

딥러닝 기반의 이종 선생 네트워크를 지원하는 주요 파라미터 최적화 흐름정보 전이학습 기술

김귀훈*, 배지훈**

Important Parameter Optimized Flow-based Transfer Learning Technique Supporting Heterogeneous Teacher Network Based on Deep Learning

Kwihoon Kim*, Ji-Hoon Bae**

요 약

본 논문에서는 이종 선생 네트워크를 지원하면서 학습속도가 더 빠른 흐름정보 기반의 전이학습 알고리즘을 제안하고자 한다. 먼저, 본 논문에서 제안하는 기법은 기존 흐름정보 전이학습에 비해서 비슷한 정확도 성능을 가지면서 낮은 복잡도를 가짐을 실험적으로 확인하였다. 다음으로, 중요한 파라미터를 최적화하는 방법을 베이시안 최적화(BO, Bayesian Optimum) 방법을 통해서 기존 기술 대비 0.1%~0.3% 높은 정확도를 얻었고, 약 230,000초~250,000초 사이의 빠른 학습 결과를 보여주었다. 따라서 제안한 이종 선생 네트워크를 지원하는 주요 파라미터 최적화 흐름정보 전이학습 기술은 기존 기술 대비 그 성능이 유사하거나 높은 정확도를 얻으면서 빠른 학습을 수행하는 딥러닝 기반의 전이학습 기술로 활용할 수 있을 것으로 기대된다.

Abstract

In this paper, we propose a flow-based transfer learning (TL) algorithm with a faster learning speed while supporting the hetero teacher network. First, the proposed technique has a similar accuracy and performance compared to the previous TL using flow information and has been identified as having lower complexity. Next, the proposed scheme through Bayesian optimization obtained 0.1% to 0.3% more accuracy than the existing technology and showed learning results that were 230,000 to 250,000 seconds faster. Therefore, it is expected that the proposed scheme can be used as deep Learning-based transfer learning technology, which achieves similar or higher accuracy compared to the existing technology.

Keywords

deep learning, machine learning, transfer learning, flow of solution procedure, Bayesian optimization

* 한국전자통신연구원 책임연구원
- ORCID: <http://orcid.org/0000-0002-0798-1687>

** 대구가톨릭대학교 인공지능·빅데이터 공학과 교수(교신저자)
- ORCID: <http://orcid.org/0000-0002-0035-5261>

• Received: Dec. 02, 2019, Revised: Jan. 22, 2020, Accepted: Jan. 25, 2020

• Corresponding Author: Ji-Hoon Bae

Dept. of AI·Big Data Engineering, Daegu Catholic University, Rep. of Korea,

Tel.: +82-53-850-2883, Email: jihbae@cu.ac.kr

1. 서 론

지난 몇 년 동안 딥러닝 기술이 크게 발달함에 따라 다양한 분야에서 딥러닝을 적용하기 시작하였다. 그러나 딥러닝의 필수 요소인 데이터 부족 문제 등이 있어서 이를 극복하기 위하여 최근 딥러닝 모델을 이용한 선생-학생 프레임워크 기반의 전이학습(TL, Transfer Learning) 기술들이 등장하였다[1]-[3]. 선생-학생 프레임워크 기반의 전이학습 기술은 사전 학습 완료된 딥 뉴럴네트워크(DNN, Deep Neural Network)를 선생 모델로 정의하고 상기 선생 모델에서 유용한 지식을 추출하여, 다른 딥 뉴럴네트워크인 학생 모델로 추출된 지식을 이전시키는 기술이다. 특히, 기존 전이학습 기술 중 [3]은 입력 공간에서 출력 공간으로 많은 층(Layer)들을 통해 순차적으로 맵핑할 때, 층 사이의 흐름(Flow)을 디스틸드 지식(Distilled knowledge)으로 정의한다. 이것은 서로 다른 두 개 층으로부터 특징(Features) 사이의 내적을 계산하여 특징이 변화해 가는 흐름 정보를 구할 수 있다. 따라서, 사전 학습 완료된 선생 모델에서 특징이 변화해가는 흐름 정보를 추출하고 이를 학생 모델로 이전하여 학습을 수행하는 것을 FSP(Flow of Solution Procedure) 전이학습이라고 한다[3].

기존의 FSP 전이학습의 경우, 학생 네트워크와 선생 네트워크 모두 구조가 동일한 RESNET (RESidual NETwork)[4] 딥러닝 모델을 이용하여 전이학습을 수행하였다. 또한, FSP 기반 흐름 정보를 이전할 경우, 학습에 필요한 다양한 파라미터들이 존재하는데, 일반적인 그리드 검색(GS, Grid Search) 방법을 이용하여 전이학습을 수행하였다. 이러한 기존 방식은 네트워크 구조가 복잡한 딥러닝 모델에 대하여 학습 시간이 오래 걸릴 수 있고, 원하는 최적의 결과를 도출하기가 어려울 수 있다.

