Д. В. Яценко

ВИЗУАЛЬНОЕ РАСПОЗНАВАНИЕ РЕЧИ С ИСПОЛЬЗОВАНИЕМ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ

Чтение по губам – задача распознавания текста по движению губ. Развитие технологий ведет к совершенствованию интерфейсов между человеком и компьютером. Система может применяться в биометрической идентификаций, распознавании речи нескольких человек, разработке слуховых аппаратов и т.д.

Введение

В настоящий момент отсутсвуют аналогичные системы с поддержкой русского языка. Данная система распознавания речи по губам имеет нейронную сетевую архитектуру, которая отображает последовательности видеофрагментов переменной длины в текстовые последовательности. В архитектуре системы используются пространственно-временные свёрточные нейронные сети, рекуррентная нейронная сеть типа долгой краткосрочной памяти и функция потерь нейросетевой временной классификации.

I. Подходы к решению задачи

Большинство движений артикулярного аппарата человека скрыты. Наиболее видимыми участками являются губы и нижняя челюсть. Между произнесенной фонемой и ее визуальным отображением нет однозначного соответствия. Визема – визуальное отображение фонемы. По Карпову считаем, что в русской речи существует 47 фонем (ударные и безударные варианты гласных) и 10 визем [1].

II. Основные элементы системы

Для организации свёрточной нейронной сети применяется 4 основных слоя: свёртка (CONV), субдискретизация (POOL), функция активации (RELU), полносвязный слой (FC). Архитектура CNN: INPUT–[CONV–RELU–POOL]x3–FC [2]. Математически оператор свёртки вычисляется по формуле 1.

$$[conv(X,W)]_{ij} = \sum_{c=1}^{C}\sum_{i`=1}^{K_i}\sum_{j`=1}^{K_j} W_{ci`j`} * X_{c,,i+i`,j+j`},$$

где conv – оператор свёртки; X – входное изображение, W – ядро свёртки, C – канал RGB, k_w, k_h – ширина и высота ядра свёртки. Для входа x и весов $w \in R^{Cxk_wxk_h}$, где $x_{ci'j'} = 0$ для i, j выходящими за границы изображения. i и j принимают значения $0 \leq i/j \leq X_{w/h} + 2 + P$, где P – отступ свёртки. Пространственно-временные свёрточные нейронные сети (STCNN) могут обрабатывать видеоданные путем свёртывания в течение промежутка времени [3].

III. Архитектура системы

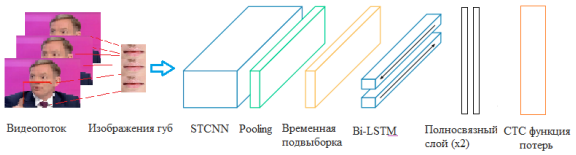

Рис. 1 – Архитектура системы распознавания речи

На вход STCNN подается последовательность T кадров (100x50 изображения губ). Далее из выделенных признаков отбирается временная выборка и обрабатывается Bi-LSTM сетью. Каждый временной шаг выход LSTM обрабатывается двумя полносвязными слоями, после чего следует softmax слой активации, который преобразует выход полносвязной сети в класс $f_{softmax} = \frac{e^{x_i^l}}{\sum_{k=0}^{n} e^{x_i^l}}$ [4]. Наконец, информация на выходе полносвязной сети подается в функцию потерь CTC [2].

IV. Выводы

Предлагаемая архитектура системы позволяет создать и обучить систему визуального распознавание речи человека.

Список литературы

1. И. Б. Тампель, А. А Карпов. Автоматическое распознавание речи. Учебное пособие. – СПб: Университет ИТМО, 2016. – 138 с.
2. Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas. Lipnet: Sentence-level lipreading. Under submission to ICLR 2017, arXiv:1611.01599v2, 2016.
3. A. Krizhevsky, I. Sutskever, G. E. Hinton. Imagenet classification with deep convolutional neural networks.
4. A. Garg, J. Noyola, and S. Bagadia. Lip reading using CNN and LSTM. Technical report, Stanford University, CS231n project report, 2016.

Яценко Дмитрий Валерьевич, магистрант кафедры ИИТ Белорусского государственного университета информатики и радиоэлектроники, d.yatsenko@sumatosoft.com.

Научный руководитель: Ивашенко Валерьян Петрович, доцент кафедры ИИТ Белорусского государственного университета информатики и радиоэлектроники, кандидат технических наук, доцент, ivashenko@bsuir.by.