ИССЛЕДОВАНИЕ ПАРАЛЛЕЛЬНЫХ СТРУКТУР НЕЙРОННЫХ СЕТЕЙ ДЛЯ ИСПОЛЬЗОВАНИЯ В ЗАДАЧАХ ПО СЕМАНТИЧЕСКОЙ КЛАССИФИКАЦИИ ТЕКСТА НА РУССКОМ ЯЗЫКЕ В УСЛОВИЯХ ОГРАНИЧЕНИЯ ВЫЧИСЛИТЕЛЬНЫХ РЕСУРСОВ (НА ПРИМЕРЕ ОПЕРАТИВНЫХ СВОДОК В СИСТЕМЕ МВД РОССИИ)

В.И. Воронов, *доцент кафедры «ИСУиА», МТУСИ, к.т.н., 111024, г. Москва, ул. Авиамоторная, 8А, vorvi@mail.ru;*
Э.В. Мартыненко, *магистрант МТУСИ, 111024, г. Москва, ул. Авиамоторная, 8А, mail@martinenko.com*

УДК 681.5.01

Аннотация. Проведено исследование параллельных структур нейросетей при решении задач семантической классификации русскоязычных текстов на примере оперативных сводок МВД РФ при ограниченности вычислительных мощностей для обучения и повседневного использования. Построены опытные модели структур нейросетей, собрана статистика и проанализирована их работа на тестовых данных. Проведен анализ производительности.

Ключевые слова: параллельные структуры нейронных сетей; сверочные нейронные сети; *KERAS*; *LSTM*; нейронные сети; *CNN*; *RNN*; *Tensor Flow*; семантическое распознавание текста на русском языке; классификация текста по категориям.

RESEARCH OF PARALLEL STRUCTURES OF NEURAL NETWORKS FOR USE IN THE TASKS ON THE RUSSIAN TEXT SEMANTIC CLASSIFICATION CONSIDERING LIMITED COMPUTING RESOURCES (ON THE EXAMPLE OF OPERATIONAL REPORTS USED IN THE RF MIA)

Vyacheslav Voronov, *associate professor of «ISMA» department, Ph.D., MTUCI, 111024, Moscow, ul. Aviamotornaya, 8A*
Eduard Martinenko, *master's student MTUCI, 111024, Moscow, ul. Aviamotornaya, 8A*

Annotation. A study of parallel structures of neural networks in solving problems of semantic classification of Russian-language texts on the example of operational reports of the Ministry of Internal Affairs of the Russian Federation with limited computing power for learning and everyday use. Experimental models of neural networks structures were built, statistics were collected and their work on test data was analyzed. Performance analysis conducted.

Keywords: the parallel structure of neural networks; convolutional neural network; *KERAS; LSTM*; neural networks; *CNN; RNN*; Tensor Flow; semantic recognition of the text in the Russian language; text classification by category.

Данное исследование направлено на экспериментальное построение параллельных структур нейронных сетей с целью определения возможности решения задач по семантической классификации текста, а также выбора наиболее оптимальной структуры для дальнейшего создания работоспособной нейронной сети при последующем внедрении и использовании в повседневной деятельности подразделений МВД России [1] (планируется использование предсказанных категорий текстовой информации в программном обеспечении для автоматической подготовки докладов и отчетов) без дополнительного вложения материальных средств или необходимости обновления уже существующей инфраструктуры автоматизированных рабочих мест (в связи с чем будет проведена оценка производительности и потребления системных ресурсов).

Учитывая служебный характер информации, вся работа с системой производится только на служебных компьютерах. Использование возможностей облачных вычислений не

представляется возможным. Ни один из компьютеров не оснащен видеокартой компании *NVIDIA*, в связи с чем все вычисления будут проводиться только на *CPU*.

Задачами проведения данного исследования являются:

- Экспериментальное построение моделей параллельных нейронных сетей.
- Оценка точности классификации на наборе обучения, наборе валидации, тестовом наборе данных.
- Оценка производительности при дальнейшей интеграции системы для удобства работы пользователей.
- Оценка потребления системных ресурсов (оперативной памяти).
- Сбор статистических данных в ходе проведения эксперимента в результате работы (обучения) различных структур нейронных сетей, предназначенных для классификации нейронных сетей.
- Проведение анализа собранных статистических данных.
- Подведение итогов и определение положительных и отрицательных сторон выбранной структуры нейронных сетей.

