ISSN : 1226-7244 (Print)
ISSN : 2288-243X (Online)
논문번호 22-01-03
j.inst.Korean.electr.electron.eng.Vol.26,No.1,19~26,March 2022
https://doi.org/10.7471/ikeee.2022.26.1.19

어텐션 기반의 지속학습에서 정규화값 제어 방법

Regularization Strength Control for Continuous Learning based on Attention Transfer

강 석 훈 *★, 박 성 현*

Seok-Hoon Kang*★, Seong-Hyeon Park*

Abstract

In this paper, we propose an algorithm that applies a different variable lambda to each loss value to solve the performance degradation caused by domain differences in LwF, and show that the retention of past knowledge is improved. The lambda value could be variably adjusted so that the current task to be learned could be well learned, by the variable lambda method of this paper. As a result of learning by this paper, the data accuracy improved by an average of 5% regardless of the scenario. And in particular, the performance of maintaining past knowledge, the goal of this paper, was improved by up to 70%, and the accuracy of past learning data increased by an average of 22% compared to the existing LwF.

요 약

본 논문에서는 LwF에서 도메인 차이에 따른 성능 하락 현상을 해결하기 위해, 각 손실값에 각각 다른 가변람다를 적용하는 알고리즘을 제안하여, 향상된 과거 지식유지가 이루어 지게 한다. 이 지식 전달 기반의 방법을 LwF와 접목하여, 과거 학습 태스크의 지식 유지 성능을 강화하였다. 가변 람다 방법을 추가적으로 적용하여, 현재 학습할 태스크를 잘 학습할 수 있도록 람다 값을 가변적으로 조절할 수 있었다. 본 논문의 제안 방법을 적용하여 학습한 결과 시나리오에 상관없이 평균 5% 정도 데이터의 정확도가 향상하였고, 특히 본 논문의 목표인 과거 지식을 유지하는 성능이 최대 70% 가까이 개선되었고, 과거 학습 데이터의 정확도가 기존 LwF 대비 평균 22% 상승하였다.

Key words : LwF, Continuous Learning, Knowledge Transfer, Variable Lambda, Catastrophic Forgetting

Ⅰ. 서론

지속적 학습환경에서 치명적 망각 현상[1-10]을 줄이기 위한 여러 방법중에 LwF(Learning without Forgetting)[11]가 있다. LwF는 과거 데이터로 학습했던 모델의 출력 값을 유지하여, 과거 학습 데이터의 지식을 우회적으로 보존하여 사용한다. LwF는 과거 네트워크 출력을 유지하여 과거 데이터와 관련된 가중치의 변화를 우회적으로 억제하지만 도메인의 차이가 커짐에 따라 가중치의 변화가 점점 심해지므로 성능의 하락이 발생하게 된다. 이런 성능 하락을 제어하기 위한 방법으로 본 논문에서는 지식 전달(Knowledge Transfer) 방법을 접목한 네트워크를 이용한다. 특히 이 가운데, AT(Attention Transfer)[12] 방법을 이용하여 Teacher가 Student에게 학습 데이터

* Dept. of Embedded Systems Engineering, Incheon National University
★ Corresponding author
Mail : hana@inu.ac.kr, Tel : +82-32-835-8760
※ Acknowledgment
Manuscript received Dec. 28, 2021; revised Mar. 18, 2022; accepted Mar. 19, 2022.
This is an Open-Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License(http://creativecommons.org/licenses/by-nc/3.0) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

에서 집중해야할 특징(Attention)을 전달함으로써 Student가 학습 데이터를 더 잘 학습할 수 있게 돕는 방식을 대상으로 한다. AT는 컨볼루션 레이어의 출력을 사용하여 네트워크가 중요하게 집중하는 부분의 정보를 전달하는 방법을 사용한다. 이때 상위 레이어부터 하위 레이어까지 다양한 레이어들의 출력을 사용하여 높은 수준부터 낮은 수준까지의 표현 정보를 사용한다. 이러한 정보를 가진 Teacher 네트워크에서 작은 규모의 Student 네트워크로 정보를 전달하여 Student 네트워크가 Teacher 네트워크의 지식을 습득할 수 있게 하는 것이다. 본 논문의 네트워크는 이전 태스크의 학습정보를 Teacher Network로 사용하고, 학습된 것을 넘겨받는 Student Network를 다음 태스크로 설정한 변형된 모델을 이용한다.

LwF는 말단의 일부 정보를 유지하여 과거 학습정보를 유지하려는 방식이다. 그러나 네트워크의 중간 가중치 값은 새로운 태스크에 의해 지속적으로 변경되게 되고, 이것으로 인해 학습이 계속 진행되면서 망각으로 이어지게 된다. 따라서 하나의 네트워크에, 유사하지만 다양한 도메인의 데이터를 학습시키려고 하는 본 논문과 같은 지속적 학습 환경에서는, 이전의 학습정보를, LwF가 보존하는 데이터까지 적절하게 잘 전달할 필요가 있는 것이다. 본 논문에서는 이를 위해, AT를 적용한 LwF에서 효율개선을 다룬다.

