Филиал Московского Государственного Университета имени М.В. Ломоносова в г. Ташкенте

Факультет прикладной математики и информатики

Кафедра прикладной математики и информатики

Мурадасилов Руслан Серверович

КУРСОВАЯ РАБОТА

на тему: «Проблема выразимости бинарных нейронных сетей»

по направлению 01.03.02 «Прикладная математика и информатика»

Курсовая работа рассмотрена и рекомендована к защите

зав.кафедрой «ПМиИ», д.ф.-м.н., профессор _____ Кудрявцев В. Б.

Научный руководитель:

к.ф.-м.н. _____ Иванов И. Е.

«____»_____ 2020 г.

Ташкент 2020

Аннотация

В данной работе рассматривается проблема выразимости бинарной нейронной сети. Определено, как устроен слой такой нейронной сети. Исследуется полнота слоя при ограничениях. Обозначена проблема аппроксимации булевых функций линейными булевыми функциями.

Аннотация

In this work

Содержание

1 Введение

В данной работе рассматривается проблема выразимости бинарной нейронной сети. Вводятся понятия, определяющие структуру слоя. Рассматривается проблема полноты нейронной сети, слой которой представляет собой комбинацию «линейная функция + некоторая нелинейность». Доказано, что такая нейронная сеть будет полна. Также доказано существование полной нейронной сети при ограничении в два слоя.

Ставится задача исследовать аппроксимацию булевых функций нейронной сетью, слой которой состоит только из линейных функций.

2 Историческая часть вопроса

За последнее десятилетие сложность и способности нейронных сетей существенно выросли, однако их потенциал до сих пор ограничивают стоимость и энергия потребления. Как известно, нейронные сети состоят из нескольких слоев взвешенных сумм, которые предсказывают нужный результат. Хранение всех значений чисел с плавающей запятой значительно увеличивает время обучения и требует большое количество памяти, что вызывает необходимость использовать только специальное оборудование, которое выдержит подобную нагрузку.

Чтобы устройство с ограниченными ресурсами могло решать такие проблемы глубокого обучения, как распознавание лиц в реальном времени, необходимо использовать в качестве весов бинарные числа, а в качестве функций активации — бинарные аналоги функций активации в непрерывном случае, то есть «бинаризовать» нейронную сеть. Это позволит хранить гораздо больший объем данных, используя, например, 32-битный контроллер. Использование битовых операций сокращает время исполнения. Размеры бинарных нейронных сетей намного меньше, чем у их вещественнных аналогов. Точность моделей также меньше, но эта разница в точности постепенно сокращается и бинарные нейронный сети становятся точнее на больших датасетах, как ImageNet.

Идея бинарной нейронной сети была впервые предложена Matthieu Courbariaux, где веса и функции активации используют только бинарные числа как и в inference, так и в алгоритме обратного распространения ошибки с использованием метода градиентого спуска (SGD) [1].

Чуть позже Mohammad Rastegari в модели XNOR-Net [2] добавляет gain term, чтобы компенсировать потерю информации во время бинаризации, который был получен из статистики весов и функций активации до бинаризации. Это улучшило общие показатели, но подсчет gain term оказался дорогостоящим.

Victor Zhou попытался обобщить квантизацию и использовал преимущество битовых операций для фиксированной точки данных, варьируя ее размерность [3]. Zhou представил DoReFa-Net, модель с перемен-

ной размерностью весов, функций активации и даже вычисления градиента во время обратного распространения ошибки со значительно улучшенным временем обучения.

Другие модели также добились положительных результатов. Zheng Tang ускорил время обучения, изучив как влияет скорость обучения на показатели нейронной сети и на колеблемость бинарных значений [4]. Идея BNN+, созданная Sajad Darabi,также улучшает скорость обучения, использую другую эффективную функцию обратного распространения вместо импульсной [5].

