딥러닝 모델 adaptation 기술의 연구 동향

양준영, 장준혁
한양대학교

요 약

딥러닝 기술은 수많은 입력 데이터에 내재하고 있는 특징을 추출 및 합성함으로써 복잡한 특징공간을 모델링할 수 있는 강점을 가지지만, 테스트 환경에서 나타날 수 있는 특정 데이터 분포에 대하여 일반화가 잘 되지 않을 경우에는 해당 데이터를 이용하여 주어진 환경에 모델을 적응시킬 수 있는 기술을 필요로 한다. 이 글에서는 DNN 모델의 adaptation 기술 연구가 가장 활발하게 진행되고 있는 음향모델링에서의 다양한 adaptation 기술을 통해 연구 동향을 알아본다.

Ⅰ. 서 론

빅데이터 시대의 도래, 컴퓨팅 파워의 발전 및 깊은 신경망(deep neural network; DNN)의 학습을 가능하게 하는 학습 알고리즘의 개발과 함께 등장한 딥러닝은 음성인식[1], 영상인식[2], 자연어 처리[3], 문자열 예측[4] 등 다양한 분야에서 뛰어난 성능을 나타내고 있다. 딥러닝 기반 모델의 가장 큰 강점은 입력 데이터에 가해지는 연속한 비선형변환을 통한 특징벡터 추출 및 합성으로 기존의 얕은 모델로는 표현할 수 없는 복잡한 특징공간에 대한 강력한 모델링 커패시티일 것이다. 그러나, 테스트 환경에서 학습시에 관측되지 않은 데이터 분포가 입력될 경우 딥러닝 모델은 일반화 능력의 부족으로 인해 성능이 하락하는 경우가 발생할 수 있다. 딥러닝 모델의 adaptation은 이처럼 테스트 환경에서 관측되는 특정 데이터 분포에 대한 모델의 성능을 향상시키기 위해 일부 데이터를 이용하여 모델을 적응적으로 추가 학습시키는 방법이다. 이에 덧붙여 테스트 환경뿐만 아니라 딥러닝 모델의 학습 과정에서부터 서로 다른 환경에 기인한 학습데이터 사이에 차별성을 두고 학습 및 테스트를 진행함으로써 모델의 성능 향상을 도모하는 방법을 adaptive training이라 한다.

딥러닝 기술을 활용한 다양한 음성 어플리케이션 중 DNN 모델의 adaptation 기술 연구가 활발히 진행되고 있는 예로 음성인식 분야에서의 음향모델링을 꼽을 수 있다. 음향모델은 단어의 의미를 구분하는 기본 단위에 해당하는 각 음소들을 하나의 Hidden Markov Model (HMM)로 모델링하여 음성신호로부터 추출된 특징벡터의 나열을 음소의 나열로 바꾸어주는 역할을 하는데, 기존의 Gaussian Mixture Model(GMM)-HMM 기반 음향모델[5]은 각 hidden state에서의 관측확률분포를 GMM으로 모델링하여 관측데이터의 우도를 최대화하도록 학습하는 생성모델(generative model)인 반면, DNN-HMM 기반 음향모델은 하나의 통합된 분별모델(discriminative model)로써 관측데이터에 해당하는 hidden state의 사후확률을 최대화하도록 학습한다[1]. DNN 음향모델의 adaptation은 딥러닝 기술의 등장과 함께 음향모델 학습의 패러다임이 GMM에서 DNN으로 변화함에 따라 이에 적합한 adaptation 기술의 필요성으로 인해 자연스럽게 DNN adaptation 기술의 연구가 가장 활발한 분야로써 자리잡게 되었다.