LwF는 이전에 학습한 데이터와 현재 학습할 데이터가 유사할수록 더 좋은 성능을 보이지만 유사하지 않은 경우에는 성능이 떨어지는 모습을 보인다[11]. 사전 학습된 데이터와 유사도가 떨어질수록, 사전 학습 데이터의 성능이 하락하는 것이 일반적이다. 대부분의 연구에서는 정규화 강도인 λ값을 각 연구 모델에 맞게 적절한 상수값으로 설정하여, 태스크간의 유사도 정보를 직접 설정하여 이용하고 있다[4]. 일반적으로 λ값이 큰 경우 과거 태스크의 성능이 높아지고 현재 태스크의 성능은 낮아지게 된다. 따라서 이전에 학습된 데이터가 제대로 인식되는 정도를 나타내는 정확도 하락을 막기 위해 λ값을 크게 설정하는 방법을 사용할 수 있지만, λ값이 커질수록 현재 태스크의 성능은 반대로 낮아지게 된다. 따라서 지속적 학습환경과 같이 이미 학습된 네트워크에 새로운 태스크가 추가되어 학습되는 경우는, 하나의 상수로 그 값을 정해서 사용할 경우 학습결과가 지속되지 못하고 대부분의 경우에 정확도가 떨어지는 결과를 가져오게 된다. 따라서 각 모델에 맞는 적절한 값을 설정하는 것이 무엇보다 중요하다. 하지만 LwF에서 적절한 λ값을 설정하는 방법에 대한 명확한 기준은 아직 많이 연구되어 있지 않다. 이 문제를 해결하기 위해, 학습하려는 데이터의 특징을 파악한뒤, 과거 지식의 망각 위험이 큰 데이터인 경우 가중치를 상대적으로 높게 설정해 과거 태스크들의 지식을 보존하고, 망각 위험이 적은 데이터인 경우 가중치를 낮게 설정해 현재 태스크들의 지식 습득에 우선을 두는, 가변 람다 방법을 이용하면 이러한 문제를 상당 부분 개선할 수 있다.

지속적 학습이 이루어지는 LwF에 지식전달의 방법을 적용하면, 이전 태스크에 대한 학습결과를 효율적으로 전달하여 사용할 수 있으므로 망각현상의 완화를 기대할 수 있다. 그러나 여기에 정규화 강도를 가변적으로 사용하기 위해서는 네트워크에서 사용하는 손실값에 대해 효율적인 조절이 필요하다. 본 논문의 기반이 되는, 지식전달을 이용한 LwF 방법에서는 지식전달 부분에서도 손실값이 생기고, LwF 부분에서도 별도의 손실값이 생기게 된다. 이들은 최적의 λ값이 각각 적용되어야 학습 정확도가 떨어지는 비율을 개선시킬 수 있다. 본 논문에서는 AT 방법을 이용한 LwF모델에서, 학습과정에서 발생하는 여러 손실값들을 정의하고, 고정된 상수값과, 복잡도 및 유사도에 따른 가변적인 정규화 강도값의 최적의 조합을 실험에 의해 조사한 뒤, 이를 위한 적용 방법을 제안하고 그 실험 결과를 제시한다.

II. 본론

2.1. 관련연구

가변 람다는 과거 출력 분포를 유지하는 강도를 조절하는 정규화 강도인 λ값을 학습 환경에 따라 조절하는 방법이다. 이것을 기존 LwF에 적용하는 방법은 크게 2가지 부분으로 나눌 수 있다. 첫 번째는 데이터의 유사도 측정으로 신경망의 활성화 출력값을 이용하여 유사도를 측정하는 것이다. 서로 다른 데이터를 입력해 출력된 값들을 비교하여 상관 계수를 측정하고 이 측정값을 유사도로 설정한다. 두 번째는 데이터의 복잡도 측정으로 Warm-up 방법을 사용하는 방법이다. 새로운 데이터로 Warm-up 과정을 거치고 정확도를 측정한 후 이전 태스크의 정확도와 비교하여 현저하게 낮게 측정되면 복잡한 데이터, 비슷하거나 약간 낮게 측정되면 단순한 데이터로 판별하게 한다. 이렇게 측정한 유사도와, 복잡도를 사용하여 최적의 λ값을 설정하는 방법이다.

가변 람다 방법에서 학습되는 첫 번째 태스크는 기본 LwF 구조를 기반으로 수행되며 태스크 2 이상부터 가변

람다를 적용하여 학습을 수행하게 한다. 새로운 태스크를 학습하기 전 상관계수 측정을 위한 전 네트워크의 출력값을 측정하고 Warm-up과정을 수행하여 복잡도 측정을 위한 현재 태스크의 정확도를 측정한다. 이후 적합한 λ값을 측정하여 학습에 사용한다.