따라서 이를 극복하기 위하여, 본 논문에서는 학생 네트워크와 서로 다른 선생 네트워크를 사용하는 이중 선생-학생 프레임워크 경우에도 가능한 FSP 기반 흐름 정보를 추출하고 빠른 학습이 가능한 알고리즘을 제안하고자 한다. 여기서, 전이학습에 필요한 중요한 파라미터에 대해서 기존 GS 방법 대신에 베이지안 최적화(BO, Bayesian Optimum)

[5] 방법을 통해서 가장 최단 시간에 최적에 가까운 결과로 수렴 가능한 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문에 대한 관련 연구를 설명하고, 3장에서는 이중 선생 네트워크를 지원하는 주요 파라미터 최적화 FSP 전이학습 알고리즘을 제안한다. 4장에서는 실제 구현을 통한 실험 결과를 통해 성능 개선 효과를 보여준다. 5장에서는 결론 및 향후 연구 계획에 관해서 설명한다.

II. 관련 연구

2.1 Dark knowledge 기반 전이학습

본 절에서는 먼저 선생-학생 프레임워크 기반의 기존 전이학습 방법으로 dark knowledge 기법에 대하여 간략히 기술하고자 한다. [1]에서는 사전에 학습 완료된 선생모델 내부에 축적된 지식을 dark knowledge로 정의하고, 다른 학생모델로 상기 지식을 효율적으로 전달하기 위한 모델 압축용 기법을 제안하였다. 본 기법에서는 선생 모델의 마지막 출력층에서의 소프트맥스(Softmax) 클래스(Class) 확률값을 완화상수(Softening factor)로 완화하여 출력된 확률값을 학생 모델로 이전하는 전이학습을 수행한다. 이러한, 완화된 클래스 확률분포는 선생 모델이 학습한 지식을 좀 더 일반화하여 전달할 수 있는 장점이 있기 때문에, 완화상수 적용 없이 클래스 확률값을 그대로 학생 모델로 전달하는 경우보다 더 높은 분류 정확도를 보여주었다.

2.2 힌트 정보 기반 전이학습

힌트(Hint) 정보[2]는 사전 학습 완료된 선생 모델의 은닉 층(Hidden layer)에서의 출력을 의미한다. [2]에서는 선생 모델의 힌트 정보와 2.1절에서 기술한 선생 모델의 dark knowledge를 단계적으로 학생 모델로 이전하는 선생-학생 프레임워크에서의 전이학습 방법을 제안하였다. [2]의 방법에서 선생 모델로부터 힌트 정보를 추출하는 은닉 층인 힌트 층(Hint layer)은 선생 모델 전체 높이에서 중간에 해당되는 은닉 층을 선택한다. 그리고 학생 모델 또한

전체 높이에서 중간에 해당되는 은닉 층을 가이드 층(Guide layer)으로 정의하고, 학생 모델의 가이드 층에서의 출력값과 선생 모델의 힌트 층에서의 출력값과의 유클리디안 거리(Euclidean distance)가 최소가 되도록 힌트 정보를 학생 모델로 이전한다. 상기의 힌트 정보 학습 완료 후, 2.1 절에서 기술한 것과 같이 완화 상수를 적용한 dark knowledge를 학생 모델로 최종 이전하는 단계별 전이학습을 수행한다. [2]에서 제안한 방법은 기존 2.1절의 dark knowledge 뿐만 아니라 추가적인 힌트 정보의 두 가지 선생 지식들을 추출하여 학생 모델로 이전하기 때문에 dark knowledge만 이용하는 전이학습 방법 보다 더 높은 정확도 성능을 보여주었다.

2.3 흐름 정보 기반 전이학습

최근, 선생 모델 내부의 계층별 특징들이 변화해 가는 흐름 정보 기반의 새로운 전이학습 방법이 제안되었다[3]. [3]에서 흐름 정보는 서로 다른 두 은닉층의 출력인 특징맵(Feature map)들 사이의 그람 행렬(Gramian matrix) 연산으로 정의하고, 이의 출력 결과를 FSP 행렬로 명명하였다. FSP 행렬은 하나의 특징맵에서 또 다른 특징맵으로 출력되는 과정을 수학적으로 도출한 것으로, 낮은 층에서부터 높은 층으로 특징이 변화해 가는 흐름정보를 FSP 행렬로 표현하였다. 따라서 [3]에서는 사전 학습 완료된 선생 모델의 낮은 층에서의 세부적 특징(Low-level feature)으로부터 높은 층에서의 추상적 특징(High-level feature)으로 변화해 가는 흐름 정보를 학생 모델로 이전하여 학습하는 방법을 수행하였다. 이때, 선생 및 학생 모델 모두 균등하게 3 등분하고 3개의 FSP 행렬들을 생성한 후, 학생 모델에서 추출된 FSP 행렬값과 선생 모델에서 추출된 FSP 행렬값 사이의 유클리디안 거리가 최소가 되도록 복수 개의 흐름 정보들을 동시에 학습하도록 수행한다. [3]의 선생-학생 프레임워크에서 채택한 선생 및 학생 모델 모두 최신의 성능이 우수한 RESNET 딥러닝 모델을 이용하였다. [3]의 결과에 따르면 CIFAR-10 및 CIFAR-100[6], CUB(Caltech-UCSD Birds)[7] 등의 신뢰성 있는 여러 가지 벤치마크 데이터 셋에 대하여, 동일한 딥러닝 모델 대비 2.2절의 전이학습 방

법보다 흐름 정보를 학습한 학생 모델의 정확도 성능이 더 우수함을 관찰할 수 있다.

