УДК 004.934: 004.588

СИСТЕМА МАШИННОГО ЧТЕНИЯ ПО ГУБАМ С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМОВ ГЛУБОКОГО ОБУЧЕНИЯ

Д.В. Яценко
Магистрант кафедры интеллектуальных информационных технологий БГУИР

В.П. Ивашенко
Доцент кафедры интеллектуальных информационных технологий, кандидат технических наук

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
ООО "СуматоСофт", Республика Беларусь
E-mail: ivashenko@bsuir.by, d.yatsenko@sumatosoft.com

Аннотация. Чтение по губам — это задача распознавания текста по движению губ. Развитие компьютерной техники ведет к совершенствованию интерфейсов между человеком и компьютером. Один из важнейших способов человеческой коммуникации — речь, поэтому способность распознавания речи является важным направлением совершенствования человеко-машинных интерфейсов. Разрабатываемая модель преобразовывает последовательность видеофрагментов с переменной длиной слова в текст, используя пространственно-временные свёртки, рекуррентную нейронную сеть и функцию потерь нейросетевой временной классификации. Визуальное распознавание речи может применяться в биометрической идентификаций, распознавании речи двух и более человек, разработке улучшенных слуховых аппаратов, управлении интеллектуальными системами и т.д.

Ключевые слова: чтение по губам, алгоритмы глубокого обучения, свёрточные нейронные сети, рекуррентные нейронные сети, Keras, Python, виземы, долгая краткосрочная память

Введение. Чтение по губам – изначально сложная задача для человека. Большинство движений (действий) человеческого артикулярного аппарата скрыты и двусмысленны без контекста. Наиболее видимыми участками движения артикуляционного аппарата являются губы, нижняя челюсть и иногда кончик языка и зубы. Следовательно, показатели точности чтения по губам человека достаточно низкие. Люди с нарушениями слуха достигают точности около 17 ± 12% для ограниченного подмножества из 30 односложных слов и 21 ± 11% для 30 сложных слов [10].

Основной сложностью машинного чтения по губам является необходимость извлечения пространственно-временных характеристик из видео.

Подходы к решению задачи распознавания речи по губам. С точки зрения речеобразования, каждый звук может быть описан уникальной комбинацией нескольких основных состояний артикуляционной системы – параметрами артикуляции [10]:

- наличие или отсутствие работы голосовых связок;
- положение языка;
- параметры раскрытия рта (ширина, высота);
- округление и вытянутость губ.

При решении задачи – создание системы распознавания русской речи по артикуляции, необходимо учитывать только видимые параметры артикуляции.

Для русской речи практически полностью скрытой остаётся динамика движения тела,

кончика и боковинок языка, небной занавески, голосовых связок. Четко различаются движения губ, кончика языка и нижней челюсти, особенно при образовании гласных, а также губных согласных. Также различия наблюдается между твёрдыми и мягкими согласными, между заднеязычными и другими негубными согласными [10].

Между произнесенной фонемой и ее визуальным отображением нет однозначного соответствия. Визуальное отображение фонемы называется виземой – состояние губ, соответствующее фонемам устной речи.

В разных языках эксперты выделяют разное количество визем (также как и фонем), в английском их 12-14, в русском 10-12 в зависимости от диктора [10]. В таблице 1 показаны базовые классы визем русской речи и соответствие фонемам. По Карпову считаем, что в русской речи существует 47 фонем, включая ударные и безударные (редуцированные) варианты гласных; согласные звуки соответствуют фонетическому алфавиту SAMPA [10].

Таблица 1

10 классов визем [10], включая нейтральное положение губ

Класс визем	Тип виземы/фонемы	Соответствующие фонемы русской речи
V1	Неогубленные гласные (широкое открытие рта)	/а!/, /а/, /э!/, /э/
V2	Неогубленные гласные (остальные)	/и/, /и!/, /ы/, /ы!/
V3	Огубленные гласные звуки	/о!/, /у/, /у!/
V4	Губные согласные	/б/, /б’/, /п/, /п’/, /м/, /м’/
V5	Губно-зубные согласные	/ф/, /ф’/, /в/, /в’/
V6	Альвеолярные фрикативные согласные	/ш/, /ж/, /ч/, /щ/
V7	Альвеолярные сонорные согласные	/л/, /л’/, /р/, /р’/
V8	Зубные согласные	/д/, /д’/, /т/, /т’/, /н/, /н’/, /с/, /с’/, /з/, /з’/, /ц/
V9	Заднеязычные согласные	/г/, /г’/, /к/, /к’/, /х/, /х’/, /й/
V10	Пауза (нейтральное положение губ)	тишина (пауза)