본 논문에서는 지식전달 방법인 AT방법을 적용한 LwF 모델을 이용하며, 여기에서 망각현상 완화를 위한 가변람다 적용방법을 제안하고, 이에 의한 학습결과 망각률을 개선하는 것을 목표로 한다. 지식 전달을 LwF에 적용하기 위해 변경된 LwF 모델 구조를 이용하고 있다. 기존 LwF는 과거 태스크의 출력을 유지하기 위해 현재 태스크를 사용해 학습을 진행한다. 그러나 지식 전달을 적용한 LwF는 Teacher, Student Network 구조를 통해 가중치가 고정된 Teacher Network의 출력을 사용하여야 한다. 지식 전달 방법중 AT를 적용한 LwF의 손실함수는 (1), (2)와 같이 나타낼 수 있다.

$$L_{total} = L_{lwf} + L_{kt} \tag{1}$$

$$L_{kt} = \| T(I) - S(I) \|_2 \tag{2}$$

L_{total}은 기존 LwF의 손실함수인 L_{lwf}와 AT의 손실함수인 L_{kt}로 이루어져있다. L_{kt}에서 $T(I), S(I)$는 Teacher, Student Network가 학습 이미지 I를 입력으로 받아 출력하는 중간 레이어의 값을 의미한다. 본 논문에서는 기존의 일반적인 AT방법과 달리 N-Channel의 특징값을 Point-wise하게 평균을 이용하지 않고 (2)의 수식으로 L_{kt}를 이용한다. 채널 기준으로 평균을 내는 방법은 출력결과가 가진 많은 정보들을 제거하게 되므로, 과거지식 유지 성능이 명백하게 떨어지게 된다. 본 논문에서는 이 수식을 사용하여 Teacher와 Student 출력의 차이를 줄이는 방향으로 학습을 진행시킨다.

2.2. 손실함수 제어를 위한 람다값 조절

기존 LwF의 손실함수는 수식 (3)과 같다. $L_o(Y_o, \hat{Y}_o)$는 모델의 출력 분포를 유지하기 위한 손실함수 값이다. Y_o는 이전 태스크로 학습된 모델에 새로운 태스크를 입력하여 만들어낸 출력 분포이고 이 값을 모델의 출력 $\hat{Y}_o$의 목표로 설정하여 과거 출력 분포를 유지하게 한다. $L_n(Y_n, \hat{Y}_n)$는 현재 태스크를 위한 Cross Entropy Loss 함수이다. L_2는 L2 정규화 값, λ는 이전 태스크에 대한 정규화 강도이다.

$$L_{lwf} = L_n(Y_n, \hat{Y}_n) + \lambda L_o(Y_o, \hat{Y}_o) + L_2 \tag{3}$$

기존 가변 람다 방법은 수식 (3)과 같이 LwF 손실값에 가중치 값 λ가 존재하여 이 값을 조절하여 학습 상황에 최적화된 값을 설정한다. 하지만 본 논문의 학습모델은 이전 태스크에 대한 지식 전달부분과 LwF가 가지는 손실이 모두 존재하게 된다. 따라서, 기존의 손실함수만으로 제대로 된 학습을 진행할 수 없게 된다. 따라서 본 논문에서는 수식 (4)와 같이 지식전달 기반 방법의 지식전달 손실이 추가된 손실 함수를 정의하여 사용한다. 본 논문에서 제안하는 가변람다 제어는 이 손실함수를 학습이 진행되는 태스크의 복잡도와 유사도에 맞게 조절하게 된다.

$$L_{total} = L_{lwf} + L_{kt} \tag{4}$$

수식 (4)의 손실함수에 람다값을 다양하게 적용해서 실험해보면, 이 값에 의한 학습유지율에 많은 차이가 난다. 따라서 수식(4)에 가변람다값을 적용하여야 하고, 적용된 결과는 수식 (5)와 같이 정의된다.

$$L_{total} = \lambda \cdot (L_{lwf} + L_{kt}) \tag{5}$$

수식 (5)는 두 가지의 손실값에 같은 람다값을 적용하는 결과를 가져오게 된다. 이 경우, 지식전달과 LwF의 학습 특성에 상관없이, 동일한 값이 두군데 모두 적용하게 되어 람다값의 범위에 따라 과거 지식 유지 정도가 완전히 달라지게 된다. 따라서 본 논문에서는, 람다값을 각 손실값에 별도로 적용하여 최적의 학습유지 결과가 나오게 조절하였다. 적용하는 손실함수는 수식 (6)과 같다.

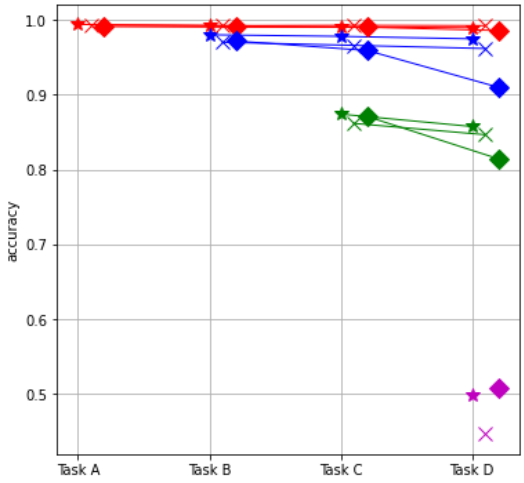
$$L_{total} = \lambda_{lwf} \cdot L_{lwf} + \lambda_{kt} \cdot L_{kt} \tag{6}$$

수식 (6) 의 $\lambda_{lwf}, \lambda_{kt}$ 는 각각 LwF 손실, KT 손실에 적용된 가중치 값이고, λ_{lwf}는 수식 (3)에 정의된 것과 같다.