Особенности входных данных

Текст представлен на русском языке, согласно параметрам, указанным в табл. 1. Дополнительной особенностью является использование специальной юридической и профессиональной терминологии, т.к. классифицируемая информация является информацией с оперативных сводок системы МВД России, а результатом классификации является содержание сводки.

Таблица 1

Параметры входных данных для обучения, используемых в данном исследовании

Параметры	**Значения**
Длина одного сообщения (слов):	Среднее – 500 Максимальное – 956
Язык:	Русский, используются специфические профессиональные и юридические термины, аббревиатуры.
Кодировка текста:	UTF-8
Количество классов: (особенностью является неравномерное распределение информации по классам)	4
Количество записей в массиве данных для проведения исследований:	9500
Дополнительную сложность набора данных представляет тот факт, что некоторые записи могут подпадать под несколько классов (они являются неоднозначными и спорными), их определение может быть дано только с использованием дополнительных данных, не представленных в наборе, что в дальнейшем может снижать точность.	

Особенности проведенного исследования

Для сохранения последовательности результатов тестирования не используются возможности библиотеки *NLTK* (или аналогов) и не проводится *Stemming and lemmatization*. Однако удаляются некоторые стоп-слова, все знаки препинания и, что более важно, все цифровые значения – такие как даты, время или государственные регистрационные знаки –

преобразуются в шаблонную форму. В проведенных мною ранее исследованиях [1] было выявлено, что простое удаление цифровых значений приводит к заметному снижению точности классификации, т.к. даты, время и иные похожие данные представляют шаблоны, несущие семантический смысл. Важен сам факт их включения текст, в том числе их расположение в предложении, но не сами конкретные значения, которые могут только увеличивать размер словаря без переноса значимой для нейронной сети информации, и каждая отдельная дата (или иная информация, которая может быть представлена в виде общего для всех значений шаблона) будет представлена новым значением в словаре, при этом не позволяя выделить полезных для проведения классификации особенностей).

Во всех структурах используется технология *Word embedding*, а также собственный класс для подготовки текстовой информации, в том числе для создания словаря, набора обучаемых классов, определения максимальной длины предложения, padding до необходимой максимальной длины текстовой информации, преобразования в массивы nympy.

При построении модели используются средства Фреймворка *Keras* версии 2.2.4 (с применением возможностей функционального подхода, так как структура является параллельной, а не последовательной) совместно с *Tensor Flow* версии 1.11.0 и последней доступной на момент написания статьи совместимой версией интерпретатора языка программирования *Python* 3.5.4 (*Win* 64-битная версия).

Конфигурация компьютера, на котором осуществляется тестирование:

Процессор: *Intel Core i*5-2300.

Оперативная память: 8 *GB* (*DDR*3-1600 / *PC*3-12800 *DDR*3 *SDRAM*).

Видеокарта: *AMD Radeon HD* 6770 (возможности ускорения вычислений с использованием графического ускорителя не используются).

Результаты, полученные в ходе моделирования структуры нейронной сети с использованием параллельных блоков свертки, *max pooling u flatten* с общим блоком *word embedding* с различными гиперпараметрами представлены в табл. 2.

Таблица 2

Тип структуры	*LSTM* (количество ячеек)	Оптимизатор	Глубина вектора *word embding*	Количество параметров для обучения	Время загрузки модели	Время обучения одной эпохи	Максимальная точность (валидации)	Максимальная точность (теста)
Parallel CNN	Нет	*Adam*	32	886,596	60	201	0,985	97,5
Parallel CNN	200	*Adam*	32	919,908	90	320	0,985	97,75

На рис. 1 показана структура нейронной сети с использованием параллельных блоков свертки, *max pooling* и *flatten* с общим блоком *word embedding*.