Основные компоненты и устройство системы распознавания речи по губам. Разрабатываемая система распознавания речи по губам имеет нейронную сетевую архитектуру для обработки речи, которая отображает последовательности видеофрагментов переменной длины в текстовые последовательности. В архитектуре системы используются пространственно-временные свёрточные нейронные сети [3] (STCNN), рекуррентная нейронная сеть типа долгой краткосрочной памяти [7] (LSTM) и функция потерь нейросетевой временной классификации [4] (CTC).

Свёрточные нейронные сети (CNN) — архитектура искусственных нейронных сетей, нацеленная на эффективное распознавание изображений, входит в состав технологий глубокого обучения. Ядро свёртки — матрица весов небольшого размера, которую «двигают» по

всему обрабатываемому слою входному изображению, формируя после каждого сдвига сигнал активации для нейрона следующего слоя с аналогичной позицией. [3]

Для организации свёрточной нейронной сети применяется 3 основных слоя:

1. Слой свёртки (CONV);
2. Слой субдискретизации (POOL);
3. Функция активации (RELU);
4. Полносвязный слой (FC).

Эти слои используются с целью построения полной архитектуры CNN. В данной системе архитектурой CNN является:

INPUT — [CONV — RELU — POOL]x3 — FC

Операция свёртки изображения — операция между матрицей изображения и ядром свёртки (фильтром), при которой каждый элемент (пиксель) в выходном изображении есть сумма произведений значения элемента ядра на значение соответствующего покрываемого элемента матрицы входящего изображения [4]. Математически операция свёртки записывается при помощи формулы 1:

$$C_{ij} = (X * K)_{ij} = \sum_{a=0}^{k-1} \sum_{b=0}^{k-1} x_{i+a,j+b}\, k_{ab}, 0 \le i, j < x - k + 1, \quad (1)$$

где: *X* – матрица размера *x* на *x*, задающая входное изображение; *K* – матрица размера *k* на *k*, которая называется ядром свёртки; *C* – результат свёртки матрицы *X* с ядром *K*.

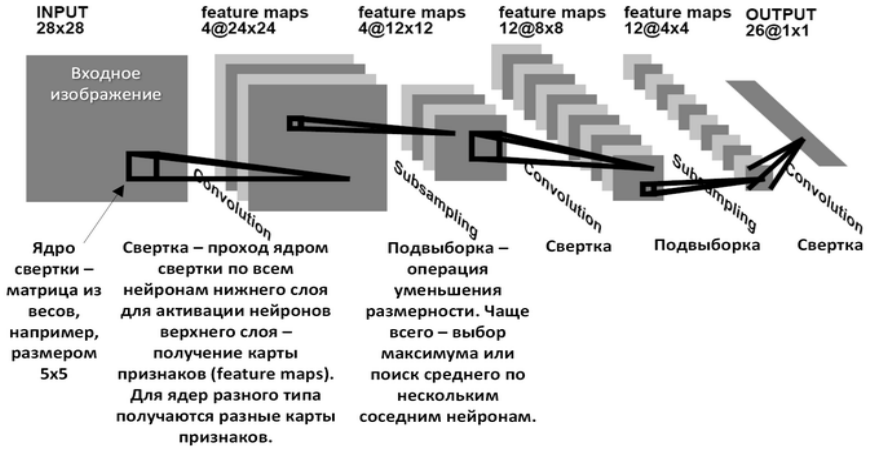

Рисунок 1. Работа свёрточной нейронной сети

Свёрточный слой, вычисляющий свёртку канала C вычисляется по формуле 2.

$$[conv(X, W)]_{ij} = \sum_{c=1}^{C} \sum_{i\prime=1}^{k_w} \sum_{j\prime=1}^{k_h} w_{ci\prime j\prime} * x_{c,i+i\prime,j+j\prime}, \quad (2)$$

где: *conv* — оператор свёртки, *x* — входное изображение, *w* — вектор весов (ядро свёртки), *C* — канал RGB, k_w — ширина ядра свёртки, k_h — высота ядра свёртки. Для входа x и весов $w \in R^{C \times kw \times kh}$, где $x_{cij} = 0$ для i, j выходящими за границы изображения. i и j принимают значения $0 \le i < X_w + 2 * P$, $0 \le j < X_h + 2 * P$, где P – отступ свёртки.

