НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
март–апрель 2016 Том 16 № 2 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
March–April 2016 Vol. 16 No 2 ISSN 2226-1494 http://ntv.ifmo.ru/en

УДК 004.522

ДВУХЭТАПНЫЙ АЛГОРИТМ ИНИЦИАЛИЗАЦИИ ОБУЧЕНИЯ АКУСТИЧЕСКИХ МОДЕЛЕЙ НА ОСНОВЕ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ

И.П. Меденников[a,b]

[a] ООО «ЦРТ-инновации», Санкт-Петербург, 196084, Российская Федерация
[b] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
Адрес для переписки: ipmsbor@yandex.ru

Информация о статье
Поступила в редакцию 19.11.15, принята к печати 18.01.16
doi:10.17586/2226-1494-2016-16-2-379-381
Язык статьи – русский

Ссылка для цитирования: Меденников И.П. Двухэтапный алгоритм инициализации обучения акустических моделей на основе глубоких нейронных сетей // Научно-технический вестник информационных технологий, механики и оптики. 2016. Т. 16. № 2. С. 379–381. doi:10.17586/2226-1494-2016-16-2-379-381

Аннотация
Предложен двухэтапный алгоритм инициализации обучения акустических моделей на основе глубоких нейронных сетей. Алгоритм предназначен для уменьшения влияния сегментов, не содержащих речь, на обучение акустической модели. Идея предлагаемого подхода заключается в уменьшении доли неречевых примеров в обучающей выборке. Оценка эффективности алгоритма выполнена на задаче распознавания английской спонтанной речи в телефонном канале (Switchboard). Применение предложенного алгоритма позволило добиться 3% относительного уменьшения пословной ошибки распознавания по сравнению с инициализацией обучения при помощи ограниченных машин Больцмана. Результаты работы могут найти применение при разработке систем автоматического распознавания речи.

Ключевые слова
автоматическое распознавание речи, глубокие нейронные сети

TWO-STEP ALGORITHM OF TRAINING INITIALIZATION FOR ACOUSTIC MODELS BASED ON DEEP NEURAL NETWORKS

I.P. Medennikov[a,b]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation
[b] STC-Innovations Ltd., Saint Petersburg, 196084, Russian Federation
Corresponding author: ipmsbor@yandex.ru

Article info
Received 19.11.15, accepted 18.01.16
doi:10.17586/2226-1494-2016-16-2-379-381
Article in Russian
For citation: Medennikov I.P. Two-step algorithm of training initialization for acoustic models based on deep neural networks. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2016, vol. 16, no. 2, pp. 379–381. doi:10.17586/2226-1494-2016-16-2-379-381

Abstract
This paper presents a two-step initialization algorithm for training of acoustic models based on deep neural networks. The algorithm is focused on reducing the impact of the non-speech segments on the acoustic model training. The idea of the proposed algorithm is to reduce the percentage of non-speech examples in the training set. Effectiveness evaluation of the algorithm has been carried out on the example of English spontaneous telephone speech recognition (Switchboard). The application of the proposed algorithm has led to 3% relative word error rate reduction, compared with the training initialization by restricted Boltzmann machines. The results presented in the paper can be applied in the development of automatic speech recognition systems.

Keywords
automatic speech recognition, deep neural networks

Акустическая модель является важным компонентом системы автоматического распознавания речи, отвечающим за описание плотности распределения акустических классов (например, фонем) на заданном участке речевого сигнала. Одним из наиболее часто используемых в современных системах автоматического распознавания речи типов акустических моделей являются акустические модели на основе

глубоких нейронных сетей (DNN) [1, 2]. Глубокие нейронные сети также активно используются во многих других предметных областях, например, в задачах распознавания образов [3].

Инициализация обучения DNN оказывает большое влияние на качество обучения. В настоящей работе предложен алгоритм инициализации обучения акустических моделей на основе глубоких нейронных сетей. Основой для него послужило наблюдение, что сегменты, не содержащие речи, составляют значительную долю в фонограммах, на которых осуществляется обучение акустических моделей. При анализе фонемной покадровой разметки фонограмм из обучающего корпуса английской спонтанной речи Switchboard[1] [4], оказалось, что около 25% кадров в разметке составляют неречевые фонемы (пауза, шум). По этой причине при обучении DNN по критерию минимизации взаимной энтропии может возникать ситуация, когда качество классификации неречевых фонем улучшается в ущерб качеству классификации речевых фонем и, следовательно, в ущерб качеству распознавания речи. Предложенный алгоритм направлен на уменьшение влияния этого эффекта и состоит из двух этапов.

1. Осуществляется предобучение DNN одним из следующих способов: при помощи ограниченных машин Больцмана [5], автоэнкодеров [6] или дискриминативного алгоритма предобучения [7].
2. Полученная на первом этапе предобученная DNN используется для инициализации обучения по критерию минимизации взаимной энтропии на сбалансированной по количеству неречевых примеров обучающей выборке. Балансировка происходит следующим образом: из обучающих примеров, соответствующих неречевым фонемам, случайным образом выбирается некоторая их часть так, чтобы количество примеров для неречевых фонем в обучающей выборке было примерно равным среднему количеству примеров для одной речевой фонемы.

DNN, полученная на втором этапе алгоритма, в дальнейшем используется для инициализации обучения по полной обучающей выборке. При этом чтобы избежать ухудшения качества классификации речевых фонем, следует уменьшать скорость обучения (learning rate, [1]) DNN, а также использовать L_2-штраф на отклонение параметров DNN от значений параметров инициализирующей нейронной сети. Это способствует улучшению качества классификации неречевых фонем без большого ущерба для качества классификации речевых фонем, что позволяет повысить точность распознавания речи.