음향모델의 adaptation은 모델을 어떤 환경에 적응시킬지에 따라 크게 두 종류로 나눌 수 있는데, 서로 다른 화자간의 발음이나 억양 등의 발성특성을 고려하여 특정 화자에 대해 모델을 적응시키는 것을 화자적응(speaker adaptation), 잔향이나 잡음, 채널 환경 등에 기인한 특정 음향환경에 대해 모델을 적응시키는 것을 환경적응(environment adaptation)이라고 할 수 있다. 음향모델의 adaptation 기술은 주로 화자적응 기술을 중심으로 개발되어왔고, 환경적응 기술은 화자적응과 동일한 방법을 사용하거나, 적응데이터의 특성상 다양한 음성·음향학적 전처리(pre-processing) 기법들과 결합되어 사용되는 경우가 많기 때문에 이 글에서는 DNN 음향모델의 화자적응 기술에 대해 다룸으로써 딥러닝 모델의 adaptation 기술 연구 동향에 대해 설명할 것이다.

Ⅱ. 본 론

이 글에서는 DNN 기반 음향모델의 화자적응 기술을 크게 두

종류로 나누어 보았다. 보조 GMM (auxiliary GMM)을 통해 추출한 특징벡터를 이용하는 방법과, DNN의 학습 과정 및 구조적인 특성을 이용한 방법이다.

1. 보조 GMM을 이용한 화자적응 기술

가. Vocal Tract Length Normalization (VTLN)

VTLN은 화자의 연령대와 성별에 따라 다른 성도(vocal tract)의 모양과 길이로 인해 다양한 음향학적 특성을 보이는 음성신호를 노말라이즈하여 서로 다른 화자의 음성신호로부터 추출한 특징벡터들이 가지는 다양한 화자 특성에 기인한 다양성을 줄이는 방법이다. VTLN은 음성신호로부터 특징벡터를 추출하기 전 음성신호의 주파수 도메인 표현에서의 frequency warping을 통해 이루어지며, 서로 다른 화자별 warping factor를 추정하는 방법으로써 기존의 GMM-HMM 기반 음향모델을 warped domain에서 추출한 특징벡터들의 우도를 최대화하도록 학습하는 과정을 필요로 한다. [6]을 통해 VTLN을 DNN-HMM 기반 음향모델에 적용한 예를 살펴볼 수 있다. 해당 연구에서는 GMM-HMM 기반 음향모델을 이용하여 추정한 화자별 warping factor를 DNN의 타겟으로 하여 학습함으로써 warping factor를 추정하는 DNN 모델을 구성하고, 해당 모델의 출력값인 warping factor의 사후확률을 DNN-HMM 기반 음향모델의 입력 특징벡터에 덧붙여 사용함으로써 화자적응을 수행하였다.

나. Feature-space Maximum Likelihood Linear Regression (fMLLR)

Maximum likelihood linear regression (MLLR)은 GMM-HMM 기반 음향모델에서 사용되는 대표적인 화자적응 방법으로 GMM을 구성하고 있는 Gaussian 성분들의 평균벡터와 공분산행렬에 아핀변환(affine transformation) 또는 선형변환(linear transformation)을 적용함으로써 특정 화자 데이터에 대한 우도를 높이는 방법이다[7]. 이 때, Gaussian 성분의 평균벡터와 공분산행렬에 동일한 선형변환행렬을 적용한다는 제약을 이용하면 모델 파라미터에 대한 변환식을 이와 동등한 의미를 갖는 특징벡터에 대한 변환식으로 표현할 수 있게 되며, 이를 constrained MLLR (CMLLR) 또는 feature-space MLLR (fMLLR)이라고 한다[7]. fMLLR은 MLLR과 달리 특징공간에서 변환을 적용하기 때문에 변환된 특징벡터를 DNN-HMM 기반 음향모델의 학습에 사용함으로써 화자적응적 학습을 수행할 수 있다. 다만, fMLLR 변환행렬 추정에 필요한 GMM의 학습 과정이 DNN과는 다른 목적함수를 최적화 기준으로 한다는 점과, DNN 모델이 이미 주어진 특징벡터들로부터 다양한 화자정보가 포함된 특징공간을 모델링하는 데에 강력하다는 점에서 fMLLR로 변환된 특징벡터를 DNN 모델의 학습에 사용하는 것은 기존 GMM 모델의 화자적응 기법으로 사용하였을 때보다 상대적으로 적은 성능 향상을 가져온다고 알려져있다[8].

