신경 베이지안: 고전적 베이지안 너머의 딥러닝 기반 사전 지식 부여 패러다임

이상우[1]°, 김지원[2], 김정희[2], 장병탁[1]
서울대학교 컴퓨터공학부[1], 네이버 랩스[2]
slee@bi.snu.ac.kr, {g1.kim, jeonghee.kim}@navercorp.com, btzhang@bi.snu.ac.kr

Neural Bayesian: The Paradigm of Prior Knowledge beyond Classical Bayesian Technique in Deep Learning

Sang-Woo Lee[1]° , Jiwon Kim[2], Jeonghee Kim[2], Byoung-Tak Zhang[1]
School of Computer Science & Engineering, Seoul National University

요 약

베이지안 기법은 기계학습 분야에서 오랫동안 각광받은 개념이지만, 신경망을 포함한 대다수의 분류기 모델의 성공에 있어서 제한적인 영향력을 보여주었다. 본 논문에서는 최근 3년 간의 딥러닝의 여러 연구 결과를 다른 관점에서 설명하기 위하여, 신경망에서 사용될 수 있는 다른 형태의 사전 지식 부여에 대한 관점, 신경 베이지안 (neural Bayesian)을 제안한다. 이 관점에 따르면 신경망에는 파라미터 확률에 대한 가정이 아니라, 파라미터간 가중치 공유와 가중치 학습 조기 중단에 의해 적절한 정규화와 전이 학습이 수행된다. 본 논문에서는 위와 같은 원리에 의해 신경망이 동작하는 여러 예를 나열한다. 특별히 우리는 이러한 원리를 더 잘 이해하기 위하여, 사전 확률 부여 기법의 일종인 신경 사전 확률(neural prior)을, 깊은 컨볼루션 신경망을 바탕으로 끊임 없이 들어오는 대용량 이미지 데이터에 대하여 순차적으로 온라인 학습하는 문제에 적용하였다. 우리가 제안한 방법은 기존의 고전적인 점진적 앙상블 기법과 베이지안 기법을 압도하며, 실험 결과의 추이는 적절한 파라미터가 존재하는 공간인 뉴럴 메니폴드 (neural manifold)의 존재를 암시한다.

1. 서 론

본 논문은 신경망 연구에서 베이지안 철학의 대안이 될 수 있는 관점인 신경 베이지안 (neural Bayesian) 을 제안한다. 베이지안 (Bayesian) 기법은 모델의 파라미터에 대한 사전 확률 (prior)를 가정하여, 적은 데이터만 가지고도 더 정확하고, 덜 상식에서 벗어나는 파라미터를 찾는, 기계학습 기법 중의 일종이다. 베이지안 이론은 기계학습 전반의 근간을 이루었으며, 또한 토픽 모델링 (topic model)을 포함한 많은 생성 모델 (generative model)에 대하여 좋은 성능을 보여주었다. 하지만, 기계 학습 연구자들의 베이지안에 대한 오랜 선호와는 달리, 신경망 (neural networks), 의사 결정 나무 (decision tree), 지지 벡터 머신 (support vector machine, SVM)과 같은 실용적인 분류 알고리즘의 성능을 높이는 데에 기여하지 못했다.

본 연구에서는 신경망, 특별히 딥러닝 분야에서 최근 3년간 소개된 관련 기념비적인 연구 결과를 정리하고, 각 기법들이 어떻게 신경 베이지안의 철학에서 묶이며, 기존 베이지안보다 나은 결과를 얻게 되는 지 소개한다. 우리는 신경망 학습에서 베이지안의 대안이 되는 슬로건인 신경 베이지안을 제안하고, 그 기작에 대해 논의한다.

2. 베이지안

베이지안 추론이란 파라미터 θ 의 사전확률 (prior) $P(\theta)$ 에 대한 정보를 사용하여, 데이터 D 에 대한 파라미터의 학습을 수행하는 방법을 총칭한다. 본 논문에서는 사전확률을 사용하여 사후 확률 (posterior) $P(\theta|D)$ 을 최대화하는 기법들을 주로 제안하는 모델과 비교한다.