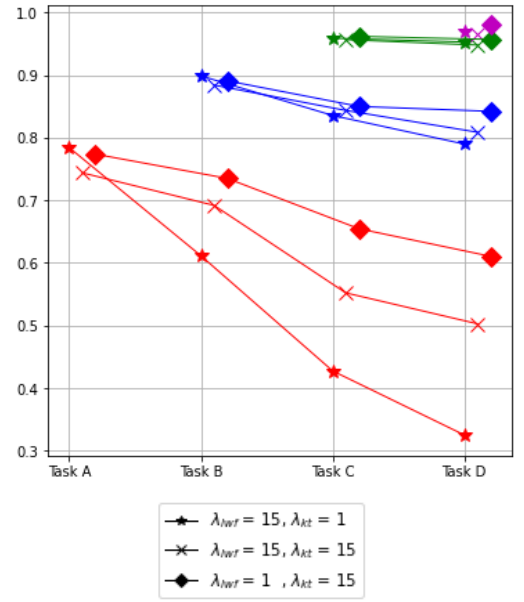
각 손실값에 적용되는 상수값의 차이에 따라 AT가 적용된 LwF의 학습정확도를 확인한 실험 결과를 그림 1에 나타내었다. 여기에서 손실함수의 두 구성요소가 미치는 결과를 확인하기 위해 3가지 경우로 나누어 실험하였다. LwF 손실의 비중이 높은 손실($\lambda_{lwf} = 15, \lambda_{kt} = 1$), 동일한 비중의 손실($\lambda_{lwf} = 15, \lambda_{kt} = 15$), KT 손실의 비중이 높은 손실($\lambda_{lwf} = 1, \lambda_{kt} = 15$)이 그것이다. 1과 15는 임의의 숫자이며, 기존 연구들에서 사용한 값들은 대부분 이 범위 내의 상수를 사용하고 있다. 이 실험에서는 동일한 모델에서 동일한 데이터가 지속적으로 학습될 때, 적용되는 정규화 강도에 따라 학습정확도가 얼마나 다르게 유지되는지를 확인하기 위함이다. 학습되는 태스크가 늘어나면서, 적용되는 정규화 강도값이 달라도 학습정확도에 변

화가 없다면, 정규화 강도값를 다양하게 변화시켜도 별 의미가 없다는 것을 뜻한다. 반대로, 적용되는 정규화 강도값의 적용에 따라, 학습정확도가 달라진다면, 효율적인 정규화 강도값의 적용이 반드시 필요한 것이라고 할 수 있다.

실험은 표 2에 나타낸 2가지 시나리오에 따라 진행하였다. 시나리오 1은 상대적으로 복잡해지는 순서이고, 시나리오 2는 상대적으로 단순해지는 순서로 데이터셋을 나열하였다.



(a) Scenario 1



(b) Scenario 2

Fig. 1. Accuracy difference according to weight ratio in AT applied LwF.

그림 1. AT적용 LwF에서 가중치 비율에 따른 정확도차이

그림 1에서, LwF loss의 비중이 높은 경우가 전체적으로 좋은 성능을 보이고 KT loss의 비중이 높은 loss는 시나리오 2에서 과거 학습 데이터의 지식 유지에 상대적으로 어려움을 보인다. 동일한 비중의 loss는 앞 두 방법의 중간 정도의 성능을 보이고 있다. AT방법을 추가하는 것은 과거 데이터 지식 유지에 분명히 도움이 되며, 특히 학습 순서가 복잡한 데이터에서 단순한 데이터로 진행되는 경우에는 효율이 크다. 그러나 그림 1의 실험결과에서 보듯이 LwF 손실보다 큰 비중을 두는 것은 역효과를 일으킬 수 있다. 따라서 과거 학습결과의 망각현상 완화를 위해서는 데이터 유사도에 따라 각 손실값에 각각 다른 가변람다값을 알맞게 적용하여야 한다.

2.3. 가변람다 적용방법

본 논문에서 가변람다가 적용되는 곳은 KT부분에서 발생하는 곳과 LwF에서 발생하는 곳 두 군데이고, 태스크의 데이터 유사도에 기반하고 있다. 그림 1에서 가중치 변화에 따라 학습결과 유지도가 크게 바뀌는 것을 확인하였듯이, 람다값의 적용은 두군데 모두에서 중요한 결과를 만들어 낸다. 본 논문에서는 데이터 복잡도와 유사도에 따라 기본 람다값을 측정하고, 이것을 태스크의 변화에 따라 이동평균을 적용하여 값을 제어한다. 이를 위한 알고리즘은 표 1에 나타내었다.

Table 1. Training Algorithm of This Paper.

