메타학습 문제를 위한 베이지안 신경망 모델 파라미터의 Amortized Inference

김주형○ 허재필
성균관대학교 소프트웨어대학
wngud0811@naver.com, jaepilheo@skku.edu

Amortized Inference of Bayesian Neural Network parameters for solving Meta-learning problems

Kim Juhyeong○ Heo Jaepil
College of Computing, Sungkyunkwan University

요 약

메타학습 연구에서는 데이터의 수가 제한된 다수의 태스크들이 주어졌을 때, 다양한 태스크에 일반화가 가능하고 유용한 정보를 추출하고자 한다. 더 나아가, 베이지안 메타학습 연구는 각각의 태스크가 주어졌을 때 그 태스크에 맞는 모델 사후분포를 빠르고 정확하게 탐색하는 것이 목표이다. 본 논문에서는 베이지안 신경망을 활용한 메타학습에서, 각각의 태스크가 주어졌을 때, 베이지안 신경망 모델 파라미터의 태스크 사후분포 추론에 잠재변수 모델을 활용하는 Amortized Inference 기법을 적용하였다. 또한 제안한 방법론을 대표적인 메타러닝 벤치마크 실험환경인 few-shot classification 문제에 적용해 기존 방법들과 성능을 비교하였다.

1. 서 론

메타학습(Meta-learning) 문제는 머신러닝 모델이 다양한 태스크 사이에서 일반적으로 활용될 수 있는 정보를 추출하는 능력을 시험하는 문제이다. 그 중에서도 Few-shot classification 문제는 모델이 기존에 관측하지 못했던 새로운 분류 Task에 대해서, 극도로 적은 수의 데이터만 가지고 얼마나 빠르고 정확하게 정의된 분류문제를 해결할 수 있는지 검증하고자 한다.

베이지안 신경망은 일반적인 빈도주의자(Frequentist) 신경망의 파라미터에 사전분포를 가정하고, 데이터를 활용해 사후분포를 추론하여 신경망 구조의 모델이 가진 불확실성을 정량화하는 머신러닝 모델이다. 신경망 모델이 과모수화(Over-parameterized)되고 구조가 복잡해짐에 따라, 근사 추론(Approximate Inference) 기법들의 근사 정확도와 연산 효율성의 한계로 비주류가 된 모델이다. 그러나 최근 베이지안 신경망 근사 추론의 문제점을 극복하기 위한 실용적인 방법들이 계속 연구되고 있다.

베이지안 신경망을 근사적으로 추론하기 위한 방법으로, 과거에는 MCMC 및 그 변형 알고리즘을 적용한, 반복적인 샘플링 방식의 근사추론 방법이 주류를 차지하였다. 이 방법은 확률적인 모델에 대한 근사 추론 정확도는 매우 우수하였지만, 연산복잡도가 파라미터의 차원과 데이터의 수에 모두 크게 영향을 받는다는 한계가 있다.

베이지안 신경망을 근사적으로 추론하는 또 다른 방법으로는, 신경망 모델에만 적용 가능한 dropout 규제 방법을 이용하여, 확률적으로 다양한 구조를 가지는 신경망 모델을 통해 신경망 모델의 분포를 표현하고, 추론과 예측에 활용하는 방식이 있다. 이 방법은 실용적이면서도 이론적으로도 잘 분석된 방법으로써, 베이지안 신경망을 구현하고자 할 때 일반적으로 사용해볼 수 있을 것이다.

하지만 현재 가장 활발히 연구되고 있는 근사 추론 방법은 변분 추론(Variational Inference) 기반 방식이다. 이 방법은 복잡한 베이지안 신경망의 사후분포를 더 다루기 쉬운 간단한 분포로 근사하고자 하고, 근사하는 분포의 파라미터를 최적화함으로써 근사 오차를 최소화하는 기법이다. 대표적인 변분 추론 기법으로써 근사 분포를 독립 정규분포로 가정하는 Mean-field Variational Inference에서는, 신경망 모델의 파라미터들이 가지는 복잡한 관계를 모델링하지 못한다는 단점이 있어 베이지안 신경망 모형에서 변분 추론의 실용성을 제한하였다. 또한 해당 기법은 과모수화된 신경망 모델에서 낭비되는 파라미터들까지 샘플링 기반 방식으로 추론한다는 점에서 연산 효율적이지 못하다고 여겨졌으나, 최근 다양한 개선 방법들이 제안되어 그 한계를 극복하고 있다.