베이지안 이론이 가장 기본적으로 사용되는 곳은 정규화이다. 정규화는 regularization의 번역으로, 선험 지식을 통해 모델이 제한된 수의 데이터에 따라 이상한 파라미터를 가지는 것을 막는 기법을 의미한다. 모델의 가중치가 커지면 과적합(overfitting)한 가능성이 높다는 정규화 이론에 따라, 많은 기계학습 방법이 목표 함수에 $P(\theta) \propto \exp(\|\theta\|_2^2)$, $P(\theta) \propto \exp(\|\theta\|_1)$ 과 같은 l2-norm, l1-norm을 적용하며, 이는 신경망 학습에서도 마찬가지이다.

이러한 방법은 전이학습으로 확장될 수 있다. 전이학습이란 기존에 학습한 데이터 및 문제를 다루는 데이터셋 (source dataset)을 학습하여 기존과 다른 스타일의 데이터나, 문제를 다루는 데이터셋 (target dataset)의 학습 성능을 높이는 문제를 말한다. [1]에서는 SVM 기반 전이 학습에서 source dataset의 최적 해인 $\hat{\theta}$ 을 사용 사전확률 $P(\theta) \propto \exp(\|\theta - \hat{\theta}\|_2^2)$ 을

적용하여 이미지 분류 성능을 높였다. 우리는 이후의 실험 보고에서 위 알고리즘을 norm prior라고 지칭한다. 위 연구를 포함하여 2000년도 중후반부터 최근까지도, 베이지안과 여기에 적용될 수 있는 생성 모델(generative model)을 바탕으로 한 전이학습 연구가 계속하여 이루어졌다. 하지만 그 전이학습 성능은 대체로 뛰어나지 않았다 [2].

3. 신경 베이지안

베이지안이 파라미터의 확률을 통하여 모델의 좋은 공간을 제시하는 원리이듯이, 신경 베이지안은 신경망에서 사용될 수 있는 패러다임이며, 구체적으로는 가중치 공유와 가중치 조기 중단을 통하여 더 정확하고 덜 상식에서 벗어나는 파라미터를 찾는 것을 원리로 가진다. 우리는 다섯 가지의 선행 연구 및 알고리즘을 통하여 이러한 기작에 대한 논증을 시작한다.

첫 번째 예는 가중치 학습 조기 중단 (early stopping)이다. 신경망에서는 특별히 조기 중단이 l2-norm을 적용하는 것보다 더 좋은 정규화 방법이며, 이는 특히 순환 신경망에서 더 두드러지게 드러난다. 이는 신경망에서 베이지안 외의 방법이 정규화에 더 적합함을 암시하며, 이 역시 신경 베이지안의 틀 안에 들어간다.

하지만, 더 훌륭한 두 번째 예는 최근에 제안된 dropout이다. Dropout은 신경망 학습을 할 때, 매 데이터를 바탕으로 가중치를 교정할 때마다, 은닉 변수의 일부를 끈 상태로 기울기 값을 계산하는 학습 방법을 의미한다. 이러한 방법에 대한 해석 중 하나는 dropout을 학습한 모델이 매 step마다 꺼지고 켜진 파라미터로 만들어진 각각의 신경망의 앙상블과 비슷한 역할을 한다는 것이다. 이 해석에서, 각 기하급수의 수의 신경망들의 가중치는 서로 공유되고 있다. 이러한 가중치를 공유하는 신경망 앙상블은 실험적으로 아주 강력한 정규화 효과를 보인다. 또한 dropout은 가중치 학습 조기 중단 없이도 과적합을 많이 막는 효과도 보여주었다.

세 번째 예와 네 번째 예는 전이 학습이다. 세 번째 예로 [3]에서는 알파벳을 사용하는 11개 언어의 음성 인식에서 각 언어에 대한 딥러닝 모델의 구조를 공유함으로서 3~5%가량의 상대 성능 향상을 만들었다. 한편, [4]에서는 120만여개 1000개 class 이미지를 포함하는 ImageNet를 학습한 딥러닝 모델의 가중치를 공유하여 다양한 다른 이미지 인식 문제의 성능을 갱신하고, 이것이 SIFT처럼 이미지 인식에 사용될 수 있는 일반적인 이미지 feature임을 논증하였다.

