Listen, Attend and Spell

William Chan et al., cs.CL, 2015

translated by. triplet02

0. 요약(Abstract)

우리는 음성 발성을 문자로 변환하도록 학습하는 신경망 구조인 Listen, Attned and Spell(LAS) 모델을 제안한다. 전통적인 HMM-DMM 모델과 달리, 이 모델은 음성 인식기(speech recognizer)의 모든 요소들(components)을 공동으로(jointly) 학습한다. 이 시스템은 두 가지 요소, 리스너(Listener)와 스펠러(Speller)로 구성된다. 리스너는 필터 뱅크 스펙트럼들(filter bank spectra)을 입력으로 하는, 피라미드형으로 쌓은 순환신경망(recurrent network, RNN)이다. 스펠러는 문자(characters)를 출력으로 하는 어텐션 기반(attention-based) 순환신경망이다. 이 신경망은 문자들 사이에 어떠한 독립적인 가정(independence assumptions)도 하지 않은 상태로 문자 시퀀스를 생성한다. 이것이 LAS 모델이 기존의 엔드투엔드 CTC 모델에서 핵심적으로 개선된 부분이다. 구글 음성 탐색 과제(Google voice search task)에서 LAS는 사전이나 언어 모델 없이 14.1%의 단어 오류율(Word Error Rate, WER)을 기록하였고, 상위 32개 빔(Beam)으로 재조정(rescoring)되는 언어 모델을 사용하여 10.3%의 WER을 기록하였다. 본 모델과 비교하면, 최고 성능(state-of-the-art) CLDNN-HMM 모델은 8.0%의 WER을 기록하였다.

1. 도입(Introduction)

심층신경망(Deep Neural Networks, DNNs)은 음성 인식기의 다양한 요소를 발전시키는 데 기여했다. 그것들은 음향 모델링(acoustic modeling)을 위한 하이브리드 DNN-HMM 음성 인식 시스템에서 널리 쓰여왔다[1, 2, 3, 4, 5, 6]. DNNs는 또한, 단어를 음소(phoneme) 시퀀스로 맵핑(map)하는 발음 모델(pronunciation model)에게 현저한 이익을 가져다주었다[7, 8]. 언어 모델링에서, 순환식 모델들(recurrent models)은 n개의 최적 목록(n-best list)으로 재평가(rescoring)함으로써 음성 인식의 정확도를 향상시켰다[9]. 전통적으로 이 요소들은 -음향, 발음, 언어 모델들은- 모두 각자 다른 목적을 갖고 개별적으로 학습되었다. 이 분야의 최근 연구는 발화 정보를 직접 문장으로 학습시킴으로써, 즉 엔드투엔드 방식으로 학습시킴으로써 이 개별적(disjoint) 학습 문제를 해결하려고 하였다[10, 11, 12, 13, 14, 15]. 이러한 방식의 두 가지 접근법은 연결주의적 시간 분류(Connectionist Temporal Classification, CTC)와 어텐션을 사용한 시퀀스 투 시퀀스이다. 두 접근법 모두 우리가 강조하고자 하는 한계를 가지고 있다: CTC는 라벨 출력이 서로 조건부 독립(conditionally independent)이라고 가정한다; 반면에 시퀀스 투 시퀀스 접근은 음소 시퀀스에만 적용되었으며[14, 15], 음성 인식을 위해 엔드투엔드로 학습되지 않았다.

이 논문에서 우리는 기존의 시도를 개선하는 모델인 Listen, Attend and Spell(LAS) 모델을 제안한다[12, 14, 15]. 신경망은 음향 시퀀스 신호를 단어 시퀀스로, 한 글자씩 변환하는 방법을 학습한다. 기존의 접근과는 다르게, LAS는 라벨 시퀀스에서의 독립성을 가정하지도 않고, HMM에 의존하지도 않는다. LAS는 어텐션을 활용한 시퀀스 투 시퀀스 학습법에 기반한다[17, 18, 16, 14, 15]. 그것은 리스너(listener)라고 불리는 인코더 RNN과, 스펠러(speller)라고 불리는 디코더 RNN으로 구성된다. 리스너는 저수준 음성 신호(low level speech signal)을 고수준 특성(high level feature)로 변환하는 피라미드형 RNN이다. 스펠러는 어텐션 기법을 사용하여 문자 시퀀스에 대한 확률 분포(probability distribution)를 지정함으로써 이러한 고수준 특성을

발성 출력(output utterance)로 변환하는 RNN이다[16, 14, 15]. 리스너와 스펠러는 함께 학습된다.