다. I-Vector

I-vector[9]는 화자인식 분야에서 정형화되어 사용되고 있는 특징벡터로, factor analysis 모델을 이용하여 서로 다른 화자 또는 채널 환경 사이에 존재할 수 있는 다양한 변화적 성분을 저차원의 벡터로 모델링하는 데에 원리를 두고 있다. 따라서, i-vector를 DNN 음향모델의 학습에 필요한 부가적인 특징벡터로써 덧붙여 사용하여 DNN이 모델링할 수 있는 특징공간의 다양성을 확장시킴으로써 화자적응적 학습을 수행할 수 있다[10]. 부가적인 특징벡터로써 i-vector와 fMLLR 변환을 비교해 볼 때, i-vector는 factor analysis 모델로부터 기인하여 서로 다른 화자 및 채널 사이의 변화적 성분을 모델링하는 데에 특화되어있다는 점과, 테스트 환경에서 추가적인 디코딩 패스를 필요로 하지 않는다는 점을 장점으로 가지고 있다. [10]에서는 i-vector를 입력 특징벡터에 덧붙여 사용하는 방법 외에 추가적인 DNN을 학습함으로써 i-vector에 포함된 화자 및 채널 성분을 이용하여 입력 특징벡터에 포함된 화자 및 채널 성분을 노말라이즈하는 기법을 제안하였다.

2. DNN 모델의 특성을 이용한 화자적응 기술

가. 아핀변환층을 이용한 화자적응 방법

DNN은 여러 층으로 구성된 아핀변환과 비선형 활성함수로 구성되어있다. 이 방법은 DNN 모델에 adaptation을 위한 아핀변환층을 추가한 뒤, 테스트 환경에서 주어진 데이터를 이용하여 해당 아핀변환층의 파라미터를 학습함으로써 특정 화자에 대한 모델의 성능을 향상시키는 간단한 adaptation 방법이다. 아핀변환층을 DNN의 입력층, 은닉층, 또는 출력층에 추가하여 adaptation을 수행하는 모델을 각각 linear input network (LON), linear hidden network (LHN), linear output network (LON)이라고 한다[11].

나. Learning Hidden Unit Contributions (LHUC)

LHUC[12] 방법은 은닉층의 활성값을 조절함으로써 adaptation을 수행하는 방법으로 화자가 달라짐에 따라 각 은닉층의 뉴런들이 활성화되는 정도가 달라진다는 가정 하에 은닉층의 활성값 벡터에 대해 화자 의존적으로 작용할 수 있는 스케일링 벡터(scaling vector)를 학습하는 방법이다. 스케일링 벡터는 scaled sigmoid function을 통해 0과 2 사이의 값으로 범위

가 제한된 뒤에 은닉층의 활성값 벡터에 각 성분별로 곱해짐으로써 각 은닉층의 활성값 벡터를 특정 화자에 대해 적응시킨다. LHUC 방법은 매우 간단하면서도 효과적인 모델 파라미터공간 적응 기술로 위에서 언급한 fMLLR 또는 i-vector를 이용한 특징공간 적응 기술과도 쉽게 결합하여 사용할 수 있다[10].

다. Hierarchy 출력층을 이용한 adaptation

음성인식에서의 DNN 음향모델은 보통 3~4천개의 타겟 클래스를 가지는 분류모델이기 때문에 적은 양의 데이터를 이용하여 adaptation을 진행할 경우 대부분의 타겟 클래스에 속한 특징벡터들의 분포가 관측되지 않는 데이터 부족 문제가 발생하게 되어 adaptation의 효과를 크게 얻을 수 없다. Hierarchy 출력층[13]은 이와 같이 많은 타겟 클래스에 기인한 데이터 부족 문제를 해결하기 위한 방법으로, 타겟 클래스를 유사한 성질을 가지는 클래스들끼리 묶어 훨씬 적은 수의 클래스로 구성된 새로운 출력층을 쌓은 뒤, 해당 구조를 이용하여 adaptation을 진행하는 방법이다. 음성인식 분야에서는 DNN 모델의 타겟 클래스 집합을 구성할 때에 결정트리를 이용하는데, 이 때 결정트리에서 상위 노드들을 적절히 선택함으로써 수천개의 tied triphone state에 해당하는 타겟 클래스의 수를 줄일 수 있다. 이 방법은 특히 adaptation에 사용할 수 있는 데이터의 양이 제한적인 경우에 효과적이다. 〈그림 1〉은 [13]에서 제시된 hierarchy output layer를 이용한 adaptation 방법을 나타내고 있는 그림이다.