표 1. 본 논문의 학습 알고리즘

Require: Dataset, $D=(D_1,\dots,D_T)$, α, β, C_1, C_2
  Teacher Network: ***Teacher***,
  Student Network: ***Student***
for each Task $t=1,\dots,T$ do
  if $t=1$ then
    Train the using normal cross entropy loss
    Calculate Task 1 accuracy and
      network activation value A_1
  else
    Copy ***Student*** to ***Teacher***
    Make output layer for Task t on ***Student***
    Do warm-up process
    Calculate Task t accuracy acc_t and
      network activation value A_t
    Calculate Correlation Coefficient x with A_{t-1}, A_t :
    if $(acc_{t-1}\times 0.5) < acc_t$ then
      $\lambda 1 = -\ln(-x+1)\times\alpha$
      Calculate New $\lambda 2$ using Previous $\lambda 2$
    else
      $\lambda 2 = -\ln(x)\times\beta$
      Calculate New $\lambda 1$ using Previous $\lambda 1$
  Train the ***Student*** using LwF
  Update A_t, acc_t

먼저 첫태스크의 경우는 태스크를 학습하며, 다음 태스크에서는 새로운 학습을 하기 전에 전 네트워크의 출

력값과 새 태스크의 웜업결과를 비교하여 태스크 정확도를 측정한다. 이 정확도에 따라 상관계수와 데이터 변화율정도를 조절하여 새로운 태스크를 위한 람다를 정한다. 이 람다는 새로운 태스크가 생길때마다 새롭게 계산하며, 급격한 변화를 막고, 추세에 따라 적용하기 위해 이동평균을 이용한다. 이 데이터 복잡도와 유사도에 따라 KT와 LwF 부분의 시작값과 증감방향 그리고 증감비율을 정하며, 새로운 태스크에 적용하게 된다. 이때 사용되는 상관계수는 0~1 사이값, 시작 람다값은 0.1~50 사이의 값의 분포를 가지며, 이론적으로 람다값은 0~∞의 값을 가질 수 있으나, 람다값이 0에 너무 가깝거나, 50을 넘어갔을 때 학습결과에 일관성을 보이지 않을수 있어서 람다값은 0.1~50 범위를 최적의 값으로 설정하였다. 값의 적용비율 또는 변화율을 정할 수 있으며, 이것은 새로운 태스크와 이전 태스크와의 유사도에 일정한 비율을 적용하여 조정할 수 있다. 이것은 실험에서 상수로 정하는 값으로 값이 너무 급변하는 것을 제한해준다.

2.4. 실험 및 결과

본 논문에서 사용하는 모든 실험의 설정은 표 2와 같다. 지속적 학습 환경에서 발생하는 치명적 망각 현상은 태스크 데이터 사이의 유사도에 영향을 많이 받는다 [3,11]. 따라서 실험 시나리오는 단순한 특징을 가진 데이터에서 점차 복잡한 특징을 갖는 데이터로 학습이 진행되는 시나리오와, 반대로 복잡한 특징을 갖는 데이터에서 점차 단순한 특징을 갖는 데이터로 진행되는 시나리오의 2가지 구성으로 진행한다. 실험에는 Mnist, Emnist, Fashion mnist, Cifar10 데이터를 사용하였다.

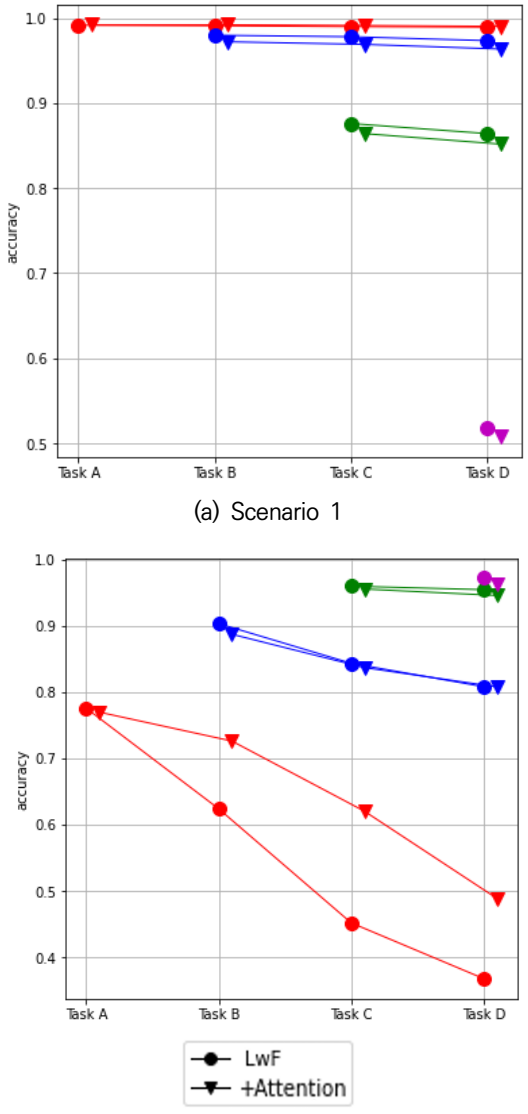
Table 2. Experimental Scenario.
표 2. 본 논문의 실험 시나리오

Scenario \ Task	Task A	Task B	Task C	Task D
1	Mnist	Emnist	Fashion	Cifar10
2	Cifar10	Fashion	Emnist	Mnist

표 2에서, 시나리오 1은 단순한 특징을 가진 데이터에서 점차 복잡한 특징을 갖는 데이터로 학습이 진행되는 시나리오이고, 시나리오 2는 복잡한 특징을 갖는 데이터에서 점차 단순한 특징을 갖는 데이터로 진행되는 시나리오이다. 학습 배치 크기는 100개, 에포크는 각 태스크별로 20회로 설정하였다.

