The Journal of the Convergence on Culture Technology (JCCT)
Vol. 8, No. 1, pp.545-550, January 31, 2022. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2022.8.1.545

JCCT 2022-1-62

딥 러닝에서 Labeling 부담을 줄이기 위한 연구 분석

An Analysis of the methods to alleviate the cost of data labeling in Deep learning

한석민*

Seokmin Han*

요 약 딥러닝은 많은 데이터를 필요로 한다는 것은 이미 널리 알려져있다. 이를 통해, 딥러닝에 쓰이는 신경망의 수없이 많은 parameter들을 학습시킨다. 학습과정에는 데이터뿐 아니라, 각 데이터별로 전문가가 입력한 label이 필요한 경우가 대부분인데, 이 label을 얻는 과정은 시간과 자원 소비가 심하다. 이 문제를 완화하기 위해, few-shot learning, self-supervised learning, weak-supervised learning등이 연구되어오고 있다. 본 논문에서는, label을 상대적으로 적은 노력으로 수행하기 위한 연구들의 동향을 살펴보고, 앞으로의 개선 방향을 제시하도록 한다.

주요어 : 딥 러닝, 유사성, 불확실성, 라벨링

Abstract In Deep Learning method, it is well known that it requires large amount of data to train the deep neural network. And it also requires the labeling of each data to fully train the neural network, which means that experts should spend lots of time to provide the labeling. To alleviate the problem of time-consuming labeling process, some methods have been suggested such as weak-supervised method, one-shot learning, self-supervised, suggestive learning, and so on. In this manuscript, those methods are analyzed and its possible future direction of the research is suggested.

Key words : Deep Learning, Labeling, Uncertainty, Similarity

Ⅰ. 서 론

최근들어, Deep Learning이 영상, 음성신호를 기반으로 Malware탐지[1], 과일분류[2], 주식 투자[3] 등 다양한 분야에 적용되어 연구가 활발히 진행되어오고 있다. Deep Learning에서는 많은 수의 perceptron layer들을 상호 연결하는 weight들을 data를 이용하여 적절한 값을 구하는 과정을 거친다. 여기서 일반적으로 굉장히 많은 수의 데이터를 필요로 한다. 이 weight값을 정해주는 과정을 학습이라고 하고, 이 학습과정을 통해 neural network의 weight들이 자동적으로 계산된다. 이를 통해 기존의 machine learning 보다 좋은 성능을 나타내어, 여러 분야에서 어떻게 적용하여야 할 것인지, 그리고 이미 적용된 분야에서는 어떻게 더 좋은 성능을 낼 것인지 연구되고 있다. 그러나, weight값을 구하기 위한 많은 데이터를 수집해야 할 뿐 아니라, 학습을 위해 쓰이는 학습데이터(training data)에 대해 데이터별로 정답(label)이 필요하다. 이 데이터별 정답은 expert

*정회원, 한국교통대학교 데이터사이언스전공 부교수 (제1저자)
접수일: 2021년 12월 29일, 수정완료일: 2022년 1월 1일
게재확정일: 2022년 1월 8일

Received: December 29, 2021 / Revised: January 1, 2022
Accepted: January 8, 2022
*Corresponding Author: seokmin.han@ut.ac.kr
Dept. of Data science, Korea National University of Transportation, Korea