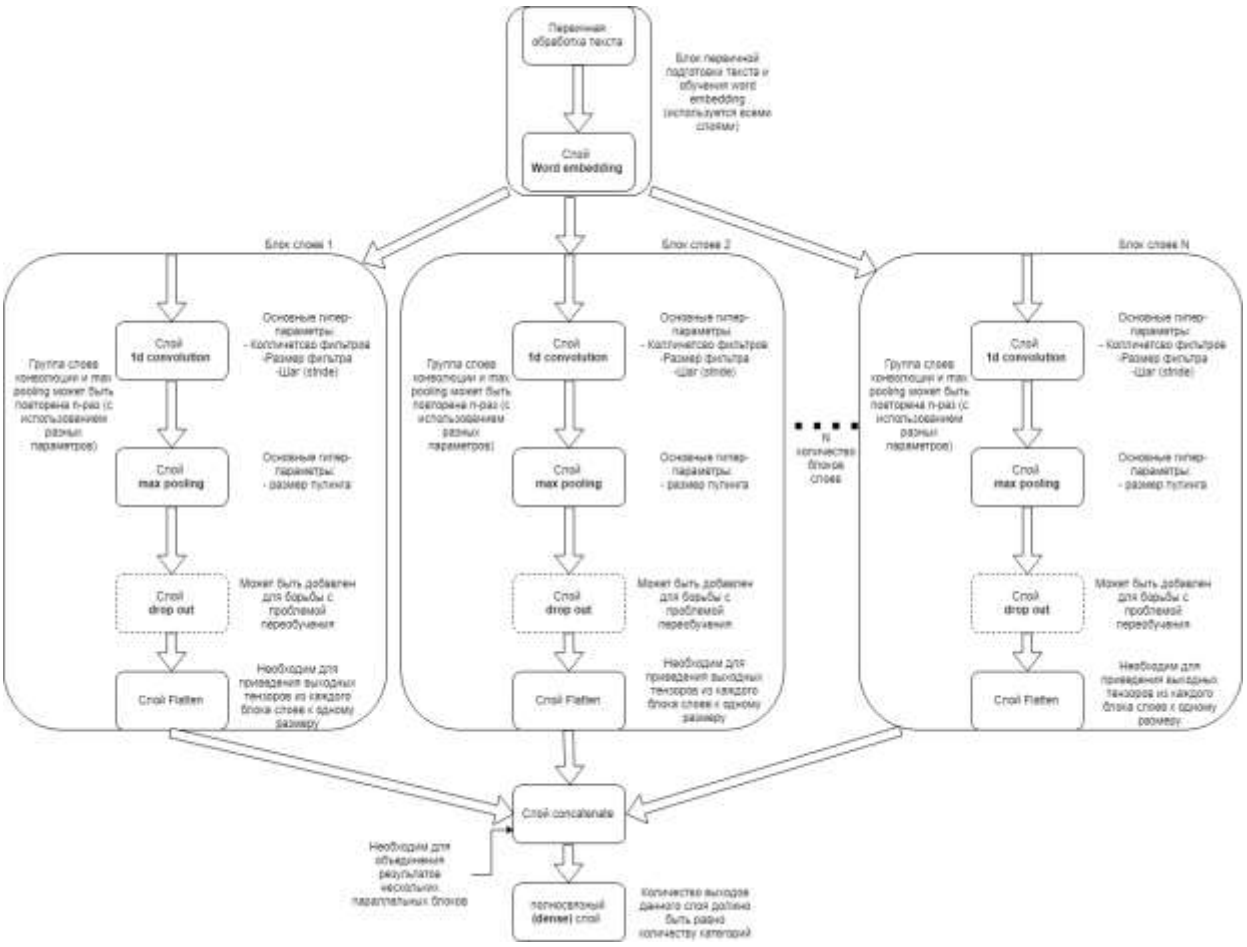

Рисунок 1

Используемая параллельная структура нейронной сети, как это видно на рис. 1, представляет из себя структуру, в которой обучение каждого блока происходит независимо друг от друга, и результаты значений объединяются с использованием слоя concatenate в последних слоях структуры нейронной сети.

При проведении исследования использована идея о комбинации слоев конволюции, *max pooling* и *LSTM* для решения задач по семантической классификации текста [3], так как они показали лучшие результаты, однако, в данной статье не предлагается использовать параллельные структуры нейронной сети.

Особенностью данной структуры является использование отдельных блоков конволюции и *max pooling*. Каждый блок представляет из себя следующие элементы:

- Слой одномерной свертки.
- Слой *max pooling*.
- Слой drop out. Данный слой является опциональным и может применяться для борьбы с проблемой переобучения сети.
- Слой *Flatten* для преобразования результатов в одномерный тензор.