4. 신경 사전 확률

신경 베이지안의 다섯 번째 예시이자, 우리가 실험에 사용할 방법은 [5]에서 제안된 것으로, 우리는 이것을 신경 사전 확률 (neural prior)라는 별칭으로 부른다. 이 방법은 기본적으로 [4]의 확장이며, 전이 학습 방법의 일종으로 target dataset의 신경망을 학습할 때 source dataset의 신경망의 가중치로 초기화한다. 이 방법을 통해 [5]은 500-class 60만 이미지를 학습한 딥러닝 모델을 바탕으로 다른 500-class 60만 이미지를 학습하는 딥러닝 모델에 전이, 4.47%의 상대 성능 개선을 얻었다. 이는 가중치 공유와 조기 중단이 적극적으로 사용되는 신경 베이지안의 더 확장된 예이다.

우리는 이러한 방법이 단순한 가중치 유사도에 대한 가정, 혹은 베이지안의 틀을 벗어난다는 것을 보이기 위하여 이를 온라인 학습 (online learning)에 적용하였다. 본 논문에서 온라인 학습이란 데이터가 끊임 없이 들어오는 상황에서 잠재적으로 기존 데이터를 대부분 버리고 일부의 계산량 만으로 효과적으로 새로운 데이터를 학습하는 방법을 의미한다. 우리가 특별히 neural prior ensemble이라는 알고리즘을 제안하며, 그 방법은 알고리즘 1에 있다. 우리는 새로운 데이터가 들어오면 이를 일정만큼 쌓아두었다가 새로운 neural network를 만들고 이들을 bagging한다. 이 때 새로운 신경망은 이전 신경망의 정보를 전이받게 된다. 이후 실험에서 neural prior ensemble은 알고리즘 1에서 얻어진 모든 신경망들을 ensemble하는 알고리즘을, neural prior는 여기서 마지막 신경망만 학습하는 알고리즘을, naïve incremental ensemble은 [3]의 가중치 초기화를 통한 전이 학습 없이 단순히 모델을 만들고 ensemble하는 알고리즘을 의미한다. naïve incremental ensemble은 learn++과 같은 유명한 방법에서 오랫동안 사용되어 온 온라인 학습 기법이다.

Algorithm Neural Prior Ensemble

repeat
- Collect N_{subset} new data D_{new}.
- Initialize a new neural network W_{new} by parameters of W_{prev}.
- Train W_{new} with D_{new}.
- Combine a weak learner W_{new} to a model. θ (i.e. $\theta \leftarrow \theta \bigcup \{W_{new}\}$)
- Refer to W_{new} as W_{prev}. (i.e. $W_{prev} \leftarrow W_{new}$)

until forever

알고리즘 1. 신경 사전 확률 앙상블

5. 실험 결과

우리는 우리가 제안한 프레임워크의 우수성을 검증하기 위하여 딥러닝 모델인 컨볼루션 신경망(convolutional neural network, CNN)을 바탕으로 이미지 인식 문제를 다루었다. 주장을 검증하기 위하여 MNIST, CIFAR-10, ImageNet 이미지 인식 데이터에 실험을 하고 그 결과를 보고한다. MNIST는 10개 class 학습 데이터 5만개, CIFAR-10는 10개 class 학습

데이터 6만개, ImageNet은 1000개 학습 데이터 50만개를 학습했으며, 각각 8개, 10개, 10개로 데이터를 균등하게 쪼개고 학습을 수행하였다.

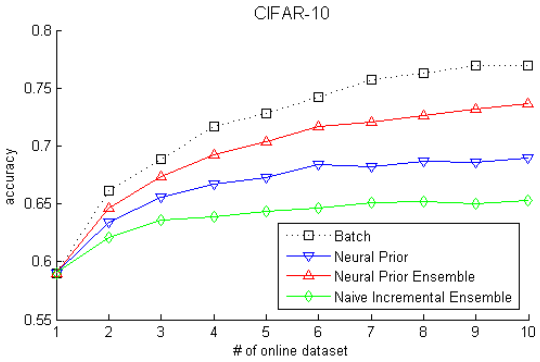

그림 1. CIFAR-10에 대한 온라인 학습 실험 결과

표 1. batch leaner와 비교한 상대 개선율 (%). 100%가 batch 성능이다.