Шаг (stride) — определяет, на сколько смещается ядро на каждом шаге при вычислении следующего пикселя результирующего изображения. Обычно его принимают равным 1, и чем больше его значение, тем меньше размер выходного изображения;

Пространственно-временные свёрточные нейронные сети (STCNN) могут обрабатывать видеоданные путем свёртывания в течение промежутка времени как показано в формуле 3.

$$[stconv(X,W)]_{ij} = \sum_{c=1}^{C} \sum_{t\prime=1}^{k_t} \sum_{i\prime=1}^{k_w} \sum_{j\prime=1}^{k_h} w_{ct\prime i\prime j\prime} * x_{c,t+t\prime,i+i\prime,j+j\prime}, \quad (3)$$

где: *stconv* — оператор пространственно-временной свёртки; k_t — кол-во временных кадров [2].

В качестве слоя активации сигмоида используется только если классов не больше двух. Для задачи, в которой необходимо распознать количество классов > 2, используется функция активации softmax. Функция softmax представлена в формуле 4. Выход модели отражает вероятность принадлежности входных данных всем классам. Сумма вероятностей по всем выходам сети равняется единице.

$$f_{softmax} = {y^l}_i = f({x^l}_i) = \frac{e^{{x^l}_i}}{\sum_{k=0}^{n} e^{{x^l}_k}}, \quad (4)$$

где: n — это количество классов.

Популярный способ субдискретизации изображения — слой *подвыборки* (*pooling* layer), выполняет уменьшение размерности сформированных карт признаков.

При архитектурах сети такого типа принято считать, что информация о факте наличия искомого признака важнее точного знания его координат. За счёт данной операции, помимо ускорения дальнейших вычислений, сеть становится более инвариантной к масштабу входного изображения.

Формально max-pooling слой может быть описан формулой 5.

$$x^l = f(a^l * subsample(x^{l-1}) + b^l), \quad (5)$$

где: x^l — выход слоя l, $f()$ — функция активации, a,b — коэффициенты, $subsample(\text{x}) = max(x)$ — операция выборки локальных максимальных значений [5].

Слой FC (полносвязный слой) выводит N-мерный вектор (N — число классов) для определения нужного класса. Работа организуется путем обращения к выходу предыдущего слоя (карте признаков) и определения свойств, которые наиболее характерны для определенного класса.

Долгая краткосрочная память (*Long short-term memory*; *LSTM*) — разновидность архитектуры рекуррентных нейронных сетей (RNN). LSTM-сеть хорошо приспособлена к обучению на задачах классификации, обработки и прогнозирования временных рядов в случаях, когда важные события разделены временными лагами с неопределенной продолжительностью и границами [7].

LSTM-блоки содержат три или четыре «вентиля», которые используются для контроля потоков информации на входах и на выходах памяти данных блоков. Эти вентили реализованы в виде логистической функции для вычисления значения в диапазоне [0; 1]. Умножение на это значение используется для частичного допуска или запрещения потока информации внутрь и наружу памяти. Например, «входной вентиль» контролирует меру вхождения нового значения в память, а «вентиль забывания» контролирует меру сохранения значения в памяти. «Выходной вентиль» контролирует меру того, в какой степени значение, находящееся в памяти, используется при расчёте выходной функции активации для блока.

Веса в LSTM-блоке (W и U) используются для задания направления оперирования вентилей. Эти веса определены для значений, которые подаются в блок (включая x_t и выход с предыдущего временного шага h_{t-1}) для каждого из вентилей. Таким образом, LSTM-блок определяет, как распоряжаться своей памятью как функцией этих значений, и тренировка весов позволяет LSTM-блоку выучить функцию, минимизирующую потери. LSTM-блоки

обычно тренируют при помощи метода обратного распространения ошибки во времени [4]. Работа LSTM сети представлена на рисунке 6 и описана математически в формуле 6.