2.4 BO 기술

BO[5]는 모델 학습의 관점에서 모델 최적화를 위한 최적의 하이퍼 파라미터(Hyper-parameter)를 탐색하기 위한 방법 중의 하나로써, 현재까지 추출한 입력값과 이에 대한 함수값들을 바탕으로 다음의 새로운 목적 함수에 대한 확률적인 추정을 수행하는 Surrogate 모델과 목적 함수에 대하여 현재까지의 확률적 추정 결과를 바탕으로 다음의 최적 입력값 후보를 추천해 주는 Acquisition 함수로 크게 구성되어 있다.

따라서 BO 방식은 이러한 두 가지 구성요소들을 이용하여, 기존의 매 회차마다 시도할 후보 하이퍼 파라미터 값들을 주관적으로 선정하고, 이를 사용하여 모델 학습을 수행한 후 검증 데이터셋에 대하여 측정된 성능 결과를 기록하는 과정을 매 번 반복하는 GS 방법보다 더 효율적으로 모델에 필요한 최적의 하이퍼 파라미터 값들을 자동으로 추출할 수 있다. 즉, 매 회차마다 새로운 하이퍼 파라미터 값을 추정할 경우 사전지식을 충분히 반영함과 동시에 전체적인 탐색 과정을 좀 더 체계적으로 수행하기 때문에 모델 학습에 필요한 하이퍼 파라미터들에 대한 최적값들을 자동으로 추출 가능하다. 따라서 본 연구에서는 이러한 BO 기술을 적용하여 전이학습에 필요한 여러 가지 하이퍼 파라미터들에 대한 최적값들을 자동으로 추출하고, 그 성능 결과를 기존의 그리드 탐색방법과도 비교한다.

III. 제안하는 이중 선생 네트워크를 지원하는 주요 파라미터 최적화 흐름정보 전이학습 알고리즘

3.1 전체 흐름도

본 논문에서 제안하는 이중 선생 네트워크를 지원하는 FSP 기반 전이학습 알고리즘 전체 흐름도는 그림 1과 같다.

크게 3 단계로 나뉘는데 첫 번째 단계는 FSP 기

반 전이학습이고, 두 번째 단계는 학생모델에 대한 미세조정학습(Student finetuning) 이다. 마지막으로 세 번째 단계는 하이퍼 파라미터 최적화(Hyper parameters optimization) 이다.

먼저, FSP 기반 전이학습 단계에서는 기존에 학습된 선생 모델을 로딩한다. 이때, 선생 모델은 학생 모델보다 층수가 깊은 VGGNET 혹은 RESNET 과 같은 큰 모델로써 이미 학습된 것을 사용한다. 고정된 선생 모델(Fixed teacher model)을 기준으로 학생 모델로 흐름 정보를 이전하기 위하여 그림 1 과 같이 3 개의 T_1, T_2, T_3 FSP 행렬들을 생성한다. 그리고, 학생 모델은 RESNET 기반 DNN을 사용하고 선생 모델과 동일한 개수의 S_1, S_2, S_3 FSP 행렬들을 생성한다. 만약, 선생과 학생 모델 사이의 3개 쌍의 FSP 행렬들에 대한 각 손실이 $Loss_1^{FSP}, Loss_2^{FSP}, Loss_3^{FSP}$ 과 같을 경우 전체 손실함수는 다음의 식 (1)과 같이 주어진다.

$$\begin{aligned}
 Loss^{FSP} &= Loss_1^{FSP} + Loss_2^{FSP} + Loss_3^{FSP} \\
 &= \lambda MSE(T_1, S_1) + \lambda MSE(T_2, S_2) + \lambda MSE(T_3, S_3)
 \end{aligned}
 \tag{1}$$

여기서 λ 는 손실함수에 대한 가중치를 의미하고,

$MSE(T_i, S_i)$ 는 i 번째 선생 모델의 FSP 행렬과 i 번째 학생 모델의 FSP 행렬 간의 차에 대한 L2 norm 을 의미한다.

따라서 입력 이미지(Input image)를 양쪽 선생 네트워크와 학생 네트워크에 입력으로 넣고 식 (1)의 손실함수를 이용하여, 선생 모델의 다중 FSP 행렬인 T_1, T_2, T_3 에 맞게 선생 모델과 다른 이종 모델의 다중 FSP 행렬인 S_1, S_2, S_3 가 유사해지는 방향으로 흐름정보 기반 지식전이를 수행한다.

다음으로, 학생모델에 대한 미세조정학습 단계에서는 기존의 FSP 기반 전이학습 단계에서 만들어진 초기 학생 모델을 초기값으로 셋팅을 한다. 입력 이미지 데이터를 기반으로 이종 학생 모델을 실제 라벨 데이터를 이용한 일반적인 지도학습 방법으로 최적화된 학생 모델(Finetuned student model)을 생성한다.