라. KL-Divergence Regularization

KL-divergence regularization term을 이용한 adaptation

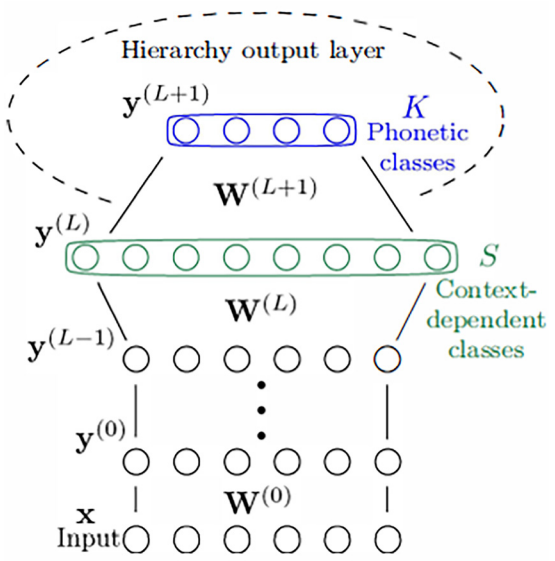

그림 1. Hierarchy output layer를 이용한 adaptation[13]

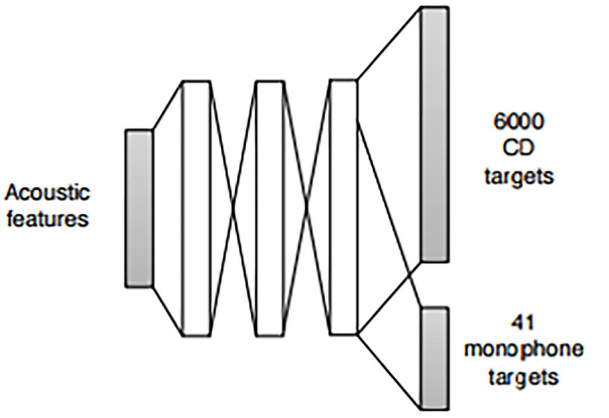

그림 2. Multi-task learning을 이용한 adaptation[16]

방법[14] 또한 adaptation에 사용할 수 있는 데이터의 양이 제한적인 경우에 추가적인 학습을 통한 adaptation이 어려운 문제를 해결하는 방법이다. 이 방법은 적응 후 모델의 출력확률분포가 적응 전 모델의 출력확률분포로부터 많이 벗어나지 않도록 규제하는 방법으로써 목적함수에 두 확률분포 사이의 KL-divergence term을 추가한 뒤 back-propagation을 진행하여 DNN 모델의 많은 파라미터들을 적은 양의 데이터를 이용하여 효과적으로 적응시킬 수 있다. 이와 같이 원래 모델의 결정특성으로부터 크게 벗어나지 않도록 하며 학습을 진행하는 방법을 conservative training이라고 한다[15].

마. Multi-Task Learning

Multi-task learning이란 유사한 성질을 가지는 서로 다른 두 task를 동시에 학습함으로써 각 task를 학습할 때에 얻을 수 있는 지식들을 결합적으로 이용하여 학습을 진행하는 방법이다. [16]과 [17]에서는 DNN 음향모델의 adaptation 방법으로써 triphone state에 대한 분류와 monophone state에 대한 분류 task를 동시에 학습하였는데, 이는 타겟 클래스의 수가 다르지만 동일한 작업을 수행하는 task를 동시에 학습함으로써 많은 타겟 클래스 수에 기인한 데이터 부족 문제를 multi-task learning을 통해 해결하는 방법을 제안하고 있다. 〈그림 2〉는 [16]에서 제시된 multi-task learning을 이용한 adaptation 방법을 나타내고 있는 그림이다.

