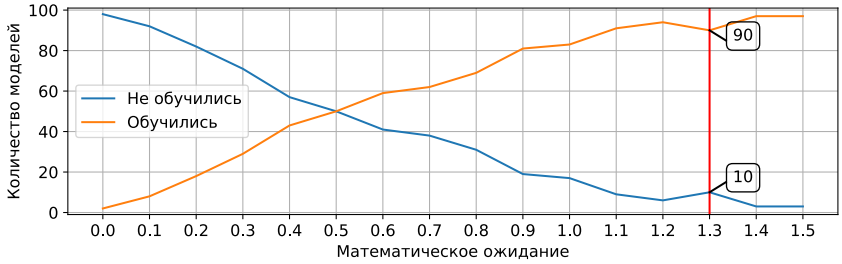

Рис. 1. Зависимость количества успешно/не успешно обученных моделей от математического ожидания начального распределения весов

Для дальнейших экспериментов производилось обучение 30-ти моделей на ранее описанном датасете аналогичным образом и с зафиксированными параметрами начального распределения значений весов сверточного слоя: математическое ожидание $a = 1.3$, дисперсия $b = 0.03$, так как при таких параметрах 90 из 100 моделей обучались успешно. Далее, для облегчения демонстрации полученных результатов, анализу будут подвергнуты только три наиболее показательные модели, так как остальные модели имеют схожие с ними характеристики.

3. Интерпретируемость обученных моделей

С целью анализа механизма принятия решений ранее обученными моделями были визуализированы веса выходного линейного слоя (см. рис. 2). Веса линейного слоя интересны по той причине, что из-за наличия взаимосвязей в свой структуре именно линейный слой осуществляет семантическую обработку входных данных.

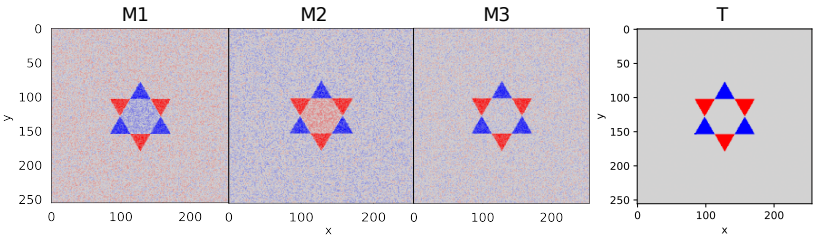

Рис. 2. Значения весов выходного линейного слоя. Красный цвет соответствует положительным значениям весов, синий – отрицательным. Рисунок с литерой «T» – веса теоретической идеальной модели

Как можно видеть по рисунку 2 в модели явно формируются области положительных и отрицательных значений весов, по которым осуществляется явное разделение между двумя классами. Также возможно заметить преобладание в центральной области отрицательных (модель M1) и положительных (модель M2) значений весов. На текущую вырожденную задачу данная особенность не влияет, так как все модели из набора полностью разделяют два класса. На основе экспериментальных моделей была построена теоретическая идеальная модель (модель T), в которой вручную были занулены веса вне области принятия решения и выставлены в максимальное/минимальное значения веса, необходимые для обеспечения разделимости классов.

4. Внедоменная работа ранее обученных моделей

Для проверки устойчивости обученных моделей был проведен эксперимент с привнесением внедоменных изменений в объекты на изображениях обучающего набора. Набор тестовых данных состоял из 336-ти изображений (168 изображений на каждый класс) размером 256×256. Треугольники каждой ориентации сдвигались попиксельно по оси x, так чтобы центр треугольника x_c находился в диапазоне $43 \leqslant x_c \leqslant 212$, а само изображение треугольника не выходило за границы кадра. Значения выхода сигмоиды после обработки изображений с изменениями приведены на рисунке 3.