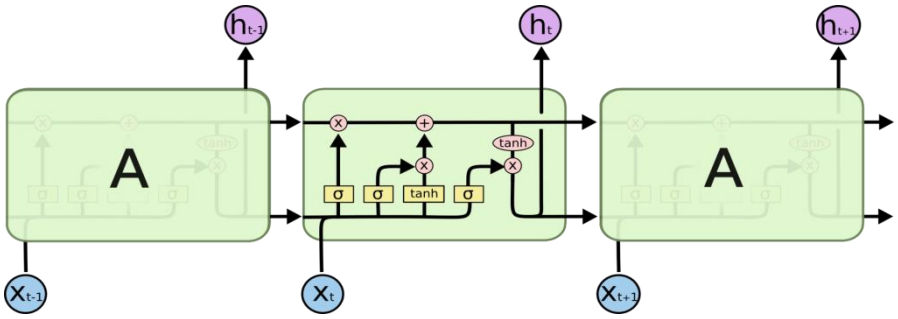

Рисунок 2. Работа LSTM сети

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f)$$

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i)$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \quad (6)$$

$$c_t = f_g \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c)$$

$$h_t = o_t \circ \sigma_h(c_t)$$

где: $\sigma(x) = \frac{1}{1+exp(-x)}$ - логистическая функция, ∘ - обозначает умножение по элементам, $x_t = \{x_1, x_2, \dots, x_T\}$— последовательность на входе, $h_t = \{h_1, h_2, \dots, h_T\}$ — выходной вектор, c_t — вектор состояний, *W*, *U* и *b* — матрицы параметров и вектор смещений, f_t — вектор вентиля забывания, вес запоминания старой информации; i_t — вектор входного вентиля, вес получения новой информации; o_t — вектор выходного вентиля, кандидат на выход [4].

Традиционная LSTM с вентилями забывания, где $c_0 = 0$ и $h_0 = 0$.

Используя Bi-LSTM сеть в задачах по разметке последовательности мы имеем доступ и к прошлым, и к будущим входным данным в течение заданного времени. Таким образом, мы можем использовать двунаправленную LSTM сеть для активации и обучения в задачах разметки последовательностей [8].

Функция потерь нейросетевой временной классификации (Connectionist temporal classification loss, CTC) представляет собой функцию оценки вывода нейронной сети, для обучения рекуррентных нейронных сетей (RNN), таких как сети LSTM, для решения проблем последовательности вывода, где время является переменной величиной. CTC широко используется в задачах распознавание речи, поскольку устраняет необходимость в предварительной подготовке данных.

Вход представляет собой последовательность наблюдений, а выходы представляют собой последовательность дискретных распределений по классам (словарь). Словарь классов расширяется специальным классом “Пустой”.

CTC вычисляет вероятность конкретной последовательности фонем на каждом временном шаге, путем суммирования всех возможных последовательностей, эквивалентных данной.

Для обучения CTC сети можно использовать алгоритм обратного распространения ошибки. Трудность обучения заключается в том, что количество наблюдений намного превышает количество классов.

Пусть V обозначает множество классов, которые модель классифицирует на одном временном шаге своего вывода, а словарь, дополненный классом “Пустой”, обозначается $\tilde{V} = V \cup \{\_\}$, где _ обозначает CTC класс “Пустой”.

Определим функцию $B : \tilde{V}^* \rightarrow V^*$, которая при получении строки удаляет соседние повторяющиеся символы и устраняет пустые символы, где * означает операцию замыкания Клини; {_} – пустой класс.

Для последовательности меток $y \in V^*$, CTC определяет $p(y|x)$ – вероятность принадлежности классу входных данных x, что описано в формуле 7.

$$p(y|x) = \sum_{u \in B^{-1}(y) s.t. |u|=T} p(u_1, \ldots, u_T|x), \quad (7)$$

где: T – это количество временных шагов в последовательности модели; $|u|$ – мощность множества u и равняется, если u конечно, числу элементов u.

Пример: Если $T = 3$, то CTC определяет вероятность строки “до” как p(ддо) + p(доо) + p(_до) + p(д_о) + p(до_) [4].

На рисунке 3 изображена архитектура системы.