На вход каждого из этих блоков подается результат слоя *word embedding*. В данной структуре для каждого блока используется общий слой *word embedding*, а также общий слой входных данных.

Все выходы соединяются в слое concatenate, задачей которого является объединение результатов работы блоков свертки.

Каждый из этих блоков может быть использован *N* количество раз, что представляет собой новый дополнительный настраиваемый гиперпараметр и является особенностью данной структуры сети. Однако для получения положительного результата в каждом отдельном блоке свертки необходимо настраивать независимые, но близкие и последовательные значения размера фильтров и *max pooling*, т.к. при использовании значений сильно отличных друг от друга будут сильно изменены размеры выходных тензоров их каждого блока.

Основным недостатком данной структуры является повышенное потребление вычислительных ресурсов и особенно оперативной памяти. Также у данной структуры низкая производительность, как это видно из результатов, представленных в табл. 2.

При построении данной структуры нейронной сети необходим функциональный подход с использованием возможностей Фреймворка *Keras*. Этот подход является более сложным в реализации и менее эффективным в отношении производительности по сравнению с созданием последовательных моделей нейронных сетей, например, как это было исследовано ранее [1].

Однако при использовании графических ускорителей компании *Nvidia* данный недостаток может быть несущественен ввиду эффективного проведения параллельных вычислений (данная особенность в ходе исследования не изучалась из-за отсутствия финансирования).

На рис. 2 представлен график точности при использовании структуры нейронной сети с применением параллельных блоков свертки, *max pooling* и *flatten* с общим блоком *word embedding.*

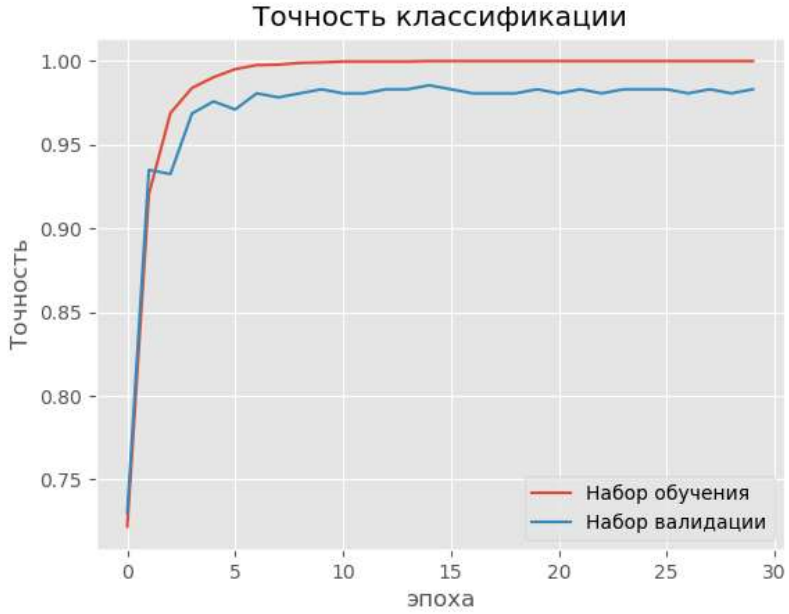

Рисунок 2

На рис. 3 представлен график потребления оперативной памяти (ОЗУ) во время исследования структуры нейронной сети с использованием параллельных блоков свертки, *max pooling* и *flatten* с разными гиперпараметрами.

На рис. 4 представлен график сравнения точности классификации различных наборов данных во время исследования структуры нейронной сети с использованием параллельных блоков свертки, *max pooling* и *flatten* с разными гиперпараметрами.

На рис. 5 представлен график времени, затраченного на обучение одной эпохи во время исследования структуры нейронной сети с использованием параллельных блоков свертки, *max pooling* и *flatten* с разными гиперпараметрами.