우리의 접근법에 대한 핵심은 어텐션 모델이 관련 정보(relevant information)을 추출해야 할 시간 단계(time step)의 수를 줄이는 피라미드형 RNN을 사용했다는 사실이다. 모델이 문자 시퀀스를 한 번에 하나씩만 출력하므로 빈도가 적거나 사전에 등록되지 않은(out-of vocabulary, OOV) 단어들은 자동으로 관리된다. 문자를 출력으로 모델링하는 방식의 또 다른 이점은 신경망이 다양한 철자 변형들을 자연스럽게 생성할 수 있게 된다는 점이다. 예를 들어, "triple a"라는 구절에 대하여 모델은 "triple a"와 "aaa"를 상위 빔(top beams)에 생성한다(**4.5장**을 보라). CTC와 같은 모델은 프레임(frames)간의 조건부 독립성 가정 때문에 이처럼 동일한 발화에 대하여 다양한 표현 방식을 생성하는 것이 곤란하다.

우리의 실험에서, 우리는 이러한 요소들이 LAS가 잘 작동하도록 하는 데 필요하다는 것을 확인했다. 어텐션 기법을 적용하지 않으면 3백만 개의 큰 학습 발화 데이터에 대한 것임에도 불구하고 모델은 학습 데이터에 현격하게 과적합했다 - 그것은 음향 특성에 집중(attention)하지 않고 학습 문장들을 기억했다. 인코더 쪽의 피라미드형 구조를 배제하면, 모델의 수렴은 크게 느려졌다 - 한 달 가량 학습하였음에도 불구하고, 오차율은 우리가 위에서 소개한 것들보다 훨씬 높았다. 이러한 두 가지 문제는 음향 신호가 수백, 수천 개의 프레임을 가질 수 있기 때문에 RNN을 학습시키는 것을 어렵게 한다는 데서 기인한다. 결국, 스펠러가 학습 문장들에 과적합하는 것을 방지하기 위해 우리는 학습 과정에서 한 샘플링 트릭(sampling trick)을 사용하였다[19].

이런 개선점들을 바탕으로, LAS는 구글 음성 탐색 과제(Google voice search task)에서 사전이나 언어 모델 없이 14.1%의 WER를 기록하였다. 언어 모델 재조정(language model rescoring)을 사용하면, LAS는 10.3%의 WER을 기록하였다. 비교하자면, 구글 최고 성능(state-of-the-art) CLDNN-HMM 시스템은 같은 데이터셋(dataset)에서 8.0%의 WER을 기록하였다.

2. 관련 연구(Related Work)

3. 모델(Model)

3.1 Listen

3.2 Attend and Spell

3.3 학습(Learning)

3.4 디코딩과 재조정(Decoding and Rescoring)

4. 실험(Experiments)

4.1 어텐션 시각화(Attention Visualization)

4.2 빔 너비의 효과(Effects of Beam Width)

4.3 발화 길이의 효과(Effects of Utterance Length)

4.4 단어 빈도(Word Frequency)

4.5 흥미로운 디코딩 예시(Interesting Decoding Examples)

5. 결론(Conclusion)

감사의 말

References

[1] Nathaniel Morgan and Herve Bourlard. Continuous Speech Recognition using Multilayer Perceptrons with Hidden Markov Models. In IEEE International Conference on Acoustics, Speechand Signal Processing, 1990.

[2] Abdel-rahman Mohamed, George E. Dahl, and Geoffrey E. Hinton. Deep belief networks forphone recognition. In Neural Information Processing Systems: Workshop on Deep Learningfor Speech Recognition and Related Applications, 2009.

[3] George E. Dahl, Dong Yu, Li Deng, and Alex Acero. Large vocabulary continuous speechrecognition with context-dependent dbn-hmms. In IEEE International Conference on Acoustics, Speech and Signal Processing, 2011.

[4] Abdel-rahman Mohamed, George E. Dahl, and Geoffrey Hinton. Acoustic modeling using deep belief networks. IEEE Transactions on Audio, Speech, and Language Processing,20(1):14–22, 2012.

[5] Navdeep Jaitly, Patrick Nguyen, Andrew W. Senior, and Vincent Vanhoucke. Applicationof Pretrained Deep Neural Networks to Large Vocabulary Speech Recognition. In INTERSPEECH, 2012.

[6] Tara Sainath, Abdel-rahman Mohamed, Brian Kingsbury, and Bhuvana Ramabhadran. DeepConvolutional Neural Networks for LVCSR. In IEEE International Conference on Acoustics,Speech and Signal Processing, 2013.

[7] Kanishka Rao, Fuchun Peng, Hasim Sak, and Francoise Beaufays. Grapheme-to-phonemeconversion using long short-term memory recurrent neural networks. In IEEE InternationalConference on Acoustics, Speech and Signal Processing, 2015.

[8] Kaisheng Yao and Geoffrey Zweig. Sequence-to-Sequence Neural Net Models for Graphemeto-Phoneme Conversion. 2015.