바. 특이값분해 (Singular Value Decomposition; SVD)

특이값분해를 이용한 adaptation 방법은 DNN의 가중치행렬이 대부분 0에 가까운 값을 가지는 sparse한 행렬이라는 가정하에 원래의 가중치행렬을 특이값분해하여 계산한 특이값들 중 값이 작은 순서대로 특이값들을 버림으로써 서로 다른 두 low-

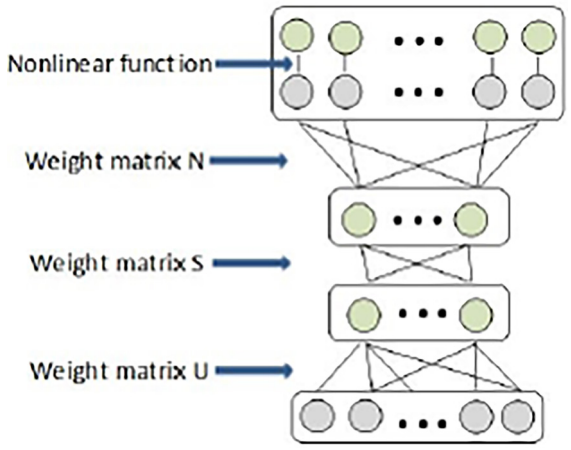

그림 3. 특이값분해를 이용한 adaptation[18]

rank 행렬(N, U)의 곱으로 근사시킨 뒤, 두 행렬 사이에 단위 행렬로 초기화한 정방행렬(S)을 삽입하여 back-propagation을 통해 해당 정방행렬을 적응시킨다[18]. 이 방법은 계산한 특이값들 중 얼마 정도를 버릴 것인지에 따라 학습해야 하는 DNN 모델 파라미터의 수를 대폭 줄일 수 있으며, full-rank 행렬을 적응시키는 방법과 비교할만한 성능을 보인다. 〈그림 3〉은 [18]에서 제시된 특이값분해를 이용한 adaptation 방법을 나타내고 있는 그림이다.

사. 매개변수화된 활성함수

매개변수화된 활성함수는 DNN 모델의 각 은닉층에서 적용되는 비선형 활성함수를 매개변수화 한 뒤 back-propagation 진행시에 함께 학습하여 활성함수의 표현력을 풍부하게 함으로써 DNN 모델의 모델링 커패시티를 확장시킬 수 있다. [19]에서는 DNN 기반의 음향모델링에서 활성함수를 매개변수화하고 주어진 데이터를 이용하여 매개변수를 적응시킴으로써 adaptation을 수행하였다.

Ⅲ. 결 론

이 글에서는 딥러닝 모델의 adaptation 기술 연구 동향을 음향모델링에서의 DNN 모델에 적용할 수 있는 다양한 화자적응 기술들을 통해 설명하였다. 기존에 사용되던 GMM-HMM 기반의 음향모델을 보조적인 모델로 사용한 방법들은 주로 GMM을 이용하여 DNN의 학습에 추가적인 정보를 제공할 수 있는 특징벡터를 추출하는 기술로 구성되어 있었으며, 이외에 분류 모델로써의 DNN의 학습 과정, 구조적인 특성 및 가중치행렬과 바이어스 파라미터를 이용한 adaptation 기술들을 알아보았다. DNN 기반의 음향모델링은 성능을 고도화하기 위하여 특징공간 적응과 모델 파라미터공간 적응 기술들을 함께 사용하는 추세이며, 이 글에서 다루지 않은 convolutional neural network (CNN)[10]나 recurrent neural network (RNN)[20] 기반의 음향모델 또는 end-to-end 프레임워크[21]에서의 adaptation 기술에 대한 연구도 활발하게 진행될 것을 기대해 볼 수 있을 것이다.