Сравнение результатов всех этих моделей на разных датасетах были подробно описаны в работе Taylor Simons и Dah-Jye Lee [6]. Наглядно видно, что на MNIST, SVHN, CIFAR и ImageNet датасетах бинарные нейронные сети практически не уступают своим вещественнным аналогам.

В данной работе рассматривается вопрос полноты бинарной нейронной сети, то есть способность в точности выразить булеву функцию. А также ставится задача аппроксимации булевых функций, то есть способность выразить булеву функцию с некоторой допустимой погрешностью. С практической точки зрения устраивают оба варианта.

3 Основные результаты

3.1 Полнота

Введем основные определения.

Определение 1. *Нейронной сетью* называются композиции функций, каждую из которых назовем слоем. Под композицией функций понимается подстановка одной функции в качестве аргумента другой.

Определение 2. *Слоем* называется композиция весовой функции и нелинейного бинарного оператора, выступающего в роли функции активации. *Входным слоем* будем называть вектор $X = (x_1, ...x_n) \in \{0,1\}^n$, а *выходным* — вектор $Y = (y_1, ...y_n) \in \{0,1\}^n$.

Определение 2. *Весовой функцией* называется линейная булева функция, применяемая ко всем выходам предыдущего слоя, включая входной слой.

Определение 3. Вектор $f^i = (f_1^i, ..., f_k^i)$, где i — номер скрытого слоя, k — количество нейронов внутри i-ого слоя, а каждая f_j^i является функцией алгебры логики от n переменных $(j = \overline{1,k})$ называется *бинарным оператором.*

Определение 4. Бинарный оператор называется *линейным*, если каждая его компонента f_j^i является линейной $(j = \overline{1,k})$. Соответсвенно, бинарный оператор называется *нелинейным*, если хотя бы одна из его компонент f_j^i является нелинейной $(j = \overline{1,k})$.

Справедлив следующий результат, относящийся к проблеме полноты бинарной нейронной сети с такой архитектурой.

Теорема 1. Бинарная нейронная сеть с линейной булевой весовой функцией и нелинейным бинарным оператором в роли функции активации полна в P_2.

Доказательство. Так как класс булевых линейных функций является предполным, то достаточно хотя бы одной нелинейной функции, чтобы нейронная сеть могла порождать P_2. По определению нелинейного бинарного оператора такая функция будет существовать, что доказывает теорему. ■

Рассмотрим вопрос существования полной бинарной нейросети с двумя скрытыми слоями. Справедлива следующая теорема.

Теорема 2. Существует бинарная нейронная сеть с линейной булевой весовой функцией и нелинейным бинарным оператором в роли функции активации с двумя скрытыми слоями, которая полна в P_2.

Доказательство. Построим бинарную нейронную сеть, реализующую СДНФ функции. Нейросеть будет полносвязной, так как в СДНФ все конъюнкции полные. Во входном слое вектор $X = (x_1, ...x_n) \in \{0,1\}^n$, на котором хотим получить значение функции. В первом скрытом слое k нейронов ($k \leqslant 2^n$), реализующих эти конъюнкции, то есть $f^1 = (f_1^1, ..., f_k^1)$ $\forall f_j^1 = \wedge$, $j = \overline{1,k}$. Весовая функция первого скрытого слоя — исключающее «или», которая позволит получать отрицания переменных при передаче их в конъюнкцию. Во втором скрытом слое один нейрон, принимающий на вход все конъюнкции СДНФ и реализующий их дизъюнкцию, то есть $f^2 = \vee$. Таким образом, функции активации (конъюнкция, дизъюнкция) — нелинейны, а весовая (исключающее «или») — линейна, и мы сохраняем структуру слоя (рис. 1).