마지막으로 하이퍼 파라미터 최적화 단계에서는 최고의 성능이 나올 수 있도록 전이 학습의 각 하이퍼 파라미터를 최적화 한다. 흐름정보 전이학습에 필요한 주요 파라미터 최적화 흐름정보 알고리즘은 BO 알고리즘을 적용하였는데, 자세한 내용은 전통적인 GS 기법과 비교하여 3.3절에서 구체적으로 설명을 할 예정이다.

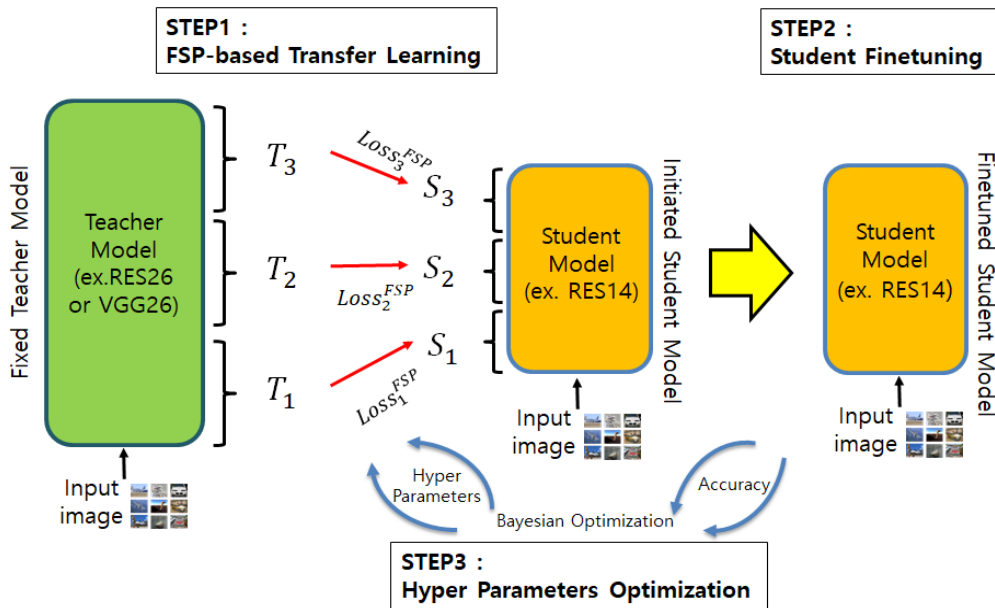


그림 1. 이종 선생 네트워크를 지원하는 FSP 기반 전이학습 최적화 전체 흐름도
 Fig. 1. Total flow of flow-based transfer learning algorithm supporting heterogeneous teacher network

3.2 이중 선생 네트워크 지원 흐름 정보 알고리즘

본 절에서는 동일 모델 FSP 기반 전이학습과 이중 모델 FSP 기반 전이학습을 구성하는 방법에 대하여 기술한다. 이때, 3.1 절에서 기술한 전이학습 알고리즘 전체 흐름도에서 단계 2의 미세조정학습 단계는 동일하고, 단계 1의 FSP 기반 전이학습은 동일 모델 및 이중 모델에 따라 그 구성이 다르다.

먼저, 동일 모델 FSP 기반 전이학습은 3.1절에서 설명한 것처럼 선생 네트워크와 학생 네트워크 각각에 대해서 12개 RES 블록과 6개 RES 블록을 낮은 층에서부터 높은 층까지 균등하게 3 등분으로 나눈다.

따라서 그림 2에 도시된 것과 같이 각 선생 모델의 FSP 행렬($T_1^{ho}, T_2^{ho}, T_3^{ho}$)과 동종의 학생 모델의 FSP 행렬($S_1^{ho}, S_2^{ho}, S_3^{ho}$)을 추출한 다음, 선생 및 학생 모델 간의 각 FSP 행렬쌍에 대한 $Loss_1^{FSP,ho}$ 및 $Loss_2^{FSP,ho}, Loss_3^{FSP,ho}$ 손실을 이용한 식 (2)의 전체 손실함수를 이용하여 동일 모델 간의 흐름정보 기반 지식전이를 수행한다.

$$Loss^{FSP,ho} = Loss_1^{FSP,ho} + Loss_2^{FSP,ho} + Loss_3^{FSP,ho} \quad (2)$$

$$= \lambda^{FSP,ho}MSE(T_1^{ho},S_1^{ho}) + \lambda^{FSP,ho}MSE(T_2^{ho},S_2^{ho}) + \lambda^{FSP,ho}MSE(T_3^{ho},S_3^{ho})$$

여기서 $\lambda^{FSP,ho}$ 는 동일 모델인 경우의 손실함수에 대한 가중치를 의미한다.