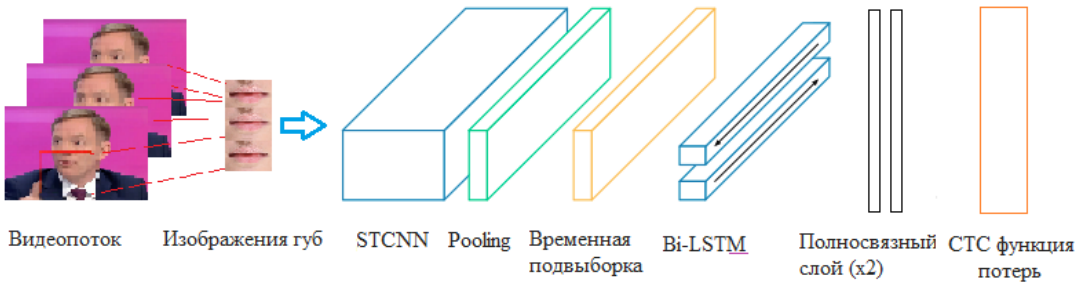

Рисунок 3. Архитектура системы распознавания речи по губам

Применение и основные этапы работы системы распознавания речи по губам.Последовательность T кадров подается на вход и обрабатывается 3-я слоями STCNN, за каждым из которых следует слой субдискретизации, или max-pooling.

Далее из выделенных признаков отбирается временная выборка (Temporal Upsampling слой) и обрабатывается Bi-LSTM сетью. Каждый временной шаг выход LSTM обрабатывается двумя полносвязными слоями, после чего следует softmax слой активации. Активационный слой преобразует выход полносвязной сети в класс из словаря. Наконец, информация на выходе полносвязной сети подается в функцию потерь CTC. На всех слоях используются функции активации ReLU [6].

Каждое изображение содержит большое количество исходной информации, которая не используется в задаче распознавания речи. Поэтому требуется обрабатывать каждое изображение и выделять область губ. Для выделения ключевых точек лица на изображении используется фреймворк Dlib [11] и предварительно обученная модель оценки положения лица shape_predictor_68_face_landmarks.dat, обученная на коллекции изображений iBUG 300W [12]. Используя эти признаки, алгоритм извлекает отцентрированные по губам изображения размером 100 × 50 пикселей [9].

Для того, чтобы увеличить количество данных для обучения проводится расширение данных. Обучение проводится на основании исходных и горизонтально отраженных изображений [1].

Список литературы

[1]. A. Garg, J. Noyola, and S. Bagadia. Lip reading using CNN and LSTM. Technical report, Stanford University, CS231n project report, 2016.

[2]. A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1725–1732, 2014.

[3]. A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pp. 1097–1105, 2012.

[4]. Felix A. Gers; Jürgen Schmidhuber; Fred Cummins (2000). "Learning to Forget: Continual Prediction with LSTM". Neural Computation. 12 (10): 2451–2471. doi:10.1162/089976600300015015.

[5]. Graham, Benjamin, "Fractional Max-Pooling", arXiv:1412.6071, 2014

[6]. R. D. Easton and M. Basala. Perceptual dominance during lipreading. Perception & Psychophysics, 32(6): 562–570, 1982.

[7]. S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.

[8]. Schuster, Mike, and Kuldip K. Paliwal. "Bidirectional recurrent neural networks." IEEE Transactions on Signal Processing 45.11 (1997): 2673-2681.

[9]. Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas. Lipnet: Sentence-level lipreading. Under submission to ICLR 2017, arXiv:1611.01599v2, 2016.

[10]. Тампель И.Б., Карпов А.А. АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ РЕЧИ. Учебное пособие. − СПб: Университет ИТМО, 2016. – 138 с.

[11]. http://dlib.net/imaging.html

[12]. https://ibug.doc.ic.ac.uk/resources/300-W/

MACHINE LIP READING SYSTEM USING DEEP LEARNING ALGORITHMS

D.V. YATSENKO
Master student of the Belarusian State University of Informatics and Radiolectronics

V.P. IVASHENKO, PhD
Associate Professor of the Department of Intelligent Information Technologies

Belarusian State University of Informatics and Radioelectronics, Republic of Belarus
SumatoSoft LTD, Republic of Belarus
E-mail: ivashenko@bsuir.by, d.yatsenko@sumatosoft.com

Abstract. Lip reading is the task of recognizing text by lips movement. The development of computer technology leads to the improvement of interfaces between human and computer. One of the most important ways of human communication is speech, so the ability to recognize speech is an important direction in improving human-machine interfaces. The model being developed transforms a sequence of variable-length video sequences into text using spatiotemporal convolutions, a recurrent neural network, and a loss neural network temporal classification function. Visual recognition of speech can be used in biometric identification, speech recognition of two or more people, the development of improved hearing aids, the management of intelligent systems, etc.

Key words: lipreading, deep learning algorithms, CNN, RNN, Keras, Python, visems, LSTM.