Экспериментальная оценка эффективности предложенного алгоритма проводилась для задачи распознавания английской спонтанной речи в телефонном канале. Для обучения использовался корпус Switchboard (300 часов речи), тестирование проводилось на подвыборке Switchboard базы HUB5 Eval 2000[2]. Для экспериментов использовался набор инструментов Kaldi ASR[3] [8, 9].

Обучены две акустические модели *dnn*1 и *dnn*2 на основе DNN с 6 скрытыми слоями по 2048 нейронов в каждом с сигмоидами в качестве функций активации. Для обучения DNN использовалась покадровая разметка на связанные состояния трифонов, сделанная при помощи модели *tri*4 из рецепта Kaldi *swbd s*5*c*. В качестве признаков были взяты логарифмы мощности выходов 23 треугольных Мел-частотных фильтров (FBANK), дополненные первыми и вторыми производными и взятые с временным контекстом в 11 кадров.

Модель *dnn*1 обучена по критерию минимизации взаимной энтропии с инициализацией обучения при помощи ограниченных машин Больцмана. Обучение модели *dnn*2 инициализировалось с использованием разработанного двухэтапного алгоритма. На первом этапе использовалось предобучение при помощи ограниченных машин Больцмана. На втором этапе проводилось обучение по критерию минимизации взаимной энтропии по сбалансированной обучающей выборке, из которой случайным образом были выброшены 98% неречевых обучающих примеров. Полученная на втором этапе глубокая нейронная сеть использовалась для инициализации обучения по полной обучающей выборке с уменьшенной в 4 раза скоростью обучения. Использовался L_2-штраф величины $4 \cdot 10^{-8}$ на отклонение параметров DNN от значений инициализирующей модели.

Модель *dnn*1 продемонстрировала пословную ошибку распознавания (Word Error Rate, WER) 16,4%, модель *dnn*2 – WER 15,9%. Полученные результаты говорят об эффективности предложенного алгоритма инициализации обучения в задаче распознавания английской спонтанной речи в телефонном канале: его применение позволило добиться 0,5% абсолютного и 3,0% относительного уменьшения пословной ошибки распознавания. Полученный результат на корпусе Switchboard оказался лучше, чем результат зарубежных исследователей, приведенный в работе [10] (16,1% WER), что свидетельствует в пользу эффективности предложенного алгоритма.

Литература

1. Hinton G., Deng L., Yu D., Dahl G., Mohamed A.-R., Jaitly N., Senior A., Vanhoucke V., Nguyen P., Sainath T., Kingsbury B. Deep neural networks for acoustic modeling in speech recognition: the shared views

[1] https://catalog.ldc.upenn.edu/LDC97S62
[2] https://catalog.ldc.upenn.edu/LDC2002S09
[3] http://www.kaldi-asr.org

of four research groups // IEEE Signal Processing Magazine. 2012. V. 29. N 6. P. 82–97. doi: 10.1109/MSP.2012.2205597
2. Dahl G.E., Yu D., Deng L., Acero A. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition // IEEE Transactions on Audio, Speech and Language Processing. 2012. V. 20. N 1. P. 30–42. doi: 10.1109/TASL.2011.2134090
3. Потапов А.С., Батищева В.В., Пан Ш. Улучшение качества распознавания в сетях глубокого обучения с помощью метода имитации отжига // Научно-технический вестник информационных технологий, механики и оптики. 2014. № 5 (93). С. 71–76.
4. Godfrey J., Holliman E., McDaniel J. Switchboard: telephone speech corpus for research and development // Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). San Francisco, USA, 1992. V. 1. P. 517–520. doi: 10.1109/ICASSP.1992.225858
5. Hinton G.E., Osindero S., Teh Y.-W. A fast learning algorithm for deep belief nets // Neural Computation. 2006. V. 18. N 7. P. 1527–1554. doi: 10.1162/neco.2006.18.7.1527
6. Vincent P., Larochelle H., Bengio Y., Manzagol P.-A. Extracting and composing robust features with denoising autoencoders // Proc. 25th International Conference on Machine Learning. Helsinki, Finland, 2008. P. 1096–1103.
7. Bengio Y., Lamblin P., Popovici D., Larochelle H. Greedy layer-wise training of deep networks // Proc. 20th Annual Conf. on Neural Information Processing Systems (NIPS 2006). Vancouver, Canada, 2006. P. 153–160.
8. Povey D., Ghoshal A., Boulianne G., Burget L., Glembek O., Goel N., Hannemann M., Motlicek P., Qian Y., Schwarz P., Silovsky J., Stemmer G., Vesely K. The Kaldi speech recognition toolkit // Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Waikoloa, USA, 2011. P. 1–4.
9. Vesely K., Ghoshal A., Burget L., Povey D. Sequence-discriminative training of deep neural networks // Proc. of the Annual Conference of International Speech Communication Association (INTERSPEECH). Lyon, France, 2013. P. 2345–2349.
10. Seide F., Li G., Yu D. Conversational speech transcription using context-dependent deep neural networks // Proc. of the Annual Conference of International Speech Communication Association (INTERSPEECH). Florence, Italy, 2011. P. 437–440.

Меденников Иван Павлович – научный сотрудник, ООО «ЦРТ-инновации», Санкт-Петербург, 196084, Российская Федерация; инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, ipmsbor@yandex.ru

Ivan P. Medennikov – scientific researcher, STC-Innovations Ltd., Saint Petersburg, 196084, Russian Federation; engineer, ITMO University Saint Petersburg, 197101, Russian Federation, ipmsbor@yandex.ru