Flipout[1]은 변분 추론 기반 방식의 베이지안 신경망 모델에 직접적으로 적용이 가능한 최적화기법의 일종이다. 기존에 변분 추론에서 각 데이터당 1번씩 수행되었던 베이지안 신경망의 파라미터 샘플링 횟수를 데이터 배치 당 1번으로 줄이고도, 파라미터의 편미분값이 가지는 높은 분산을 비약적으로 감소시키는 방법을 제안하였다. 뿐만 아니라, 순전파와 역전파에 소요되는 연산량이 같은 모델 구조를 가진 빈도주의자 신경망에 비해 최대 2배 정도 밖에 걸리지 않는 실용성을 보이는데 성공하였다. 이 방법은 베이지안 신경망을 근사 추론할 때, 은닉 특징이 아닌 파라미터를 대상으로 수행하는 변분 추론의 경우에 한해 현존하는 최고의 구현 방법론 중 하나이다.

Amortized Variational Inference는 변분 추론 기법들 중에서도 별도의 모델로 근사 분포의 파라미터 또는 잠재 변수를 예측하여 추론하는 기법들을 지칭하는 용어이다.[2] 이 기법은 Variational Auto-Encoder 모델에서 사용되어 널리 알려졌다. 일반적으로 잠재 변수 추론 시에 한 개의 조합의 잠재 변수를 모든 데이터에 대해 광역적으로 적용하던 것과 달리, Amortized Variational Inference는 각 데이터를 별도의 잠재 변수 모델에 입력하여 데이터별 잠재 변수를 추론하고 활용하고자 한다.

최근 머신러닝 학계에서 Amortized Variational Inference 기법에 대한 관심이 커짐에 따라, Amortized Variational Inference로 베이지안 모델이 가진 파라미터의 사후분포를 추론하는 시도가 다양하게 이루어지고 있

다. 하지만 본 연구자가 가진 지식의 범위 내에서는, 기존의 연구는 대부분 간단한 형태를 가진 모델의 파라미터를 추론하는 데에 그쳤다. 예를 들어 딥러닝 모델의 마지막 층이 가진 파라미터만 예측하거나 선형 베이지안 모델의 파라미터를 추론하는 등의 방식이었다. 하지만 본 연구에서는 기존 연구들에서 더 나아가 더 복잡한 파라미터들 간의 관계를 가지는 베이지안 신경망 모델의 파라미터를 추론하고자 시도하였다.

2. 문제 정의

파라미터 θ로 구성된 모델 f_θ가 있다고 할 때, 본 연구에서 다룬 메타학습 문제를 정의하면 다음과 같다.

$$\min_{\theta} \mathrm{E}_{T\sim p(T)}\left[\mathrm{E}_{D\sim p_T(D)}[L(D;f_\theta)]\right]$$

위 식에서 $p(T)$는 태스크의 분포를 의미하고, 이는 전체 클래스에서 분류 문제를 구성할 일부 클래스만 표본 추출하는 방식으로 경험적으로 표현할 수 있다. 또한 $p_T(D)$는 임의의 태스크 분포에서 태스크를 표본 추출하고, 해당 태스크로 정의되는 데이터셋의 분포이다. 그러므로 본 문제는 태스크 분포에서 태스크를 표본 추출하고, 추출된 태스크별 데이터셋의 분포에서 다시 데이터를 표본 추출하여 모델 f_θ가 주어졌을 때의 손실함수 L에 대해 최소화하는 문제이다. 다중작업학습(Multi-task Learning)과 같은 일부 메타학습 문맥에서는 손실함수 L을 각 태스크마다 상이하게 가정하기도 하지만, 본 연구에서는 단순히 모든 태스크에 대해 L의 형태가 같다고 가정하였다.