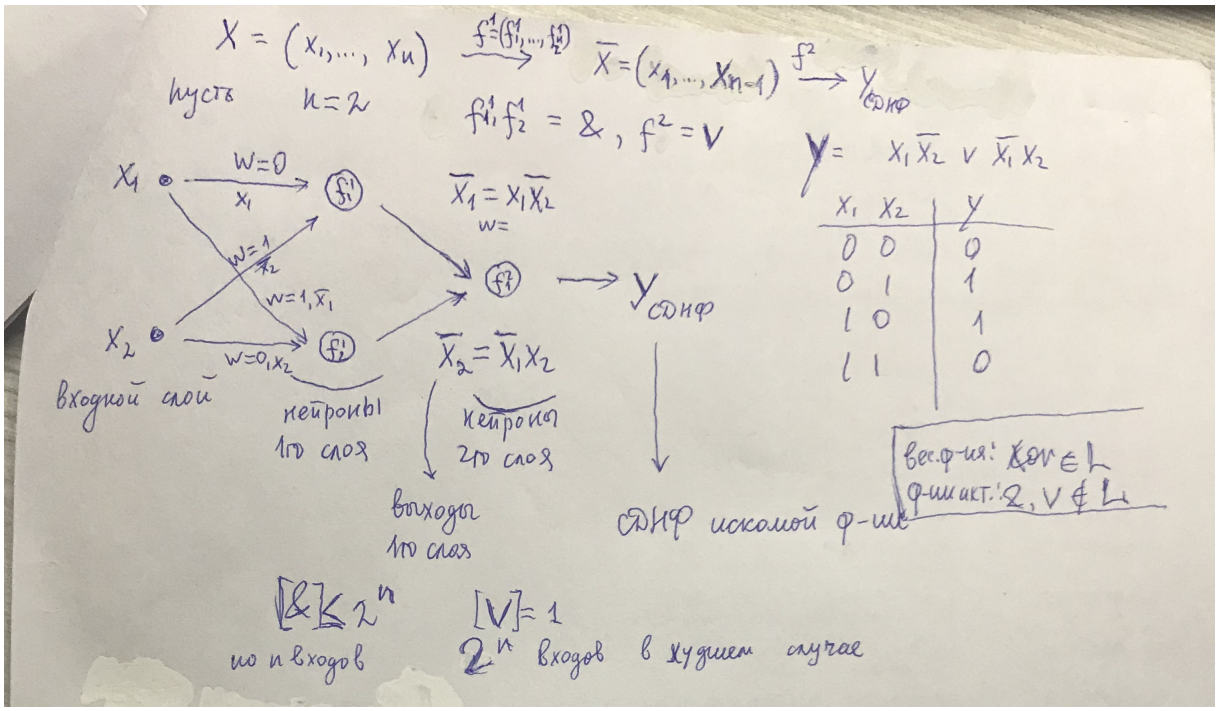

Рис. 1

Так как СДНФ функции конечна и единственна, то по ней мы сможем однозначно восстановить все функции алгебры логики, и такая нейронная сеть будет полна, что и требовалось доказать. ■

Исторически, проблема выразимости нейронных сетей с вещественными значениями прослеживается от 13-й проблемы Гильберта, ставящей вопрос "существует ли непрерывная функция трех переменных, которая не может быть представлена через композицию непрерывных функций двух переменных". Она была решена в 1957 г. В.А. Арнольдом; он показал, что любая непрерывная функция трех переменных представляется в виде композиции непрерывных функций двух переменных. В том же 1957 г. А.Н. Колмогоров доказал более сильную теорему: для реализации функций многих переменных достаточно операций суммирования и композиции функций одной переменной. Эта теорема в 1987 году была переложена Хехт–Нильсеном для нейронных сетей: любая функция нескольких переменных может быть представлена двухслойной нейронной сетью с прямыми полными связями с N нейронами входного слоя, $(2N + 1)$ нейронами скрытого слоя с ограниченными функциями активации (например, сигмоидальными) и M нейронами выходного слоя с неизвестными функциями активации [7, 8, 9, 10].

В бинарных нейронных сетях справедлив аналогичный результат.

Теорема 3. Бинарная нейронная сеть с линейной булевой весовой функцией и нелинейным бинарным оператором в роли функции активации с тремя слоями полна в P_2.