한편, 이중 모델 FSP 기반 전이학습 또한 선생 네트워크와 학생 네트워크 각각에 대해서 12개 RES 블록과 24개 VGG 블록을 낮은 층에서부터 높은 층까지 균등하게 3 등분으로 나눈다.

$$Loss^{FSP,he} = Loss_1^{FSP,he} + Loss_2^{FSP,he} + Loss_3^{FSP,he} \quad (3)$$

$$= \lambda^{FSP,he}MSE(T_1^{he},S_1^{he}) + \lambda^{FSP,he}MSE(T_2^{he},S_2^{he}) + \lambda^{FSP,he}MSE(T_3^{he},S_3^{he})$$

여기서 $\lambda^{FSP,he}$ 는 이중 모델인 경우에 손실함수에 대한 가중치를 의미한다.

이중 모델 FSP 기반 전이학습은 선생 네트워크와 학생 네트워크가 서로 다르지만 동일한 비율로 나눌 수 있는 경우에, 다중 흐름정보 추출 시 동종 모델 기반의 전이학습과 유사하게 적용할 수 있음을 알 수 있다. 따라서 VGG 네트워크가 아닌 다른 이중 모델일 경우에도 마찬가지로 다중 흐름정보 추출 시 동종 모델과 유사한 비율로 나눌 수 있다.

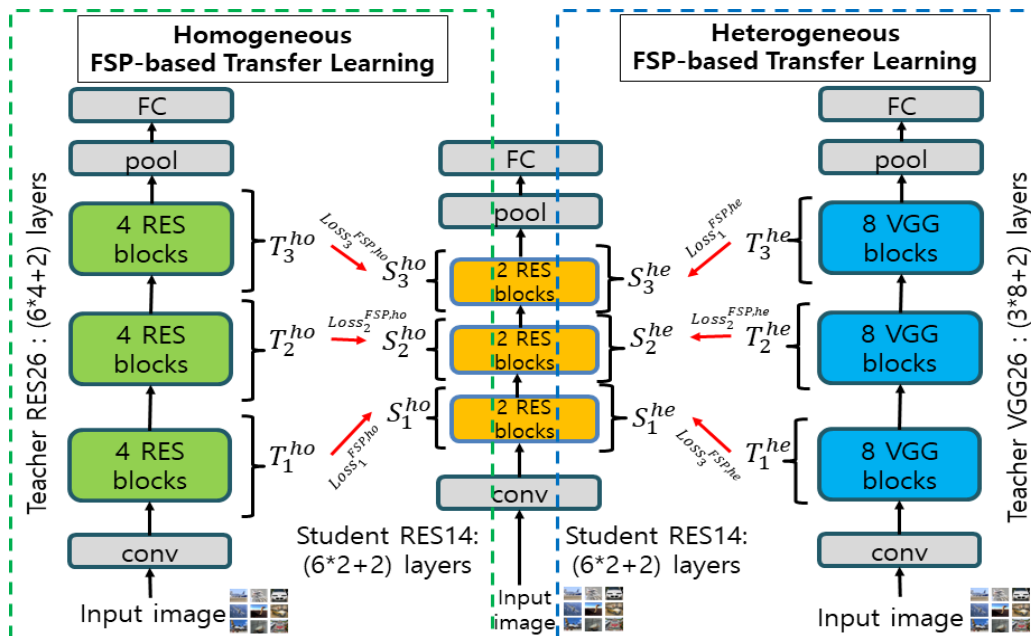


그림 2. 동일 모델 및 이중 모델 FSP 기반 전이 학습 알고리즘 흐름도
Fig. 2. Flow of homogeneous and heterogeneous model flow-based transfer learning algorithm

3.3 주요 파라미터 최적화 흐름 정보 알고리즘

본 절은 흐름정보 전이학습에 필요한 주요 파라미터 최적화 흐름정보 알고리즘에 대해서 설명한다. 본 논문에서는 크게 GS 기법과 BO 기법의 2가지 방법을 고려한다.

먼저, 알고리즘 1은 흐름정보 전이학습에 대한 기존의 GS 기반 주요 파라미터 최적화를 설명한다. 본 논문에서 고려하는 흐름정보 전이학습은 단계 1의 FSP 행렬을 이용한 지식전이 수행 시 주로 손실 함수에 대한 가중치와 초기 학습율에 영향을 받는다.

알고리즘 1. GS 기반 흐름정보 전이학습 주요 파라미터 최적화

```

1) GS 기반 Pseudo 코드

For (i = 0, j = 0 ; i<max_i and j<max_j; i++, j++)
    param_i = param_i_init + i*param_i_step
    param_j = param_j_init + j*param_j_step
    Load teacher model
    Do FSP-based knowledge transfer with param_i and param_j
    Save student model
    Do fine tuning
    result = Evaluate student model
    Accuracy[i][j] = result

i_sel, j_sel = maxarg (i,j)(accuracy[i][j])
Choose parameter param_i_sel = param_i_init + i_sel*param_i_step
Choose parameter param_j_sel = param_j_init + j_sel*param_j_step
    
```