	MNIST	CIFAR-10	ImageNet
neural prior	77.9	55.7	46.7
Neural prior ensemble	**90.3**	**81.7**	**71.6**
naïve incremental ensemble	45.6	35.1	33.1

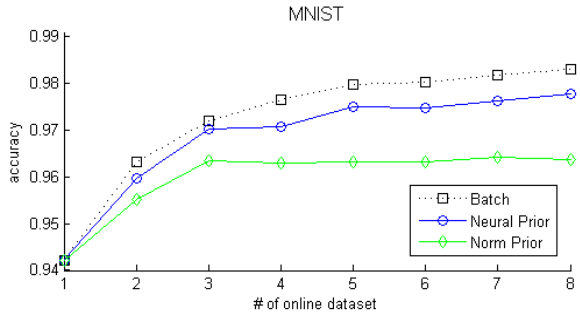

그림 2. MNIST에 대한 온라인 학습 실험 결과

그림 1은 CIFAR-10 우리가 제안한 알고리즘을 비교 모델들과 비교한 것이다. 먼저, neural prior 알고리즘은 기존의 naïve incremental ensemble 방법보다 성능이 좋다. 이는 전체 중의 일부의 데이터만으로는 CNN의 표상을 충분히 학습시키는 데에 무리가 있기 때문인 것으로 보인다. 또한 neural prior ensemble은 neural prior보다 성능이 좋다. 표 1은 MNIST와 CIFAR-10, ImageNet의 성능 결과를 보여준다. 학습 데이터가 어려워질수록 성능이 좋지 않아지는 것을 볼 수 있다.

Neural prior는 단순히 베이지안 철학에서처럼 기존에 있는 가중치와의 거리에 패널티를 주는 방법이 아니다. 그림 2는 multilayer perceptron을 바탕으로 neural prior와 norm prior[1]을 비교한 결과이다. norm prior 방법은 이전에 있는 weight을 l2 prior로 주는 [4]의 방법과 비교될 수 있다. 하지만 우리의 방법이 더 잘한다. 이것은 가중치 공유 및 조기 중단 기법이 베이지안 기법이 할 수 있는 것 이상의 동작을 함을 의미한다.

6. 논의 및 결론

본 논문에서 자세히 설명하지는 않았지만, 본 저자들은 이러한 일련의 현상들이 신경 메니폴드(neural manifold)를 함의한다는 가설을 가지고 있다. 이에 따르면, 신경망 가중치의 에러 탐색 공간에는 신경망이 데이터를 적당히 설명하는 특별한 subspace가 존재한다. 우리는 딥러닝에서 신경망 가중치가 어떠한 에러 탐색 공간을 가지고 있는 지 잘 알지 못한다. 예컨대, 최근 연구 결과에 따르면 CNN의 local minima는 global minima와 아주 가깝고 또한 global minima는 아주 큰 plateau를 가지게 될 것이라고 한다. 하지만 그러한 결과를 상상하는 것은 어려운 일이다. 우리는 후속 연구로 신경 메니폴드의 아이디어와 기존 기계학습 연구의 관계를 논의하고, 그 기작을 규명할 것이다.

많은 기계학습 모델의 기작이 베이지안 이론에 의하여 설명되었다. 그러나 최근 신경망은 안타깝게도 그 기작을 설명할 수 없는 이론적인 기반이 부족하며, 단지 합리적으로 고안된 휴리스틱들에 의하여 좋은 성능을 보이고 있을 따름이다. 본 논문이 제안한 신경 베이지안의 설득력과는 관계 없이, 딥러닝을 포함한 신경망 연구는 새로운 통합된 이론을 필요로 한다.

Acknowledgement

이 논문은 네이버 랩스의 재원으로 수행되었으며, 정부(미래창조과학부) 의 지원을 (NRF-2010-0017734-Videome, R0126-15-1072-SW스타랩, 10035348-mLife, 10044009-HRI.MESSI) 을 일부 받았음.

참 고 문 헌

[1] T. Tommasi, F. Orabona, B. Caputo, “Safety in Numbers: Learning Categories from few examples with Multi Model Knowledge Transfer, In *CVPR*,” 2010.

[2] R. Salakhutdinov, J. Tenenbaum, A. Torralba, “Learning with Hierarchical-Deep Models,” *IEEE T. PAMI*, 2013.

[3] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, J. Dean, “Multilingual Acoustic Models Using Distributed Deep Neural Networks,” In *ICCASP* 2013.

[4] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition,” In *ICML*, 2014.

[5] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, “How transferable are features in deep neural networks?,” In *NIPS*, 2014.