[9] Tomas Mikolov, Karafiat Martin, Burget Luka, Eernocky Jan, and Khudanpur Sanjeev. Recurrent neural network based language model. In INTERSPEECH, 2010.

[10] Alex Graves, Santiago Fernandez, Faustino Gomez, and Jurgen Schmiduber. ConnectionistTemporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In International Conference on Machine Learning, 2006.

[11] Alex Graves. Sequence Transduction with Recurrent Neural Networks. In International Conference on Machine Learning: Representation Learning Workshop, 2012.

[12] Alex Graves and Navdeep Jaitly. Towards End-to-End Speech Recognition with RecurrentNeural Networks. In International Conference on Machine Learning, 2014.

[13] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, RyanPrenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Ng. Deep Speech:Scaling up end-to-end speech recognition. In http://arxiv.org/abs/1412.5567, 2014.

[14] Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. End-to-end Continuous Speech Recognition using Attention-based Recurrent NN: First Results. In Neural Information Processing Systems: Workshop Deep Learning and Representation Learning Workshop, 2014.

[15] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio.Attention-Based Models for Speech Recognition. In http://arxiv.org/abs/1506.07503, 2015.

[16] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation byJointly Learning to Align and Translate. In International Conference on Learning Representations, 2015.

[17] Ilya Sutskever, Oriol Vinyals, and Quoc Le. Sequence to Sequence Learning with NeuralNetworks. In Neural Information Processing Systems, 2014.

[18] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares,Holger Schwen, and Yoshua Bengio. Learning Phrase Representations using RNN EncoderDecoder for Statistical Machine Translation. In Conference on Empirical Methods in NaturalLanguage Processing, 2014.

[19] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks. In http://arxiv.org/abs/1506.03099, 2015.

[20] Tara N. Sainath, Oriol Vinyals, Andrew Senior, and Hasim Sak. Convolutional, Long ShortTerm Memory, Fully Connected Deep Neural Networks. In IEEE International Conference onAcoustics, Speech and Signal Processing, 2015.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with DeepConvolutional Neural Networks. In Neural Information Processing Systems, 2012.

[22] Leonard E. Baum and Ted Petrie. Statistical Inference for Probabilistic Functions of FiniteState Markov Chains. The Annals of Mathematical Statistics, 37:1554–1563, 1966.

[23] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In International Conference onMachine Learning, 2001.

[24] Minh-Thang Luong, Ilya Sutskever, Quoc V. Le, Oriol Vinyals, and Wojciech Zaremba. Addressing the Rare Word Problem in Neural Machine Translation. In Association for Computational Linguistics, 2015.

[25] Sebastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. On Using VeryLarge Target Vocabulary for Neural Machine Translation. In Association for ComputationalLinguistics, 2015.

[26] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and Tell: A NeuralImage Caption Generator. In IEEE Conference on Computer Vision and Pattern Recognition,2015.

[27] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov,Richard Zemel, and Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generationwith Visual Attention. In International Conference on Machine Learning, 2015.

[28] Oriol Vinyals, Lukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey E. Hinton.Grammar as a foreign language. In http://arxiv.org/abs/1412.7449, 2014.

[29] Oriol Vinyals and Quoc V. Le. A Neural Conversational Model. In International Conferenceon Machine Learning: Deep Learning Workshop, 2015.

[30] Sepp Hochreiter and Jurgen Schmidhuber. Long Short-Term Memory. Neural Computation,9(8):1735–1780, November 1997.

[31] Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. Hybrid Speech Recognition withBidirectional LSTM. In Automatic Speech Recognition and Understanding Workshop, 2013.

[32] Salah Hihi and Yoshua Bengio. Hierarchical Recurrent Neural Networks for Long-Term Dependencies. In Neural Information Processing Systems, 1996.

[33] Jan Koutnik, Klaus Greff, Faustino Gomez, and Jurgen Schmidhuber. A Clockwork RNN. InInternational Conference on Machine Learning, 2014.

[34] Navdeep Jaitly, Vincent Vanhoucke, and Geoffrey Hinton. Autoregressive product of multiframe predictions can improve the accuracy of hybrid models. In INTERSPEECH, 2014.

[35] Hasim Sak, Andrew Senior, Kanishka Rao, and Francoise Beaufays. Fast and Accurate Recurrent Neural Network Acoustic Models for Speech Recognition. In INTERSPEECH, 2015.

[36] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In Neural Information ProcessingSystems, 2013.

[37] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, NagendraGoel, Mirko Hannenmann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, GeorgStemmer, and Karel Vesely. The Kaldi Speech Recognition Toolkit. In Automatic SpeechRecognition and Understanding Workshop, 2011.

[38] Jeffrey Dean, Greg S. Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Quoc V. Le, Mark Z.Mao, Marc'Aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, and Andrew Y. Ng. LargeScale Distributed Deep Networks. In Neural Information Processing Systems, 2012.