알고리즘 2. BO 기반 흐름정보 전이 학습 주요 파라미터 최적화

```

2) BO 기반 Pseudo 코드

Define blackbox_fn(param_i, param_j)
    Load teacher model
    Do FSP-based knowledge transfer with param_i and param_j
    Save student model
    Do fine tuning
    Accuracy = Evaluate student model
    Return Accuracy

while (1)
    Do BayesianOptimum with blackbox_fn(param_i, param_j), max_trials
    Receive Accuracy
    if Accuracy > max_accuracy    max_accuracy = Accuracy
    else    max_accuracy_count ++
    end
    if max_accuracy_count > limit_max_accuracy_count or trials > max_trials
        Choose parameter param_i_sel
        Choose parameter praram_j_sel
        Stop BayesianOptimum
    end
    
```

따라서 여기서는 동종 모델에 대한 식 (2)의 가중치 $\lambda^{FSP,ho}$ 와 초기 학습율에 대한 2개 파라미터를 param_i와 param_j로 설정하여 최적의 파라미터를 구한다. 만약, 구하기를 원하는 전이학습 파라미터가 더 많아지면 알고리즘 1에 추가할 수 있다.

이때, 두 파라미터에 대한 최대 값인 max_i와 max_j 값이 커지면 원하는 최적 파라미터를 얻을 확률이 높아지는 반면에 복잡도가 커진다. 따라서, 알고리즘 1에서는 높은 연산량을 요구하는 GS 방식으로 정확도(Accuracy)가 최대가 되는 두 파라미터 param_i_sel과 param_j_sel을 구하게 된다.

다음으로, 알고리즘 2는 흐름정보 전이학습에 대한 BO 기반 주요 파라미터 최적화를 설명한다. 여기서도 알고리즘 1과 같이 이중 모델에 대한 식 (3)의 가중치 $\lambda^{FSP,he}$ 와 초기 학습율에 대한 2개 파라미터를 param_i와 param_j로 설정하여 최적의 파라미터를 구한다. 만약, 구하기를 원하는 전이학습 파라미터가 더 많아지면 알고리즘 2에 추가할 수 있다.

이때, 알고리즘 1과 달리 param_i와 param_j가 그리드 방식으로 순차적으로 적용되지 않고, 정확도를 최대로 가지는 컨벡스(Convex) 함수를 찾아서 최적의 정확도를 갖는 param_i와 param_j를 찾게 된다.

그리고 최대 시도수(max_trials)로 탐색하는데 최대 정확도 한계수(limit_max_accuracy_count) 만큼 탐색한 다음 더 이상 정확도가 향상되지 않으면 최적의 파라미터라고 판단한다. 이때, 최대 정확도를 갖게 하는 파라미터 param_i_sel과 param_j_sel을 알고리즘 1과 달리 적응적으로 구하게 된다.

IV. 실험 결과

4.1 성능 지표

제안된 이중 선생-학생 프레임워크로 구성된 전이학습의 성능에 대한 결과를 도출하기 위해서, 다음과 같은 성능 지표들을 사용한다.

- 정확도 : CIFAR-10 테스트 데이터를 사용하여, 10개 클래스를 얼마나 정확하게 구하는지를 판별한다.
- 학습 시간 : 최적의 정확도를 찾기 위해서 학습하는데 소요되는 시간을 실험적으로 산출한다.

- 복잡도 : 각 모델의 복잡한 정도를 나타내는 척도이다. 복잡도를 측정하기 위해서 여기서는 FLOP (Floating Operations)을 사용한다.

4.2 실험 환경 및 결과

본 절에서는 3장에서 기술한 전이학습 알고리즘에 대한 실제 구현 및 실험 결과에 대해서 설명한다. 제안된 전이학습 알고리즘에 대한 성능을 실험하기 위한 주요 환경인 소프트웨어 요구사항은 다음과 같다. 먼저, 소프트웨어 운영체제(OS, Operating System)로는 우분투(Ubuntu) 16.04 버전을 사용하였고, 프로그래밍 언어로는 파이썬(Python) 3.6 버전을 이용하였다. 다음으로, 제안된 전이학습 모델 구현을 위한 개발 툴로는 텐서플로우(Tensorflow) 1.12 버전을 사용하였다.

본 논문에서는 동종 및 이종 딥러닝 모델들을 이용하는 전이학습 간의 성능 비교를 위하여 선생 모델로는 26층 RESNET과 상기 RESNET 보다 네트워크 구조의 복잡도가 상대적으로 낮은 26층 VGGNET[8]을 사용하였다. 본 논문에서 고려한 VGGNET 구조는 [8]의 원래 VGGNET 구조를 기반으로 26층 RESNET과 동일한 층 수 및 컨볼루션(Convolution) 필터 구조를 가지도록 재 구성하였다. 학생 모델로는 14층 RESNET을 사용하였다.