또한 잠재 변수를 z, 데이터 D가 주어졌을 때 변분 분포를 $q_\varphi(z|D)$라 한다면 Amortized Variational Inference의 문제 정의는 다음과 같다.

$$\min_{\varphi} \mathrm{E}_{D\sim p(D)}\left[\mathrm{E}_{z\sim q_\varphi(z|D)}\left[\ln\frac{q_\varphi(z|D)}{p(z,D)}\right]\right]$$

이는 데이터 분포에서 얻을 수 있는 각 데이터마다, 변분 분포를 조정해 실제 데이터와 잠재변수에 대한분포에 대한 근사 오차를 최소화하는 문제이다. 본 연구에서는 메타학습 문제를 Amortized Variational Inference의 문제를 품으로써 해결하고자 하였고, 두 문제를 결합하였다. 이는 다음과 같이 표현할 수 있다.

$$\min_{\varphi}\mathrm{E}_{T\sim p(T)}\left[\mathrm{E}_{D\sim p_T(D)}\left[KL\left(q_\varphi(\theta|D)||p(\theta|D)\right)\right]\right]$$
$$\equiv \min_{\varphi}\mathrm{E}_{T\sim p(T)}\left[\mathrm{E}_{D\sim p_T(D)}\left[\mathrm{E}_{\theta\sim q_\varphi(\theta|D)}\left[\ln\frac{q_\varphi(\theta|D)}{p(\theta|D)}\right]\right]\right]$$

이는 일반적인 메타학습 문제의 변분적 하한(Variational Lower Bound)임을 어렵지 않게 보일 수 있을 것이다. 변분 파라미터 φ, 변분 분포 $q_\varphi(\theta|T)$이 구체적으로 무엇을 의미하는지는 3절 방법론 부분에서 서술하였다.

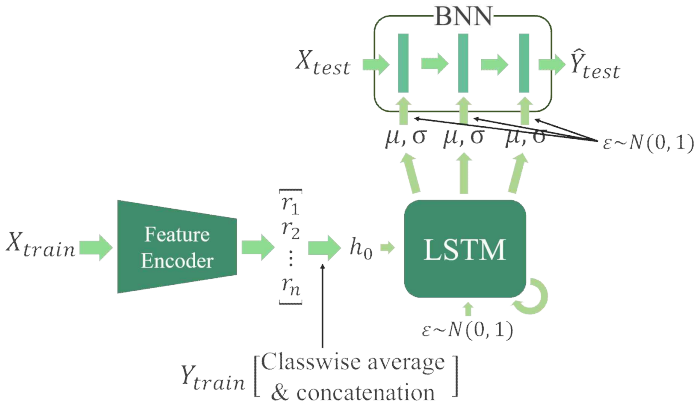

그림 1 본 연구에서 제안하는 모델의 도식.

3. 모델

본 연구에서는 베이지안 신경망의 파라미터들이 층별로 순차적인 구조를 가지므로, 변분 분포 $q_\varphi(\theta|T)$를 LSTM 모델로 표현하고자 하였다. 따라서 φ는 LSTM의 모델이 가진 파라미터이다. 먼저 베이지안 신경망 모델의 구조를 정하고, 베이지안 신경망의 각 파라미터의 평균과 분산 파라미터를 LSTM 모델로 한 sequence step당 베이지안 신경망 한 층의 파라미터를 한 번에 추출하는 방식으로 베이지안 신경망 모델 파라미터의 근사 분포를 얻었다. 독립 정규분포를 근사 분포로 가정하는 Mean-field Variational Inference와 다른 부분은, LSTM 모델을 통해 각 정규분포의 평균과 분산 파라미터 간의 복잡한 상호작용을 모델링할 수 있다는 것이다.