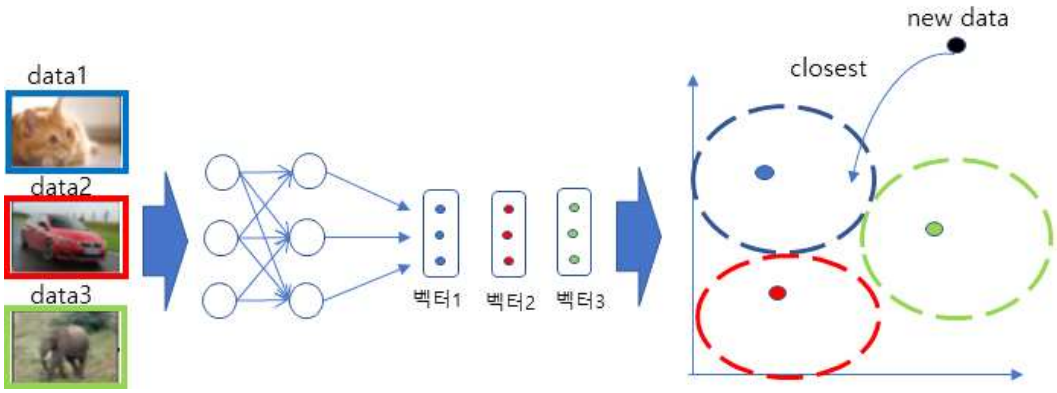

그림 1. few-shot learning의 개념도
Figure 1. The illustration of few-shot learning

또는 oracle로 일컬어지는 존재에 의해 부여되는데, 일반적으로 사람이 수행한다. 많은 데이터를 필요로 하는 deep learning의 특성상 데이터 labeing을 위해 상당한 노동력과 시간이 필요하다. 이런 노력과 시간을 줄이기 위해, few-shot learning, self-supervised learning, weak-supervised method, semi-supervised method등, labeling 에 드는 자원을 줄이려는 연구가 진행되어 왔다. 본 논문에서는, labeling을 가능한 한 줄이면서도 성능을 유지하기 위한 여러 방법들의 현황을 살펴보도록 한다.

II. 본 론

1. Few-shot learning과 self-supervised learning 연구 동향

1) Few-shot learning

Few-shot learning은 각 class별로 단지 몇 장(few shot)만의 데이터를 중심으로 학습하여 원하는 업무(task)를 해결하는 것을 목표로 한다. 대부분의 경우, 자동차(car)와 얼굴(face), 개(dog)와 고양이(cat)등, 서로 범주가 다른 데이터를 각각 학습하고, query data라는 형식으로 서로 다른 범주의 데이터를 pair로 입력, 같은 부류의 데이터인지 아닌지 구별해내는 방식으로 작동한다. 이때, 서로 다른 범주의 데이터를 구분하는 classifier로 deep learning neural network를 적용한다. 기존의 supervised learning과 다르게, class별로 단지 몇 개의 training data만 갖고 있는 것으로 가정하므로, classifier의 학습 방향은 각 데이터의 class를 구별하는 것이 아닌, 입력된 두 data가 같은 부류에 속하는 것인지 아닌지를 알아내는 것을 목표로 한다. 따라서, 두 data가 같은 부류에 속하는 경우는 가까운 거리로, 각각 다른 부류에 속하는 경우는 먼 거리로 표현하도록 학습하는 것이 특징이다. 거리를 비교해야 하므로, 거리 비교를 위한 데이터의 특징벡터(feature vector)가 매우 중요하다. 이 vector가 출력 그 자체이거나, 출력에 핵심적 요소가 되는 경우가 많다. 이전 연구방법 방법으로는 각 데이터를 벡터로 표현해서 상호간 거리를 측정하는 Metric Learning[4][5][6] 방식이 대표적이다. 이 개념은 그림 1에 표시되어있다. 그림 1에서와 같이, neural network에 데이터가 입력되고, 각 데이터에 해당하는 특징벡터(feature vector)가 출력된다. 이 neural network는 같은 범주의 데이터는 특징 벡터간의 거리가 가깝고, 다른 범주의 데이터는 거리가 멀도록 학습시킨다. 그러면, 새로운 입력이 주어진 경우, 동일한 neural network를 통해 특징벡터를 출력하고, 이 출력된 특징벡터와 가까운 범주로 할당된다.