Доказательство.

3.2 Аппроксимация

В практике расчетов, связанных с обработкой экспериментальных данных, вычислением $f(x)$, разработкой вычислительных методов, встречаются следующие две ситуации:

1. Как установить вид функции $y = f(x)$, если она неизвестна? Предполагается при этом, что задана таблица ее значений, которая по-

лучена либо из экспериментальных измерений, либо из сложных расчетов.

2. Как упростить вычисление известной функции $f(x)$ или же её характеристик, если $f(x)$ слишком сложная?

Ответы на эти вопросы даются теорией аппроксимации функций, основная задача которой состоит в нахождении функции $\varphi(x)$, близкой (т.е. аппроксимирующей) в некотором нормированном пространстве к исходной функции $y = f(x)$. Функцию $\varphi(x)$ при этом выбирают такой, чтобы она была максимально удобной для последующих расчетов.

Одним из замечательных свойств нейронных сетей является способность аппроксимировать и, более того, быть универсальными аппроксиматорами. С помощью нейронных сетей можно аппроксимировать сколь угодно точно непрерывные функции многих переменных.

Что касается бинарных нейронных сетей, здесь этот вопрос ещё хорошо не изучен. Важно доказать, насколько хорошо можно аппроксимировать любую булеву функцию линейными, то есть важен вопрос линейной аппроксимации в бинарном случае.

Введем метрику, вычисляющую процент несовпадений по следующей формуле:

$$\sum |y_{real} - y_{pred}| \cdot 100\%,$$

где y_{real} — столбец значений предсказываемой функции, y_{pred} — столбец значений функции, полученной в результате аппроксимации.

4 Заключение

В данной работе исследована проблема полноты бинарной нейронной сети. Доказана теорема о полноте нейронной сети со слоем «линейная функция + некоторая нелинейность». Также доказан результат о существовании полной нейронной сети при ограничении в два слоя с помощью СДНФ.

Обозначена важность проблемы аппроксимации булевых функций линейными функциями. Ставится цель изучить линейную аппроксимацию в бинарном случае.

5 Список использованной литературы

[1] Courbariaux, M.; Bengio, Y. BinaryNet: Training Deep Neural Networks with Weights and Activations Constrained to +1 or −1. arXiv:1602.02830.

[2] Rastegari, M.; Ordonez, V.; Redmon, J.; Farhadi, A. XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 525–542.32.

[3] Zhou, S.; Ni, Z.; Zhou, X.; Wen, H.; Wu, Y.; Zou, Y. DoReFa-Net: Training Low Bitwidth Convolutional Neural Networks with Low Bitwidth Gradients. arXiv 2016, arXiv:1606.06160.

[4] Tang, W.; Hua, G.; Wang, L. How to Train a Compact Binary Neural Network with High Accuracy? In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017

[5] Darabi, S.; Belbahri, M.; Partovi Nia, V.; Courbariaux, M. Regularized Binary Network Training. arXiv:1812.11800

[6] Simons, T.; Lee, D. A Review of Binarized Neural Networks. Electrical and Computer Engineering, Brigham Young University, Provo, UT 84602, USA.

[7] Колмогоров А. Н. О представлении непрерывных функций нескольких переменных суперпозициями непрерывных функций меньшего числа переменных. ДАН СССР, 1956, Т.108, №2, С. 179-182

[8] Арнольд В.И. О функции трех переменных. ДАН СССР, 1957, Т.114, №4, С. 679-681

[9] Колмогоров А. Н. О представлении непрерывных функций нескольких переменных в виде суперпозиций непрерывных функций одного переменного и сложения. ДАН СССР, 1957, Т.114, №5, С. 953-956

[10] Hecht-Nielsen R. Kolmogorov's mapping neural network existence theorem. IEEE First Annual Int. Conf. on Neural Networks, San Diego, 1987. Vol. 3. — P. 11—13.