LSTM 모델의 초기 은닉 특징 h_0로는 각 태스크 데이터를 별도의 특징 추출 신경망을 통해 추출하여 사용하였다. 각 sequence step에서 input은 정규분포 노이즈로써, 이는 베이지안 신경망 파라미터의 평균과 분산 파라미터에 영향을 주는 hyperprior라고 할 수 있다. 따라서 본 방법론은 일종의 Hierarchical Bayesian approach로 여겨질 수 있고, 여러 근원으로 인한 모델의 불확실성이 존재한다. 이는 Hierarchical Bayesian의 장점인 강건함(Robustness)을 유도하기 위해 이런 방식으로 모델을 구성한 것이다. 일반적으로, 베이지안 신경망 모형은 파라미터의 사전분포 선택에 따라 일반화성능이 극도로 민감하게 반응한다고 알려져 있다. 사전분포 선택으로 인한 모델의 민감성을 줄이기 위해 LSTM 모델을 이용해, 사전분포의 파라미터에 대한 사전분포인 hyperprior를 모델링하여 모델 초기화에 대한 강건성을 높이고자 하였다.

각각의 학습과정마다, LSTM 모델의 출력으로부터 얻어진 베이지안 신경망 파라미터 분포의 파라미터를 바탕으로 베이지안 신경망 모델의 표본을 표본 추출하였고, 경사를 계산한 후 내부최적화 과정(Inner-gradient step)을 수행하고자 하였다. 그리고 그 파라미터를 다시 추론 과정에서 사용하고, 역전파를 통해 LSTM 모델의 파라미터에 대한 편미분 추정치를 얻어 경사기반 최적화 알고리즘을 적용하였다. 또한 이때 얻어진 평균과 분산 파라미터는 표준정규분포와의 쿨백-라이블러 발산 값을 계산하여 규제 효과를 위해 역전파과정에서 반영하였다.

4. 구현 상세

본 연구에서는 은닉 특징 수준이 아닌 파라미터 수준의 베이지안 신경망 근사추론 기법을 적용하였고, 매 학습 단계마다 베이지안 신경망이 가진 파라미터의 수에 비례하는 표본 추출 횟수가 요구되었다. 표본 추출 과정의 효율을 최대화하기 위해, Flipout 기법[1]을 적용하였다. 제안된 방식대로, Fully connected 구조를 가진 신경망 층 파라미터의 평균과 분산 파라미터에 노이즈를 곱한 후 순전파를 진행하는 방식으로 구현하였다.

LSTM 모델의 초기 은닉 특징을 태스크 데이터로부터 추출하기 위해, 별도의 특징 추출 모델을 사용하였다. 본 연구에서는 모든 실험 과정에서 EfficientNet을 사용하여 태스크 데이터로부터 특징을 추출하였고, 같은 클래스별로 그 특징 값을 평균 내었다. 그 후, LSTM 모델이 클래스 배치 순서를 고려하지 않도록 무작위적인 순서로 병합하여 LSTM 모델의 초기 은닉 특징 h_0로 사용하였다.

본 연구에서 정의하는 최적화 문제는 변분 파라미터 φ에 대한 최소화문제이지만, 본 연구의 문제 정의와 유사한 기법들인 Black-box 방식과 Optimization-based 방식의 메타학습 기법들은 태스크별 파라미터에 대한 내부 최적화과정을 추가로 수행한다. 본 연구의 본래 목표는 내부 최적화과정을 완전히 없애도록 베이지안 신경망 모델을 End-to-End로 추론하는 것이었지만, 다수의 실험결과 본 방법론에서 그것은 경험적으로 불가능하였다. 기본적인 수준의 분류 문제를 푸는 베이지안 신경망 모델의 파라미터를 추론하는 경우에도 훈련 데이터셋에서의 분류 정확도조차 1/<클래스 수>를 기록하며, 정상적인 학습이 불가능하였다. 따라서 본 연구에서도 태스크별 파라미터라고 할 수 있는 θ에 대한 내부최적화과정을 수행하게 되었으나, 반복 횟수를 최소화하고자 하였다.

평균과 분산 파라미터와 표준정규분포와의 쿨백-라이블러 발산 값을 손실함수에 반영할 때의 scale은 hyperparmeter로써, 실험과정을 통해 직접 탐색하였다.

