

ОДЕСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ імені І.І.МЕЧНИКОВА

Факультет математики, фізики та інформаційних технологій

Кафедра математичного забезпечення комп'ютерних систем

Дипломна работа

на здобуття ступеня вищої освіти «бакалавр»

на тему Розробка системи машинного читання по губах з використанням

алгоритмів глибокого навчання

The development of machine lip-reading system using deep learning algorithms

Виконав: студент денної форми навчання

напрямку підготовки	123 – Комп'ютерна інженерія	.
---------------------	-----------------------------	---

(шифр і назва напрямку підготовки, спеціальності)

Яворський Євгеній Юрійович

(прізвище, ім'я, по-батькові)

Керівник _____ к.ф.-мат.н., доц. Шпінарева І.М.

(науковий ступінь, вчене звання, прізвище та ініціали, підпис)

Рецензент _____ к.техн.н., доц. Волощук Л.А.

(науковий ступінь, вчене звання, прізвище та ініціали)

Рекомендовано до захисту:

Протокол засідання кафедри

№ _____ Від « _____ » _____ 2020 р.

Завідувач кафедри

Захищено на засіданні ЕК №

протокол № від « » 2020 р.

Оцінка _____ / _____ / _____

(за національною шкалою, шкалою ECTS, бали)

Голова ЕК

(підпис)

Є.В. Малахов

(прізвище, ініціали)

(підпис)

Н.Ф. Казакова

(прізвище, ініціали)

Одеса - 2020

АНОТАЦІЯ

У сучасному світі технології неперервно рухаються у напрямку автоматизації і облегшення людського життя. Усі розробки у сфері машинного навчання і нейронних мереж, спрямованні на розв'язання задач, на які легко здатна людина, але є складними з математичної точки зору. Одною з таких задач є трансліювання людської мови у вигляді тексту, але інколи виникають ситуації у яких ця задача стає дуже складною, при гучній обстановці, коли людина говорить надто швидко або не достатньо чітко вимовляє слова, або за якихось обставин, звук загалом відсутній. У таких випадках підвищити якість розпізнавання мови може читання по губах. Ця задача не є тривіальною як для людини так і для комп'ютера. Тому саме по собі читання по губах не дає високих результатів в розпізнаванні мови, але разом з іншими технологіями підвищує якість розпізнавання мови.

Метою роботи є розробка системи зчитування мови по губах за допомогою нейронної мережі з невеликою кількістю вимовлених слів на даному відрізку відеоряду.

У дипломній роботі за допомогою бібліотек для машинного навчання keras створена згорткова нейронна мережа, яка має п'ять шарів згортки, два прихованих повнозв'язних шари з функцією активації reLU і вихідний шар з функцією активації softmax. У якості навчальної множини було використано набір даних MIRACL, який складається з 3000 зразків 20 унікальних слів. Після навчання нейронної мережі, тестування системи показало точність 55%.

Перспективами розвитку є навчання згорткової нейронної мережі на множині з більшою кількістю унікальних слів і зразків.

Використати дану нейронну мережу можливо як доповнення до іншої технології розпізнавання мови або як самостійну технологію читання по губах у сфері безпеки або послуг.

ABSTRACT

In our days, technologies focused on automatization and making human lives easier. All developments in machine learning and neural networks focused on solving tasks, that are easily solved by people, but are very hard to be computed. One of such tasks is to translate human speech into text, but there are situations when this task becomes much harder, for example loud noise, fast speech, bad pronunciation, or no sound on video at all. In these cases lip reading is the way to increase accuracy. This task itself is much harder for people and computer to solve, and do not have good enough solutions, but as a part of combination with other technologies it increase accuracy of speech recognition.

Graduate works purpose is to create a system that lip-reads few words from video, using neural network.

In graduate work convolutional neural network had been created, using library for machine-learning Keras. It consists of five convolutional layers, two hidden layers with reLU activation function and output layer with softmax activation function. MIRACL dataset had been used for neural networks learn. It includes 3000 samples of 20 unique words. After neural network learning process, tests showed 55% system prediction accuracy.

This system can be used as an addition to another speech recognition technology, or as solo self-found system in security and service fields of usage.

АННОТАЦИЯ

В современном мире технологии непрерывно двигаются в направлении автоматизации и облегчения жизни человека. Все разработки в сфере машинного обучения и нейронных сетей направлены на решение задач, которые легко решаются человеком, но сложны с математической точки зрения. Одной из таких проблем является преобразование речи человека в текст, но иногда возникают ситуации, в которых эта задача становится очень сложной. Например, в громкой обстановке, когда человек говорит слишком быстро или неразборчиво, или звук по тем или иным причинам полностью отсутствует. В таких случаях повысить качество распознавание речи может чтение по губам. Это не тривиальная задача как для человека, так и для компьютера. Поэтому само по себе чтение по губам не дает высоких результатов в распознавании речи, но вместе с другими технологиями повышает качество распознавания речи.