2) self-supervised learning

Deep learning에 사용되는 self-supervised learning은, 궁극적으로는 self-supervised의 뜻과 같이 미리 주어진 labeling 없이 학습을 진행할 수 있는 것을 목표로 한다. labeling이 되어있지 않은 데이터들로 학습이 완료된 task에서 사용된 neural network의 weight을, 이미 학습이 완료된 task와는 다른 별도의 task에 fine-tuning 후 적용하는 방식이 주로 쓰이고 있다[7][8][9]. 미리 학습이 완료된 task는 labeling이 되어있지

않은 데이터로 학습하였으므로, 판별(classification)보다는 데이터 자체의 분포 또는 표현방식을 학습한 상태로 볼 수 있다. 이 task에서 쓰인 weight을 사용하여 별도의 task를 위한 label을 스스로 만들어내는 것으로 해석할 수 있다는 점에서, 별도의 labeling 없이 학습을 진행하는 비지도 학습(unsupervised learning)과는 약간의 차이가 있다고 보기도 한다. 그림 2에서와 같이, 미리 학습된 task를 pretext task 라고 하고, 이 학습된 task에 이용된 weight를 복사하여 다른 task에 적용하는 과정을 전이(transfer)라고 부른다. 그리고, 이 전이 과정을 거쳐 적용되는 별도의 trask를 downstream task라고 부른다. 이 개념은 그림 2에 나타나 있다.

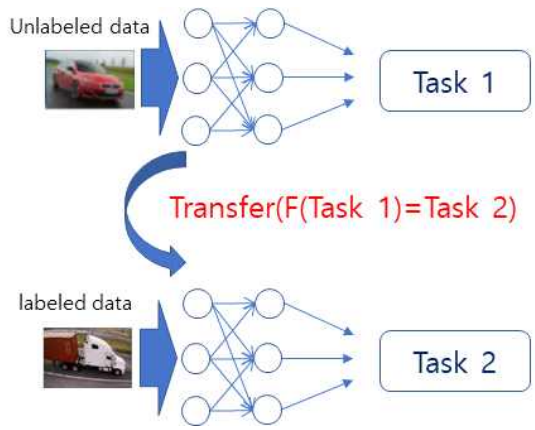

그림 2. transfer를 이용한 pretext learning의 개념
Figure 2. The concept of pretext learning using transfering

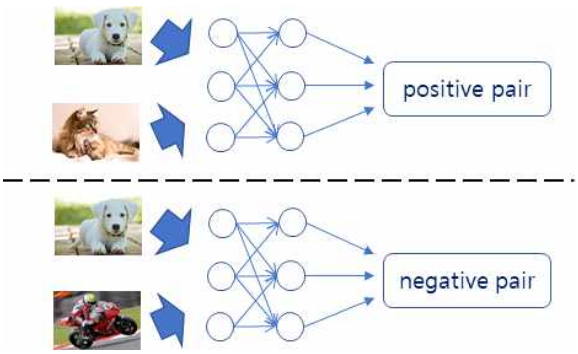

그림 3. Positive pair와 negative parir를 이용한 contrastive learning 의 개념
Figure 3. The concept of contrastitve learning using positive pair and negative pair

self-supervised learning의 또 다른 흐름은, contrastive learning[10][11]이라는 방식이다. 이 방식은, 범주가 비슷한 부류의 데이터에 대한 classification 결과가 범주가 다른 부류의 데이터에 대한 classification결과보다 정답에 가까울 확률이 높다는 것을 가정한 방법이다. pretext task와 downstream task를 이용한 방식보다 학습량에 있어 유리하다고 알려져 있다. 그림 3 에서와 같이, 유사한 범주에 속하는 데이터짝(pair)와, 다른 범주에 속하는 데이터짝으로 학습을 진행하는데, 이는 few-shot learning과 비슷한 점이 있다.