5. 결과

본 절에서는 본 연구에서 제안한 방법을 Omniglot 데이터셋과 miniImagenet 데이터셋에서 실험하고 성능을 측정하여, 대표적인 Few-shot classification 연구들과 그 성능을 비교하였다. 각각의 실험환경에서 내부최적화를 1회 수행하며 학습한 경우에는 learning rate 1의 SGD를 수행하였고, 5회 학습한 경우에는 learning rate 0.5의 SGD를 수행하였다.

표 1 Omniglot 데이터셋에서의 결과 비교

	5-way		20-way	
	1-shot	5-shot	1-shot	5-shot
Matching Network[4]	98.1%	98.9%	93.8%	98.7%
Prototypical Network[5]	98.8%	99.7%	**96.0%**	**98.9%**
MAML[6]	98.7%	**99.9%**	95.8%	**98.9%**
Meta-SGD[7]	**99.5%**	**99.9%**	95.9%	**98.9%**
본 연구(내부최적화 1회)	68.5%	72.6%	29.3%	34.0%
본 연구(내부최적화 5회)	87.4%	86.1%	54.3%	54.7%

표 2 miniImagenet 데이터셋에서의 결과 비교

	5-way	
	1-shot	5-shot
Matching Network[4]	46.6%	60.0%
Prototypical Network[5]	49.4%	**68.2%**
MAML[6]	48.7%	63.1%
Meta-SGD[7]	**50.4%**	64.0%
본 연구(내부최적화 1회)	38.9%	48.9%
본 연구(내부최적화 5회)	35.8%	52.1%

실험 결과 본 방법에서 제안한 방법의 벤치마크 성능은 현존하는 최고의 Few-shot classification 방법들의 성능에 비하면 상당히 부진한 성능을 보였다. 하지만 1-shot에 비교해 5-shot의 경우에서 성능이 증가하는 것으로 보아, 본 연구에서 제안하는 방법론이 태스크 데이터에서 유의미한 context를 추출하고 주어진 문제를 푸는데 활용한다는 것을 경험적으로 확인할 수 있었다.

6. 결론 및 향후 연구

본 연구에서 제안한 방법을 통해 정량적으로는 현존하는 최고의 메타학습 기법들을 능가하는 성능을 보이는 것이 불가능하였지만, 메타학습 문제를 해결하는 데에 베이지안 모델을 사용하고, 모델의 파라미터의 좋은 초기 분포를 Amortized Inference 기법을 이용한 추론이 가능하다는 것을 실험을 통해 검증하였다.

향후 연구 방향으로는 하이퍼파라미터에 대한 더 정밀한 조정, LSTM이 아닌 다른 형태의 모델 탐색 등을 시도해볼 수 있을 것이다. 또한 본 연구에서는 다루지 못했던, 베이지안 모델의 파라미터를 Amortized Inference로 추론할 때 불확실성 정량화 능력에 대한 분석도 수행해볼 수 있을 것이다.

7. 참고문헌

[1] Yeming Wen, Paul Vicol, Jimmy Ba, Dustin Tran and Roger Grosse, “Flipout: Efficient Pseudo-Independent Weight Perturbations on Mini-Batches”, International Conference on Learning Representations, 2018.

[2] Cheng Zhang, Judith Butepage, Hedvig Kjellstrom and Stephan Mandt, “Advances in Variational Inference”, in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 8, 2019.

[3] Garnelo et al, “Neural Processes”, International Conference on Machine Learning, 2018.

[4] Vinyals et al, “Matching networks for one shot learning”, Advances in Neural Information Processing Systems, 2016.

[5] Jake Snell, Kevin Swersky and Richard S. Zemel, “Prototypical Networks for Few-shot Learning”, Advances in Neural Information Processing Systems, 2017.

[6] Finn et al, “Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks”, International Conference on Machine Learning, 2017.

[7] Zhenguo Li, Fengwei Zhou, Fei Chen and Hang Li, “Meta-SGD: Learning to Learn Quickly for Few-Shot Learning”, arXiv:1707.09835v2.