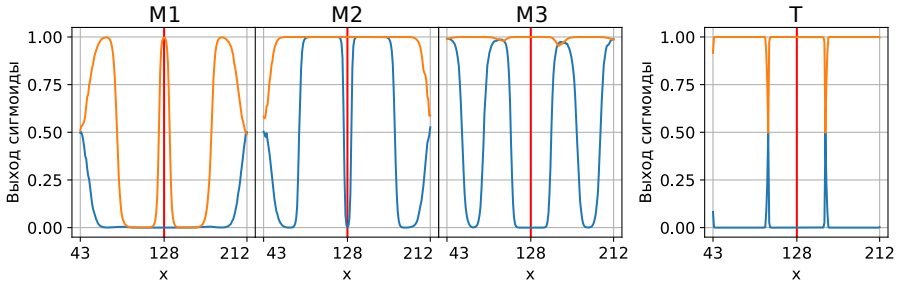

Рис. 3. Отклик модели при сдвиге по оси x доменных объектов. Синяя линия соответствует классу 1, оранжевая – классу 2, красной линией обозначено положение треугольников в обучающем наборе данных

Исходя из результатов проведенного эксперимента возможно отметить, что модель M1 и модель M2 являются противоположными друг другу. Так, у модели M1 шире область принятия решения в сторону класса 1, а у M2 – в сторону класса 2. Также стоит отметить, о наличии пика около первоначального положения треугольников в обучающем наборе данных. Из характеристик этого пика следует вывод, что модели M1 и M2 дадут ложное срабатывание, даже при малом отклонении исходного треугольника от обучающего домена. Обратная ситуация наблюдается у модели M3. Как можно видеть пик в сторону класса 1 у нее значительно шире чем у M2. Это говорит о том, что модель сможет корректно отработать при большем отклонении целевого объекта. Поведение модели M3 больше всего похоже на поведение идеальной теоретической модели T, причем можно заметить, что M3 не совсем полностью сфокусировалась на классе 2, так как присутствуют небольшие выбросы причем именно в тех местах, где у модели T выход сигмоиды становится равен 0.5.

Для проверки гипотезы о влиянии первоначального распределения значений весов нейросетевой модели был проведен эксперимент по введению внедоменного объекта. В качестве внедоменного объекта был выбран круг с центром совпадающим с центрами треугольников из обучающего набора данных и изменяющимся радиусом r в диапазоне $1 \leqslant r \leqslant 127$, то есть было сформировано 127 тестовых изображений. Графики зависимости выхода сигмоиды от радиуса круга приведены на рисунке 4.

Как можно видеть для моделей M1 и M2 при возрастании охвата центральной области кругом значение сигмоиды монотонно изменяется к тому значению, к которому сеть была предрасположена (см. результаты предыдущего эксперимента). Обратная ситуация наблюдается у модели M3. Здесь значение сигмоиды не всюду изменяется монотонно и даже флуктуирует около величины 0.81 до тех пор, пока размер круга не достигает радиуса вписанной в треугольник окружности, то есть

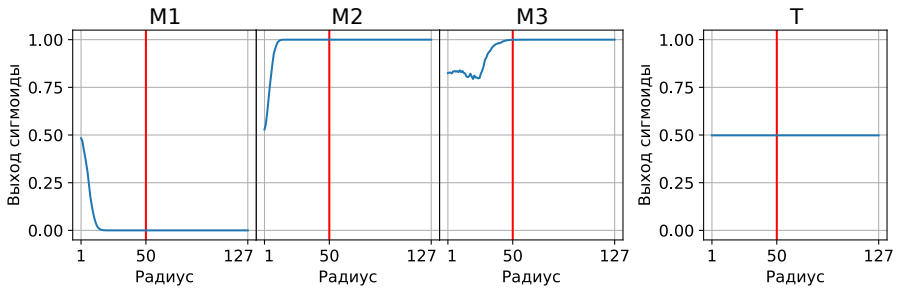

Рис. 4. Отклик модели при изменении радиуса внедоменного объекта (круга)

ключевых областей задействованных в классификации. В тоже время выход модели T при расширении внедоменной фигуры остается неизменным со значением 0.5, то есть модель T не реагирует на внедоменный объект, так как отсутствуют шумы в весах сдвигающие итоговый выход и расширяющийся круг одновременно достигает все ключевые рецепторные области.