Цель работы – разработать систему считывания речи по губам с помощью нейронной сети, на отрезке видеоряда, с небольшим количеством слов.

В дипломной работе с помощью библиотек для машинного обучения keras создана свёрточная нейронная сеть, состоящая из пяти слоев свертки, двух скрытых слоев с активационной функцией reLU, и выходным слоем с активационной функцией softmax. В качестве обучающего множества был использован набор данных MIRACL, состоящий из 3000 образцов и 20 уникальных слов. После обучения нейронной сети тестирование показало точность 55%.

Перспективы развития – обучение свёрточной нейронной сети на обучающем множестве с большим количеством образцов и уникальных слов.

Использовать данную нейронную сеть можно, как дополнение к другой технологии распознавания речи, так и как самостоятельную систему, в сфере безопасности или обслуживания.

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ, УМОВНИХ ПОЗНАЧЕНЬ ТА ТЕРМІНІВ	6
ВСТУП	7
1 ОГЛЯД СИСТЕМ ТА МЕТОДІВ РОЗПІЗНАВАННЯ МОВИ.....	9
1.1 Огляд аналогічних систем автоматичного читання по губах	9
1.2 Огляд методів розпізнавання.....	13
1.2.1 Прихована марковська модель	13
1.2.2 Нейронна мережа	15
1.3 Уточнення постановки задачі.....	18
2 ПРОЕКТУВАННЯ СИСТЕМИ РОЗПІЗНАВАННЯ МОВИ	19
2.1 Опис згорткової нейронної мережі.....	20
2.2 Вибір DataSet.....	23
2.3 Вибір засобів реалізації.....	23
3 ПРОГРАМНА РЕАЛІЗАЦІЯ СИСТЕМИ РОЗПІЗНАВАННЯ МОВИ..	25
3.1 Опис коду	25
3.2 Інструкція роботи програми	27
4 ТЕСТУВАННЯ СИСТЕМИ РОЗПІЗНАВАННЯ МОВИ.....	30
ВИСНОВКИ.....	32
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ	33
ДОДАТОК А КОД МОДЕЛІ НЕЙРОННОЇ МЕРЕЖІ.....	34
ДОДАТОК Б КОД ОБРОБКИ ВІДЕО З ВЕБКАМЕРИ	36

ПЕРЕЛІК СКОРОЧЕНЬ, УМОВНИХ ПОЗНАЧЕНЬ ТА ТЕРМІНІВ

Скорочення:

CTC – Connectionist Temporal Classification, нейро-мережева темпоральна класифікація

DA-RNN – Dual-Stage Attention-Based Recurrent Neural Network

HCNN – Hahn Convolutional Neural Network

SDK – Software Development Kit

V2P – video2phoneme

LSTM – Long short-term memory Довга короткочасна пам'ять

ЗНМ, CNN – згорткова нейрона мережа

КШН – коефіцієнт швидкості навчання

НМ – нейрона мережа

ШНМ – штучна нейронна мережа

Читання по губах, важлива частина людського спілкування. Як показали експерименти 1976 року, сприймання інформації порушується якщо накласти на рухи губ не оригінальний звуковий ряд. Для людини читання по губах – це важлива навичка, яку вона використовує на рефлексорному рівні, особливо добре розвинута у глухо-німих людей, а використання цієї навички на свідомому рівні має велику сферу застосування, наприклад: розвідка, безпека, перекладач для глухих та інші.

У області комп'ютерних технологій сфера застосування читання по губах теж має великий практичний потенціал. Наприклад, медичні слухові апарати, біометрія, камери стеження, які навіть в галасливій обстановці зможуть прочитати по губах розмови для забезпечення державної безпеки. Так само, читання по губах підвищує точність розпізнавання мови по засобу аналізу звуку в середньому зменшуючи похибку на 3-5%. Якщо з урахуванням похибки сучасних алгоритмів, у яких похибка приблизно 15%, зменшує її на 20-30%.

Незважаючи на те, що і сфера застосування, і важливість самої технології дуже велика, за багато років, вчені так і не змогли досягти особливих успіхів в цій області. Навіть для відносно простої з фонетичної точки зору, англійської мови, і на сьогоднішній день, точність розпізнавання навіть найкращих систем залишає бажати кращого.

Розпізнати мову за допомогою міміки людини, це складне завдання, як для самої людини, так і для комп'ютера. Багато фонем зовні дуже схожі. Адже, багато букв формуються не тільки положенням губ, але і формою і положенням язика, який у багатьох буквах розташовується за зубами або непомітний в зв'язку з поганою видимістю і однотонністю зображення, тому не дивно, що людина дуже погано справляється з читанням по губах. Навіть найкращі в читанні серед інвалідів зі слуху справляються з цим завданням тільки з точністю в 16 $\pm$ 12% для однокладових слів і 22 $\pm$ 11% для

багатоскладових, і вловлювати зміст і будувати речення доводиться з припущень про контексті сказаного так, що про точні сприйняття всієї інформації мови йти не може.