실험은 두가지 방법으로 진행한다. 본 논문의 기반이 되는 AT 적용 LwF 방법이 기본적인 LwF보다 효율이 개선되었음을 먼저 보이고, 그 뒤에 본 논문의 방식이 이 두가지 방법보다 더욱 개선된 결과를 나타내고 있음을 보여, AT 적용의 LwF에서 가변람다 방식이 효과적임을 보인다. 각각의 결과는 그림 2와 그림 3에 나타낸다.

기존의 LwF와, AT 방법을 적용한 LwF에서 가변람다를 적용하지 않았을 경우의 결과는 그림 2와 같다. (a)는 시나리오 1의 결과이고, (b)는 시나리오 2의 결과이다. 그림 2의 시나리오 1의 결과에서, 점점 복잡한 특징을 갖는 데이터로 학습되는 경우, LwF의 결과가 약간 우세하지만 대동소이한 차이를 보이고 있다. 이는 LwF 손실값에, 과거 학습데이터의 지식 유지를 위한 추가적인 손



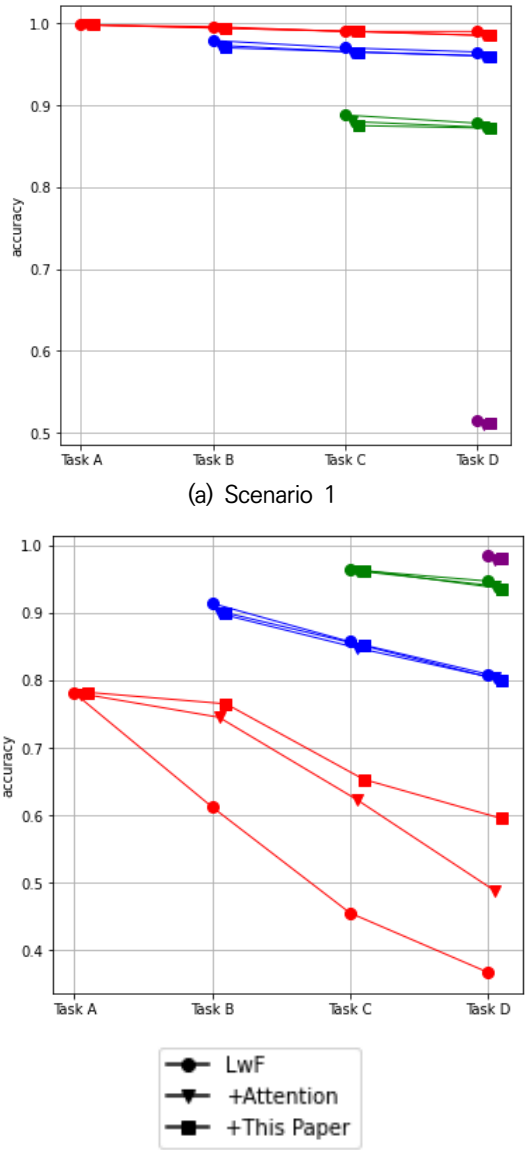
(a) Scenario 1

(b) Scenario 2

Fig. 2. Comparison graph of experimental results between LwF with AT and conventional LwF.
그림 2. AT가 적용된 LwF와 기존 LwF의 실험결과 비교 그래프

실이 추가됨에 따라 생기는 제약때문에 발생하는 성능의 변화이다. 이러한 현상은 기존 LwF 손실의 가중치를 크게 설정했을 때에도 발생하는 현상이다. 시나리오 2의 경우, 점점 단순한 데이터로 학습이 진행되었으며, LwF와 비교하여 과거 학습데이터 지식 유지 능력이 더 높은 것을 알 수 있다.

그림 2의 결과에서 알 수 있듯이, 데이터가 복잡해 지는 경우는 제약이 가장 적은 기본 LwF와 AT를 적용한 방법 모두 결과에 변화가 적은 반면, 데이터가 단순해 지는 경우에는 AT를 적용한 방식의 학습 유지 능력이 개선되어, 기본 LwF보다, AT를 적용하는 것이 더 나은 방향임을 알 수 있다.