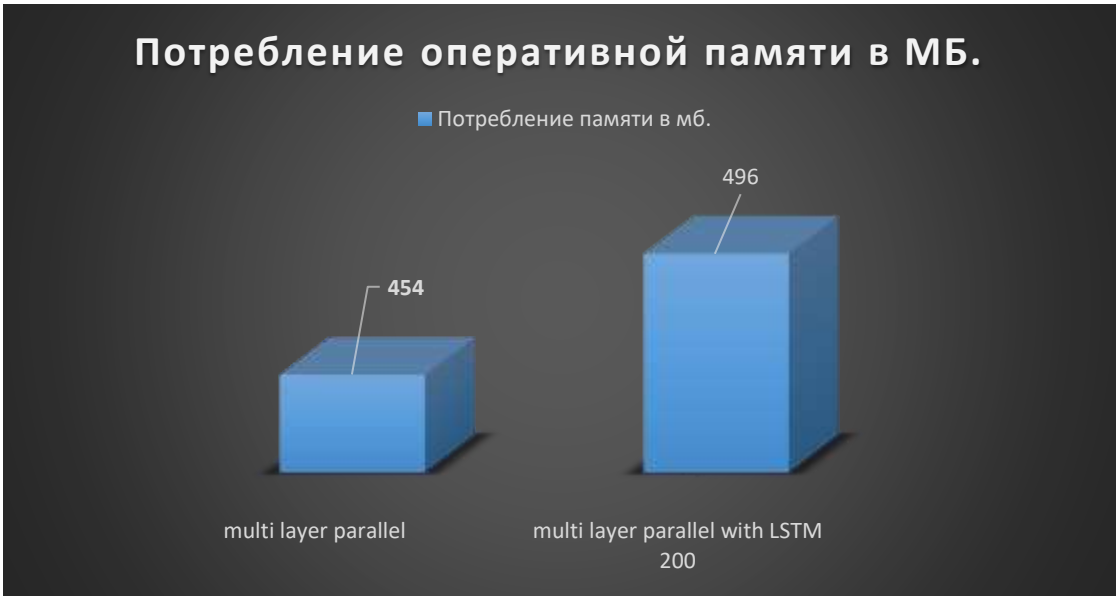

Рисунок 3

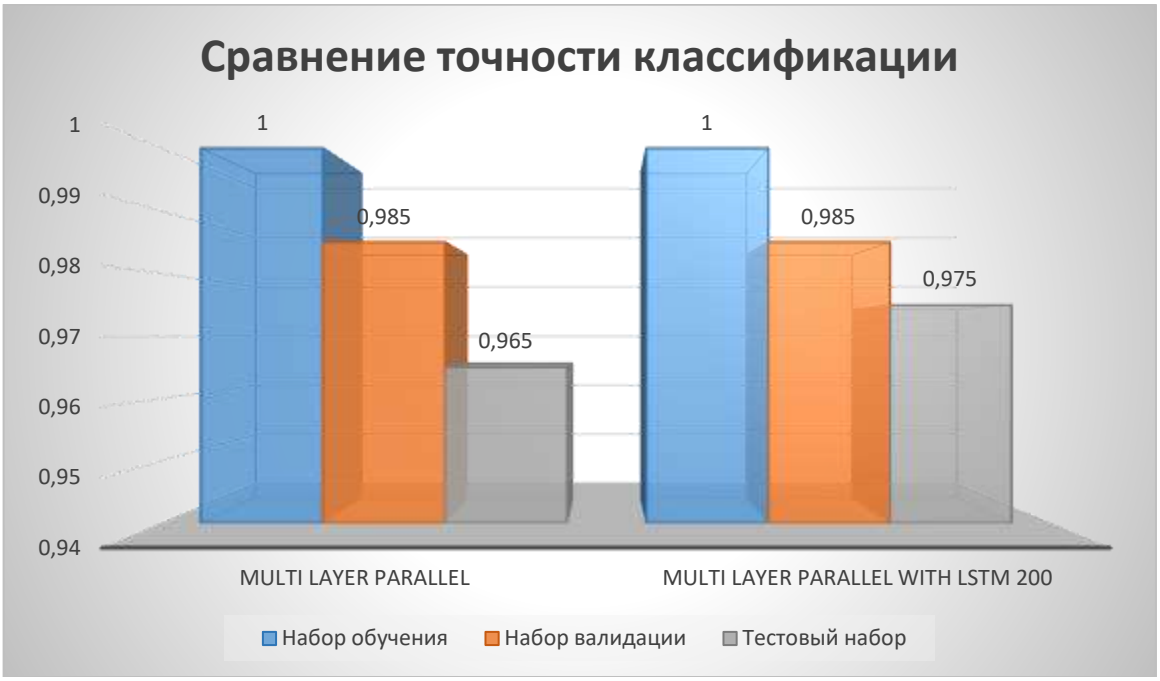

Рисунок 4

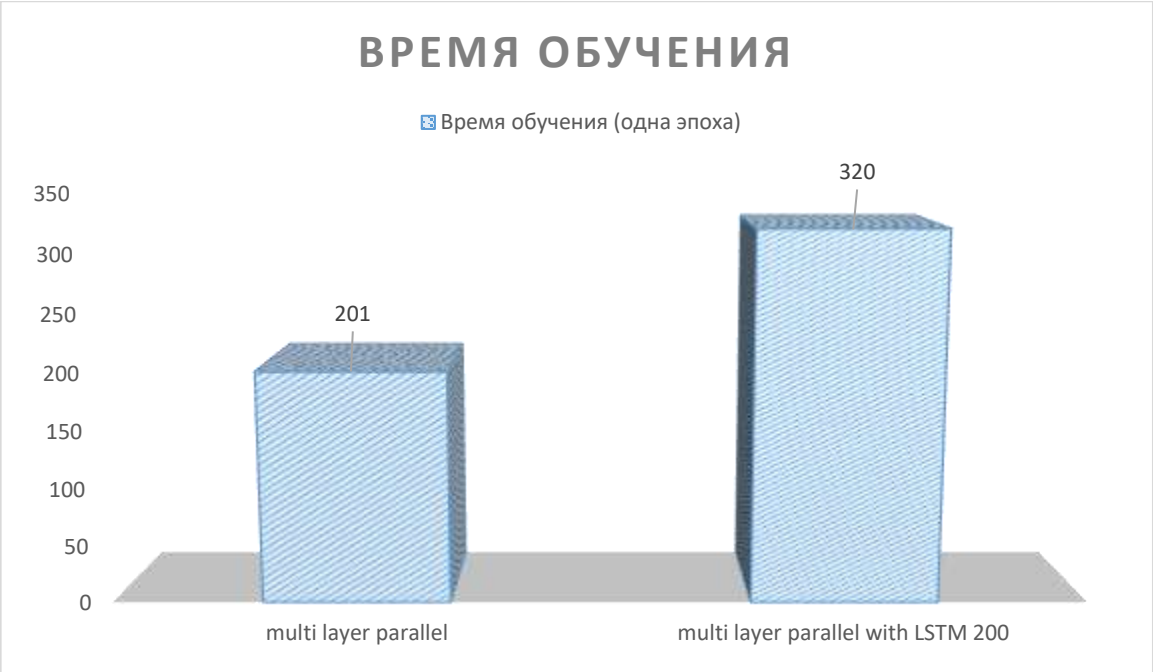

Рисунок 5

Для проведения тестирования была построена модель с тремя параллельными блоками свертки со следующими гиперпараметрами:

- Для слоев одномерной свертки выбраны последовательно увеличивающиеся размеры фильтров:
  4,6,8, функция активации *Relu*.
- Для слоев Drop out значение было установлено в 0,5.
- Размер max *pooling* 2.
- В конце каждого блока применен слой flatten ().
- Результаты всех блоков объединяются в слое concatenate.
- Для полносвязанных слоев использована функция активации *softmax*.
- В качестве оптимизатора для модели использован Adam. Также было проведено тестирование *Adadelta*, однако, при этом получены более низкие значения точности классификации.

Анализ полученных в ходе экспериментов данных показывает очень ровные значения точности на всем периоде обучения, так как это видно на рис. 2. После достижения моделью максимальной точности значительной проблемы переобучения не возникает. При этом точность классификации намного выше, чем у структур, в которых используются только слои *Word Embeding* и *LSTM* [2]. Также данная структура обладает высокими показателями точности, как это видно на рис. 4. Применение слоев *LSTM* практически не влияет на точность классификации данных для валидации и в ходе обучения. Однако использование слоев *LSTM* в данной структуре нейронной сети позволяет увеличить точность на тестовых данных, как это видно на рис. 4 и в табл. 2. При этом значительно увеличивается потребление моделью оперативной памяти, как это видно на рис. 5, и практически в два раза увеличивается время обучения одной эпохи, как это показано в табл. 2.