표 1. 전이학습 정확도 (선생 지식이 없는 학생모델 vs 동일 전이학습 vs 이종 전이학습)

Table 1. Accuracy (Student only vs. the homogeneous TL vs. the heterogeneous TL)

Method	1st	2nd	3rd	Avg
S-RES14 without T	90.23	90.30	90.43	90.32
T-RES26, S-RES14	91.12	90.55	91.10	90.92
T-VGG26, S-RES14	90.75	90.57	90.80	90.70

S: student, T: teacher, RES26: 26-layer RESNET, VGG26: 26-layer VGGNET, RES14: 14-layer RESNET

표 1은 각 기법에 대한 정확도를 비교한 결과들을 보여준다. 첫 번째(S-RES14 without T)는 전이학습을 사용하지 않는 학생 모델만 있는 경우이고, 두 번째(T-RES26, S-RES14)는 선생과 학생 모델 모두

동일한 네트워크 구조를 사용하는 경우이며, 세 번째(T-VGG26, S-RES14)는 선생과 학생이 서로 다른 이종 모델인 경우를 나타낸다. 3가지 경우를 비교해보면, 첫 번째의 선생 지식이 없는 학생만 있는 경우보다, 두 번째와 세 번째 전이학습을 사용한 경우 모두가 각각 0.6%, 0.4% 정도 더 좋은 정확도 성능들을 보여 주었다. 또한, 세 번째의 선생과 학생이 이종 모델인 경우가 두 번째의 동종 전이학습과 비교하여 그 정확도 성능이 0.2% 정도 떨어지는 것을 관찰할 수 있다.

하지만, 그림 3의 복잡도 성능을 비교한 결과들을 살펴보면, 동종 딥러닝 모델들을 이용한 전이학습 대비, 이종 전이학습 경우가 약 2000 FLOP 정도 복잡도가 낮음을 알 수 있다. 이는 정확도와 복잡도가 서로 트레이드 오프(Trade-off) 관계에 있음을 알 수 있다.

따라서 표 1과 그림 3의 결과를 토대로 이종 모델 전이학습은 낮은 복잡도에서도 동일 모델 전이학습과 비슷한 정확도 성능을 보여줄 수 있음을 실험적으로 관찰할 수 있었다.

표 2는 GS 기법 및 BO 기법으로 전이학습을 수행한 경우에 정확도를 비교한 결과들을 보여준다. 동일 모델의 경우 GS 기법 및 BO 기법 성능은 각각 91.12%, 91.35%의 결과로써, BO 기법이 GS 기법보다 0.2% 정도 더 높은 정확도를 보여주었다. 또한, 이종 모델인 경우에도 GS 기법 및 BO 기법이 각각 90.85%, 90.96%의 결과로써, BO 기법이 0.1% 정도 더 높은 정확도를 보여주었다.

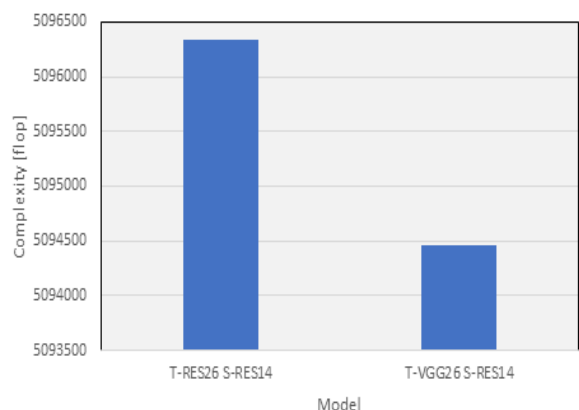


그림 3. 복잡도 (동일 전이학습 vs 이종 전이학습)
Fig. 3. Complexity (the homogeneous TL vs. the heterogeneous TL)

표 2. 정확도(그리드 검색 기법 vs. 베이지안 최적화 기법, 동일모델 전이학습 vs. 이종모델 전이학습)
 Table 2. Accuracy(GS vs. BO, the homogeneous TL vs. the heterogeneous TL)

Method	1st	2nd	3rd	Avg
T-RES26 S-RES14 (GS)	91.37	91.02	90.98	91.12
T-RES26 S-RES14 (BO)	91.35	91.31	91.38	91.35
T-VGG26 S-RES14 (GS)	91.00	90.83	90.71	90.85
T-VGG26 S-RES14 (BO)	90.98	90.96	90.94	90.96

따라서 표 2의 실험결과로부터 동일 모델 및 이종 모델에 상관없이 기존 전이학습에 적용하였던 GS 기법에 비해 본 논문에서 고려한 BO 기법이 더 좋은 정확도 성능을 보여주었다. 이는 GS 기법이 파라미터를 일정 간격으로 나누어 찾기 때문에 최적값을 도출하는데 한계가 있는 반면, BO 기법은 높은 정확도가 나올 가능성이 높은 파라미터를 적응적으로 찾기 때문이다.

그림 4는 GS 기법 및 BO 기법으로 전이학습을 수행한 경우에 학습 시간을 비교한 결과를 보여준다. 동일 모델인 경우에는 흐름정보 전이학습 시 GS 기법 및 BO 기법이 각각 411,665초, 184,498초로써, BO 기법이 약 230,000초 정도 빠른 학습 결과를 보여주었다. 마찬가지로, 이종 모델인 경우에도 GS 기법 및 BO 기법이 각각 403,208초, 147,576초로써, BO 기법이 약 250,000초 정도 빠른 학습 결과를 보여주었다.