(a) Scenario 1

(b) Scenario 2

Fig. 3. Comparative experiment result graph.
그림 3. 최종 실험 결과 그래프

본 논문에서는, 이 개선된 AT 기반의 LwF에 표 1과 같은 방법의 가변람다를 AT 부분과 LwF 부분 모두에 적용하여, 그 효율을 더욱 개선하였다. 그림 3에 본 논문에 의한 가변람다 적용 방법을 기본 LwF는 물론, AT만 적용된 LwF의 결과와도 비교하여 나타내었다. 표 3에는 각 실험 결과에 대한 태스크별 최종 정확도를 나타내었다.

본 논문의 방법이 LwF에 비해서 시나리오 1의 경우는 약간의 차이만 있을뿐 유사한 성능을 보이고 있고, AT방법을 적용한 결과와도 거의 같은 성능을 보이고 있다. 제약이 상대적으로 적은 LwF와 비교했을 때, 본 논문의 방법이 제약이 더 많음에도 불구하고 거의 비슷한 결과를 보이고 있음을 알 수 있다. 시나리오 2의 경우, 최근 학습한 결과는 3가지 방법이 모두 유사한 결과를 보였지만, 본 논문의 목표인 오래된 학습결과에 대한 성능 유지 성능이 상당히 개선된 것을 보이고 있다. 기본 LwF 방법과 비교했을 때, 가장 먼저 학습한 Task A에 대해 약 70%의 성능개선을 보이고 있고, AT 방법만 적용되었을 때와 비교해서도 약 20%의 성능이 개선된 것을 알 수 있다. LwF와 지식 전달 기반 방법을 결합하고 여기에 각 손실값에 가변적인 람다값을 적용한 것이 과거 지식 유지에 뛰어난 성능이 있음을 나타내는 것이다.

Table 3. Final Accuracy by Task for Experimental Results.
표 3. 실험 결과에 대한 태스크별 최종 정확도

Scenario	Method	Task A	Task B	Task C	Task D
1	LwF	98.9	97.3	86.4	51.7
	+Attention	99.0	97.2	86.4	50.8
	This Paper	99.1	97.2	85.7	49.3
2	LwF	36.7	80.8	95.4	97.2
	+Attention	48.8	80.9	95.1	97.1
	This Paper	59.5	83.1	95.1	97.1

태스크가 길어질때에 대한 성능도 그림 4와 같이 검증하였다. 그림 4는 시나리오 2에서 태스크가 2배이상 길어질때에 대한 결과이다. 시나리오 1의 형태로 길어진 경우는 제약이 가장 적은 LwF가 근소하게 우세하다.

그림 4의 경우는 본 논문의 방법을 LwF와도 비교하고, 수동으로 설정한 최상의 결과(+Value로 표시)와도 비교하였다. 본 논문의 결과는 태스크가 지속될수록 LwF와 비교하여 훨씬 나은 성능을 보이고 있다. 수동으로 설정한 결과는 알고리즘에 의하지 않고, 각 태스크마다 최상의 결과가 나오게 수동으로 값을 설정한 결과이다. 본 논문의 결과가 대부분 수동결과와 거의 유사한 성

능을 보이고 있으며, 미세하게 조정된 부분이 적용될 수 있어서, 일부 태스크에 대해서는 오히려 더 좋은 성능이 나오기도 한 것을 볼 수 있다.

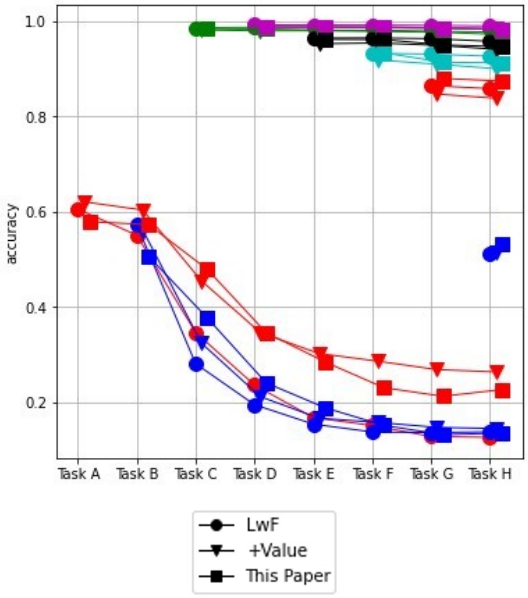


Fig. 4. Experiment result Graph in Case of long Tasks.
그림 4. 태스크가 길어진 경우 실험 결과 그래프

본 논문의 결과는 다른 지식 전달 기반 방법들과 비교 실험에서도 13~18% 향상된 성능을 보였다. LwF는 제약이 가장 적기 때문에 새 태스크에 대한 결과가 가장 좋다. 그러나 과거 태스크의 성능은 현저하게 떨어지게 된다. 본 논문의 방법은 새로운 데이터의 학습 능력의 하락을 최소화하여, 새 태스크에 대해서는 기존의 LwF와 큰 차이가 없으면서, 과거 학습 데이터의 지식을 효과적으로 보존할 수 있는 결과를 보인다고 할 수 있다.