Также возможно использование структуры, в которой применяются как отдельные блоки *word embedding* и входных данных, так и общие, как это показано на рис. 1. В ходе исследования установлено, что использование отдельных блоков *word embedding* является нецелесообразным, так как на точность классификации положительно это не влияет, при этом значительно увеличивается потребление памяти и время работы. В каждом отдельном блоке

Word Embedding должны быть одинаковые настройки глубины вектора, так как для дальнейшей конкатенации полученных результатов и подачи на вход следующего слоя размер выходного тензора должен быть одинаковым.

Также следует отметить ограничение максимального разброса значений размера и количества фильтров и размера слоя max pooling в каждом последовательном блоке, так как при изменении данных параметров меняется размер выходного тензора, каждый из которых будет подвергаться операции конкатенации. При больших различиях это может привести к артефактам классификации из-за различных пространственных размеров выходных тензоров. Оптимальным для данной структуры является использование последовательных инкрементно изменяемых значений размера фильтра, как, например, в первом блоке размер фильтра 2, во втором – 3 и т.д.

Данная структура обладает большим потенциалом при построении модели нейронной сети для решения задач по классификации текста в зависимости от конкретных текстовых данных, так как она, по сути, добавляет дополнительный гиперпараметр: количество параллельных блоков свертки, в каждом из которых могут быть использованы свои уникальные гиперпараметры только для этого конкретного блока. Однако это все дается ценой производительности и потребления системных ресурсов.

Плюсы:

- Высокие потенциальные показатели точности классификации.
- Возможность обучения отдельных блоков свертки max pooling с разными гиперпараметрами независимо друг от друга.
- Гладкий график точности во время обучения и меньшая проблема переобучения нейронной сети.
- Возможность более тонкой настройки, так как количество блоков и их параметры являются новым гиперпараметром специфичным для данной структуры, который может быть установлен для решения конкретных задач по классификации.

Минусы:

- Низкие показатели производительности.
- Высокие показатели потребления оперативной памяти.
- Есть ограничение на используемый размер фильтра, так как в случае сильного различия при конкатенации результатов работы разных блоков может происходить искажение данных из-за различных размеров полученных тензоров.

Исследованная параллельная структура нейронной сети может быть применена для решения задач семантической классификации текста по категориям, причем она показала очень хорошие результаты. Данная структура обладает преимуществами в точности проведения классификации над последовательной структурой с использованием слоев *Word Embedding* и *LSTM* [2], а также обладает новыми настраиваемыми гиперпараметрами, присущими только параллельным структурам: количеством блоков свертки и дополнительной гибкости отдельной настройки каждого из блоков, параметры которых могут быть подобраны под конкретный набор данных. Данная структура может быть использована в повседневной деятельности для решения задач по классификации текстовой информации при использовании слоев *Word Embedding* и в целом отвечает требованиям производительности без использования средств ускорения (при использовании небольшого количества блоков свертки и небольшой глубине вектора *Word Embedding*) и может быть построена с использованием функциональных возможностей Фреймворка *Keras*.

Основным недостатком является низкая производительность и высокое потребление системных ресурсов особенно при использовании большого числа блоков свертки, большой глубины вектора *Word Embedding* и использовании большого количества ячеек *LSTM*. Однако данная структура является перспективной и гибкой. В дальнейшем целесообразно проведение дополнительных исследований, особенно, с использованием средств ускорения параллельных

вычислений, таких как использование графических ускорителей фирмы *Nvidia*, кластерных и облачных вычислений.

Литература

1. Воронов В.И., Мартыненко Э.В. Применение рекуррентной нейронной сети с длинной краткосрочной памятью для классификации информации из оперативных сводок системы МВД России // Телекоммуникации и информационные технологии, 2018. – Т. 5. – № 1. – С. 131-135.
2. Radhika K., Bindu K.R. A text classification model using convolution neural networkand recurrent neural network // International journal of pure and applied mathematics, 2018. – No. 15. 1549-1554.
3. Xingyou W. Combination of convolutional and recurrent neural network for sentiment analysis of short text // Beijing Language and Culture University, Beijing, China, 2017. – С. 2428-2437.