참 고 문 헌

[1] Dahl, George E., et al. "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition." IEEE Transactions on Audio, Speech, and Language Processing 20.1 (2012): 30-42.

[2] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

[3] Mikolov, T., and J. Dean. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems (2013).

[4] Graves, Alex. "Generating sequences with recurrent neural networks." arXiv preprint arXiv:1308.0850 (2013).

[5] Rabiner, Lawrence R. "A tutorial on hidden Markov models and selected applications in speech recognition." Proceedings of the IEEE 77.2 (1989): 257-286.

[6] Serizel, Romain, and Diego Giuliani. "Vocal tract length normalisation approaches to DNN-based children's and adults' speech recognition." Spoken Language Technology Workshop (SLT), 2014 IEEE. IEEE, 2014.

[7] Leggetter, Christopher J., and Philip C. Woodland. "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models." Computer Speech & Language 9.2 (1995): 171-185.

[8] Parthasarathi, Sree Hari Krishnan, et al. "fMLLR based feature-space speaker adaptation of DNN

acoustic models." Sixteenth Annual Conference of the International Speech Communication Association. 2015.

[9] Dehak, Najim, et al. "Front-end factor analysis for speaker verification." IEEE Transactions on Audio, Speech, and Language Processing 19.4 (2011): 788-798.

[10] Miao, Yajie, Hao Zhang, and Florian Metze. "Speaker adaptive training of deep neural network acoustic models using i-vectors." IEEE/ACM Transactions on Audio, Speech, and Language Processing 23.11 (2015): 1938-1949.

[11] Yao, Kaisheng, et al. "Adaptation of context-dependent deep neural networks for automatic speech recognition." Spoken Language Technology Workshop (SLT), 2012 IEEE. IEEE, 2012.

[12] Swietojanski, Pawel, and Steve Renals. "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models." Spoken Language Technology Workshop (SLT), 2014 IEEE. IEEE, 2014.

[13] Price, Ryan, Ken-ichi Iso, and Koichi Shinoda. "Speaker adaptation of deep neural networks using a hierarchy of output layers." Spoken Language Technology Workshop (SLT), 2014 IEEE. IEEE, 2014.

[14] Yu, Dong, et al. "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition." 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013.

[15] Albesano, Dario, et al. "Adaptation of artificial neural networks avoiding catastrophic forgetting." The 2006 IEEE International Joint Conference on Neural Network Proceedings. IEEE, 2006.

[16] Bell, Peter, and Steve Renals. "Regularization of context-dependent deep neural networks with context-independent multi-task training." 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015.

[17] Huang, Zhen, et al. "Rapid adaptation for deep neural networks through multi-task learning." Proc. Interspeech. 2015.

[18] Xue, Jian, et al. "Singular value decomposition based low-footprint speaker adaptation and personalization for deep neural network." 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014.

[19] Zhang, C., and P. C. Woodland. "DNN speaker adaptation using parameterised sigmoid and ReLU hidden activation functions." 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016.

[20] Miao, Yajie, and Florian Metze. "On speaker adaptation of long short-term memory recurrent neural networks." Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)(To Appear). ISCA. 2015.

[21] Graves, Alex, and Navdeep Jaitly. "Towards End-To-End Speech Recognition with Recurrent Neural Networks." ICML. Vol. 14. 2014.

약 력

양 준 영

2016년 한양대학교 공학사
2016년~현재 한양대학교 음성/음향/오디오신호처리 연구실 석박통합과정 재학중
관심분야: 음성인식

장 준 혁

1992년~1998년 경북대학교 공학사
1998년~2000년 서울대학교 공학석사
2000년~2004년 서울대학교 공학박사
2000년~2005년 Netdus Corporation, Chief Engineer
2004년~2005년 University of California, Santa Barbara, Unisted States, Postdoctoral Fellow
2005년~2005년 KIST, Research scientist
2005년~2011년 인하대학교 전자공학과 조교수
2011년~현재 한양대학교 융합전자공학부 부교수
관심분야: 딥 러닝, 음성인식 및 통신 전처리기술, 오디오신호처리, 바이오신호처리