Ⅲ. 결론

본 논문에서는 LwF에서 도메인 차이에 따른 성능 하락 현상을 해결하기 위해 지식 전달 방법을 LwF와 접목시켰고, 여기에 각 손실값에 각각 다른 가변람다를 적용하는 알고리즘을 제안하여 향상된 과거 지식유지가 이루어 지고 있음을 보였다. 지식 전달 방법의 대표적 방법을 적용한 결과이다. 본 논문의 방법은 AT에 기반을 두고, 과거 지식 유지 성능이 상대적으로 높지만 새로운 지식을 습득하는 능력이 부족한 AB(Activation Boundaries) [13] 적용 방법과 과거 지식 유지 성능이 상대적으로 낮지만 새로운 지식을 습득하는 성능이 크게 떨어지지 않는 AT 적용 방법의 타협점인 변형 방법을 이용한 것으로 볼 수 있다. 이 지식 전달 기반의 방법을 LwF와 접목하여 과거 학습 태스크의 지식 유지 성능을 강화하였으며, 가변 람다 방법을 추가적으로 적용하여 현재 학습할 태스크를 잘 학습할 수 있도록 람다 값을 가변적으로 조절하였다. 본 논문의 제안 방법을 적용하여 학습한 결과 시나리오에 상관없이 평균 5% 정도 데이터의 정확도가 향상하였고, 특히 본 논문의 목표인 과거 지식을 유지하는 성능이 최대 70% 가까이 개선되었고, 과거 학습 데이터의 정확도가 기존 LwF 대비 평균 22% 상승하였다. 따라서 본 논문의 방법이 기존 지식 전달 적용 방법에서 발생하는 새로운 데이터 학습 성능 저하 현상을 완화하며 효율적으로 과거 지식을 유지할 수 있음을 보여준다. 지속적 학습 환경은 일관된 특징을 가진 데이터들이 들어오는 환경이 아닌 치명적 망각 현상이 발생하기 쉬운 다양한 특징을 가진 데이터들이 들어올 수 있는 환경이다. 이 환경에서 과거 지식 유지 성능을 강화하는 지식 전달 기반의 방법과 학습 환경을 고려한 가변 람다 방법을 적용함으로써 본 논문의 제안 방법은 다양한 학습 환경에서 인공 신경망의 효율적인 학습을 위한 솔루션이 될 수 있다.

References

[1] R. M. French, “Catastrophic forgetting in connectionist networks,” *Trends in cognitive sciences*, vol.3, no.4, pp.128-135, 1999.
DOI: 10.1016/s1364-6613(99)01294-2
[2] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, “Continual lifelong learning with neural networks: A review,” *Neural Networks*, vol.113, pp.54-71, 2019.
DOI: 10.1016/j.neunet.2019.01.012
[3] F. Zenke, B. Poole, and S. Ganguli, “Continual learning through synaptic intelligence,” *Proceedings of the 34th International Conference on Machine Learning*, vol 70, pp.3987-3995, 2017.
DOI: 10.5555/3305890.3306093
[4] Y. Hsu, Y. Liu, A. Ramasamy, and Z. Kira, “Re-evaluating continual learning scenarios: A categorization and case for strong baselines,” arXiv:1810.12488, 2019.
[5] J. Yoon, E. Yang, J. Lee, and S. J. Hwang, “Lifelong learning with dynamically expandable

networks," arXiv:1708.01547, 2017.
[6] H. Shin, J. K. Lee, J. Kim, and J. Kim, "Continual learning with deep generative replay," arXiv: 1705.08690, 2017.
[7] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *NIPS Workshop*, arXiv:1503.02531, 2014.
[8] K. McRae, and PA. Hetherington, "Catastrophic interference is eliminated in pretrained networks," *Proceedings of the 15h Annual Conference of the Cognitive Science Society*, pp.723-728, 1993. DOI: 10.1.1.30.4449
[9] R. M. French, "Catastrophic forgetting in connectionist networks," *Trends in cognitive sciences* 3.4, pp.128-135, 1999. DOI: 10.1016/S1364-6613(99)01294-2
[10] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol.114, no.13, pp.3521-3526, 2017.
[11] Z. Li and D. Hoiem, "Learning without forgetting", *IEEE transactions on pattern analysis and machine intelligence*, vol.40, no.12, pp.2935-2947, 2017. DOI: 10.48550/arXiv.1612.00796
[12] S. Zagoruyko, and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," arXiv:1612.03928, 2016.
[13] B. Heo, M. Lee, S Yun, and JY. Choi, "Knowledge transfer via distillation of activation boundaries formed by hidden neurons," *Proceedings of the AAAI Conference on Artificial Intelligence,* Vol.33, No.1, pp.3779-3787, 2019. DOI: 10.48550/arXiv.1811.03233

BIOGRAPHY

Seok-Hoon Kang (Member)

1989 : BS degree in Electronic Communications Engineering, Hanyang University.
1995 : PhD degree in Electronic Communications Engineering, Hanyang University.

2004~ : Professor, Embedded Systems Engineering, Incheon National University.
A.I., Deep Learning, Mobile/Embedded System, Wearable System, Natural Language Processing

Seong-Hyeon Park (Member)

2020~ : MS degree in Embedded Systems Engineering, Incheon National University.
A.I. Deep Learning, Embedded Systems Engineering