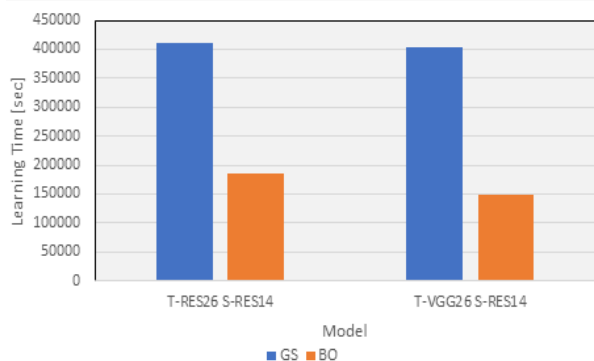


그림 4. 평균 학습시간 (그리드 검색 기법 vs. 베이지안 최적화 기법, 동일모델 전이학습 vs. 이종모델 전이학습)
 Fig. 4. Mean learning time (GS vs. BO, the homogeneous TL vs. the heterogeneous TL)

그림 4의 실험결과로부터 동일 모델 및 이종 모델에 상관없이 기존 GS 기법에 비해 BO 기법이 더 빠른 학습 성능을 보여주었다. 이는 GS 기법은 파라미터를 일정 간격으로 나누어 모든 경우의 수를 전부 계산한 후에 최고 정확도를 찾아가는 방식인 반면, BO 기법은 높은 정확도가 나올 가능성이 높은 파라미터를 적응적으로 찾아가는 방식이기 때문에 학습속도가 훨씬 빠르다.

V. 결론 및 향후 과제

본 논문에서는 상대적으로 복잡도가 낮은 다른 선생 네트워크 구조를 사용하는 경우에도 기존 동종 모델을 사용하는 전이학습과 유사한 정확도를 가지면서 학습속도가 더 빠른 이종 전이학습 방법을 제안하였다. 이를 위하여 VGGNET 선생 모델로부터 흐름정보를 추출하고 이를 다른 네트워크 구조를 가지는 RESNET 학생모델로 이전하는 FSP 기반의 이종 전이학습 프레임워크에 BO 기법을 적용하였다. 본 연구에서 제안한 전이학습 방법에 대한 성능을 평가하기 위하여 CIFAR-10 데이터셋에 대하여 분류 정확도 및 학습시간, 학습모델의 복잡도의 성능 지표들을 사용하였다.

본 연구의 결과에 따르면, 상대적으로 복잡도가 낮은 선생 네트워크를 사용하더라도 동종 모델을 이용하는 기존 FSP 전이학습 대비 유사한 정확도를 가지면서 빠른 학습이 가능함을 실험적으로 관찰할 수 있었다.

References

- [1] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network", NIPS 2014 Deep Learning Workshop, arXiv preprint arXiv:1503.02531, pp. 1-19, Mar. 2015.
- [2] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets", in Proc. 5th Int. Conf. Learning Represent. (ICLR), San Diego, USA, pp. 1-13, May 2015.

- [3] J. Yim, D. Joo, J. H. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization, and transfer learning", in Proc. of 2017 IEEE Conf. Comput. Vision Pattern Recogn. (CVPR), Honolulu, USA, pp. 7130-7138, Jul. 2017.
- [4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", in Proc. IEEE Conf. Comput. Vision Pattern Recogn (CVPR), Las Vegas, USA, pp. 1-12, Jun. 2016.
- [5] Shahriari Bobak and Kevin Swersky, et al., "Taking the human out of the loop: A review of bayesian optimization", Proceedings of the IEEE, Vol. 104, No. 1, pp. 148-175, Jan. 2016.
- [6] "The CIFAR-10 dataset and CIFAR-100 dataset", <http://www.cs.toronto.edu/kriz/cifar.html>, [accessed: May 2019].
- [7] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset", Technical report, 2011.
- [8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", in Proc. 5th Int. Conf. Learning Represent. (ICLR), San Diego, USA, 2015, pp. 1-14, May 2015.

배 지 훈 (Ji-Hoon Bae)



2000년 2월 : 경북대학교
전자·전기공학부(공학사)
2002년 2월 : 포항공과대학교
전자·전기공학과(공학석사)
2016년 2월 : 포항공과대학교
전자·전기공학과(공학박사)
2002년 2월 ~ 2019년 8월 :

한국전자통신연구원 책임연구원
2019년 9월 ~ 현재 : 대구가톨릭대학교 인공지능·
빅데이터공학과 조교수
관심분야 : 인공지능, 딥러닝/머신러닝, 레이다 영상 및
신호처리, 최적화 기법

저자소개

김 귀 훈 (Kwihoon Kim)



1998년 2월 : KAIST 전기 및
전자공학과(공학사)
2000년 2월 : KAIST 전기 및
전자공학과(공학석사)
2019년 2월 : KAIST 전기 및
전자공학부(공학박사)
2000년 2월 ~ 2005년 6월 :

LG데이콤 주임연구원
2005년 7월 ~ 현재 : 한국전자통신연구원 책임연구원
관심분야 : 지능형 에지 컴퓨팅, 인공지능, 딥러닝,
강화학습