В заключительной части данной работы был проведен дополнительный эксперимент с внедоменной фигурой: на входы сетей подавались 10 тыс. изображений размером 256×256, каждое из которых содержит круг со случайным радиусом в диапазоне $1 \leqslant r \leqslant 127$ и случайным расположением. Итоговые гистограммы плотностей распределений значений выхода сигмоиды для каждой из моделей представлены на рисунке 5.

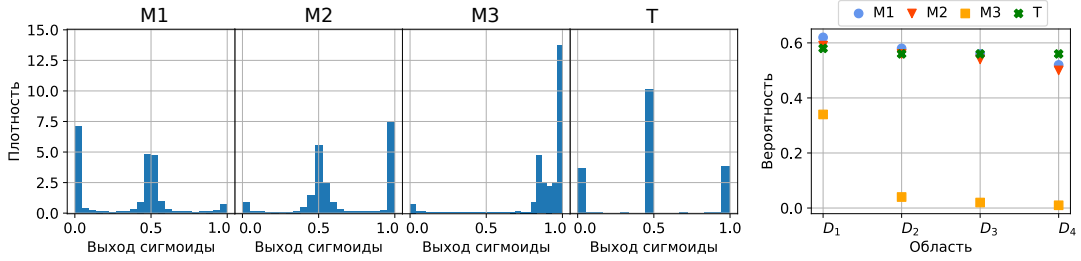

Рис. 5. Гистограммы распределений значений выхода сигмоиды. Справа – диаграмма вероятностей попадания отклика сети в область неопределенности решения модели: $D_1 = [0.1, 0.9]$, $D_2 = [0.2, 0.8]$, $D_3 = [0.3, 0.7]$, $D_4 = [0.4, 0.6]$

По представленным гистограммам у моделей явно наблюдается смещение в классификации. Так модели M1 и M2 наиболее близки по распределению к теоретической модели T, однако имеют явный перекос к одному из классов. В то же время модель M3 по результатам данного эксперимента сильнее всего отличается от модели T и не обладает пиком около 0.5, это означает, что именно эта модель больше всего дает ложных срабатываний в эксперименте. Так же по диаграмме вероятностей попадания в область неопределенности решения модели (см. рис. 5 справа) можно отметить, что модель T является наиболее устойчивой по сравнению с остальными моделями при сужении области неопределенности решения.

5. Заключение

В данной работе построена и изучена микромодель механизма принятия решения нейросетью, в том числе при ее функционировании вне домена обучающих

данных. Проанализировано влияние параметров начальной инициализации параметров сети на ее способность к обучению. Установлено, что начальное распределение весов влияет на характер ложных срабатываний, причем основные ошибки в принятии решения вне домена обучения привносят значения тех весов, которые не являются ключевыми, для формирования ответа на обучающем наборе данных. Следовательно, повысить устойчивость функционирования сети возможно избавлением от этих весов (см. результат по модели T). Развитие предложенного подхода планируется в последующих работах. В первую очередь необходимо найти способ автоматического обнаружения «шумящих» весов.

Список литературы

1. Liang Y., et al. Explaining the black-box model: A survey of local interpretation methods for deep neural networks // Neurocomputing. 2021. Vol. 419. P. 168-182.
2. Li Z., et al. A survey of convolutional neural networks: analysis, applications, and prospects // IEEE transactions on neural networks and learning systems. 2021.
3. Vaswani A. et al. Attention is all you need // Advances in neural information processing systems. 2017. Vol. 30.
4. Zhou K., et al. Domain generalization: A survey // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022.