Читання по губах – одне із завдань машинного зору, яке зводиться до обробки відео з кадрів для зчитування просторово-часових лицьових маркерів, але завдання сильно ускладнюють постійний рух особи, низька якість відеоматеріалу, а також неточність технологій машинного зору, які відстежують стан і рух особи в кадрі. Таким чином останні розробки в цій сфері змогли розпізнати тільки деякі окремі слова, але не цілісні речення.

Метою роботи є розробка системи зчитування мови по губах за допомогою нейронної мережі з невеликою кількістю вимовлених слів на даному відрізку відеоряду.

В рамках даної роботи були поставлені наступні задачі:

- провести огляд існуючих підходів автоматичного читання по губах;
- виділити і описати основні етапи роботи алгоритмів читання по губах;
- розробити систему автоматичного читання по губах, засновану тільки на візуальній інформації;
- перевірити якість роботи запропонованої системи.

Розпізнавання мови засобом зорового сприйняття читання по губах дуже важлива технологія і має широку сферу використання. Але сама по собі задача читання по губах є дуже складною як для людини так і для комп'ютера. Навіть найостанніші розробки в цій сфері не дають досить точні результати, що б на високому рівні переводити візуальну інформацію по міміці особи в текст. Але при використанні інших технологій – таких як аудіо розпізнавання мови, можуть підвищити точність, особливо в разі галасу або при недостатній гучності вихідного аудіо матеріалу.

В ході дипломної роботи розглянути і порівняні кращі з існуючих на сьогоднішній день рішень для візуального розпізнавання мови. Майже всі використовують одну і ту ж модель, але кожна має особливість, яка так чи інакше дозволяє досягти кращих результатів на різних навчальних множинах. Так LipNet показує точність більш 93% але на множині, що складається з 50 унікальних слів. V2P від DeepMind google хоч і не може похвалитися такою точністю, за те може розпізнавати більше 127 тисяч унікальних слів.

У дипломній роботі спроектована і реалізована система зчитування мови по губах за допомогою згорткової нейронної мережі з невеликою кількістю вимовлених слів на даному відрізку відеоряду. Модель згорткової нейронної мережі навчена та протестована на множині MIRACL. Тестування системи показало точність 55%.

Перспективами розвитку є навчання згорткової нейронної мережі на множині з більшою кількістю унікальних слів і зразків.

Використати дану нейронну мережу можливо як доповнення до іншої технології розпізнавання мови або як самостійну технологію читання по губах у сфері безпеки або послуг

За результатами дипломної роботи підготовлені і опубліковані тези на конференцію [10]

СПИСОК ВИКОРИСТАННИХ ДЖЕРЕЛ

1. LipNet [Електронний ресурс] – Режим доступу до ресурсу: <https://habr.com/ru/post/398901/>.
2. Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel: Backpropagation Applied to Handwritten Zip Code Recognition, Neural Computation, 1(4):541-551, Winter 1989.
3. Romanuke, Vadim. Appropriate number and allocation of ReLUs in convolutional neural networks (англ.) // Research Bulletin of NTUU “Kyiv Polytechnic Institute”
4. MIRACL data set [Електронний ресурс] – Режим доступу до ресурсу: <https://sites.google.com/site/achrafbenhamadou/-datasets/miracl-vc1>.
5. Keras guides [Електронний ресурс] – Режим доступу до ресурсу: <https://keras.io/guides/>.
6. Lip Reading with Hahn Convolutional Neural Networks Abderrahim Mesbah , Hicham Hammouchi , Aissam Berrahou , Hassan Berbia , Hassan Qjidaa , Mohamed Daoud
7. LARGE-SCALE VISUAL SPEECH RECOGNITION Brendan Shillingford, Yannis Assael, Matthew W. Hoffman, Thomas Paine, Cían Hughes, Utsav Prabhu, Hank Liao, Hasim Sak, Kanishka Rao, Lorraine Bennett, Marie Mulville, Ben Coppin, Ben Laurie, Andrew Senior, Nando de Freitas DeepMind & Google
8. Tensorflow [Електронний ресурс] – Режим доступу до ресурсу <https://www.tensorflow.org/overview>
9. Apache spark [Електронний ресурс] – Режим доступу до ресурсу <https://spark.apache.org/>
10. Яворский Е. Ю., Шпинарёва И. М разработка системы машинного чтения по губам с использованием алгоритмов глубокого обучения//// Информатика, інформаційні системи і технології - XVII всеукраїнська конференція студентів та молодих вчених: Одеса, 22 квітня 2020. – с.111