2. active learning & semi-supervised learning 연구 동향

Active learning은, 데이터베이스에서 학습에 적합한 데이터들을 골라내는 작업에서부터 시작하고, 이 과정이 가장 중요하다. 다양한 기존 연구에서 데이터를 골라내는 방법을 제시하고 있다[12-20]. 이렇게 골라진 데이터들에 대해, expert 또는 oracle이란 존재가 labeling을 수행한다. 이렇게 얻어진 labeling과 데이터를 기반으로 classifier를 학습시키는 것이 주된 프로세스이다. 대부분의 경우, 라벨이 있는 데이터가 일부 있거나 또는 전혀 없는 상태로 초기상태(initial state)를 구성한다. 라벨이 있는 데이터가 전혀 없는 경우라면 variance 등의 metric을 통해 불확실성 또는 유사성을 계산하고 그 결과에 따라 순서를 정한다(ranking). 라벨이 있는 데이터가 일부라도 있는 경우라면 이를 이용해 network를 pre-training 시킨 후, 가장 대표성이 높은 데이터를 선정하여 oracle에게 labeling을 요청한다. 대표성이 높은 데이터라는 뜻은, 유사성가 높거나 불확실성이 높은 데이터를 뜻한다. 라벨이 주어진 데이터는 labeled data가 되고, 이를 이용하여 기존의 labeled data에 더하여 network를 학습시킨다. 이 과정을 반복적으로 요청하는 것이 active learning의 일반적인 형태이다. 여기서, labeling 횟수와 요청량을 최소화하는 것이 목표다.

지금까지 많은 종류의 연구에서 대표성이 높은 데이터 샘플(informative samples)을 찾는 것을 목표로 하는데, 대표성을 계산하기 위해 많이 쓰이는metric이 불확실성과 유사성라고 할 수 있다.

1) 유사성(similarity)

유사성을 이용한 방법은, 각 데이터들을 비교하여, 상대적으로 다른 데이터들과 가장 비슷한 데이터를 가장

대표성을 갖는 데이터로 가정하고 이를 찾아낸다. 데이터베이스 내의 데이터들을 특정 metric을 이용해 상호 비교하는 과정이 필요하다. 유사성을 측정하기 위해서는 그 중심이 되는 구심점이 필요한데, 이는 자연스럽게 clustering 과정을 거치게 된다. cluster의 중심에 가까운 데이터또는 그 중심 자체가 representative sample이 된다. 유사성을 계산하기 위해 사용되는 metric은 mutual information[12] , cosine 유사성[13] 등과 같은 수학적으로 널리 쓰이는 metric 뿐 아니라, GOLD estimator[14] 와 같은 독자적인 개념의 metric을 사용하는 경우도 있다. 이 개념을 그림 4에 나타내었다. 그림 4 에서와 같이, 각각의 class의 중심점을 계산할 수 있고, 이 중심점에서부터 각 data들의 거리를 유사성으로 정의한다.

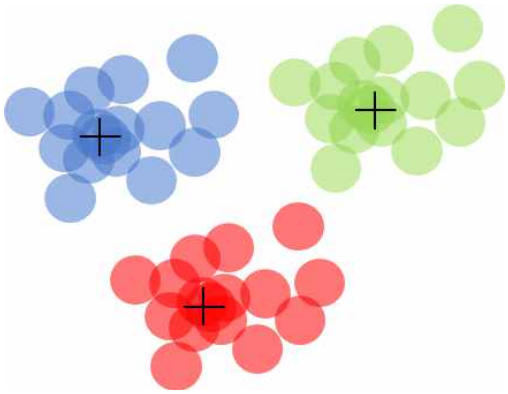

그림 4. 유사성의 개념. 각 클래스의 중심이 십자로 표시되어있다.
Figure 4. The concept of Similarity. Each class has the corresponding color. The center of each class is represented as a cross. Data are clustered around the center.

2) 불확실성(uncertainty)

불확실성을 이용한 방법은, 각 데이터들을 비교하여, 서로 가장 거리가 멀고, 특정 class에 할당하기 어려운 data를 골라내는 작업이다. 분산(variance)과 같은 수학적 metric을 이용하기도 하지만[15][16], neural network의 weight값의 불확실성을 확률적으로 이용하기도 하고[17], classifier의 출력값의 확률을 이용하여 데이터의 불확실성을 정의하기도 한다[18]. 또한, neural network의 특성상 학습 결과에 따라, 같은 데이터의 경우에도 classification 결과가 조금씩 변하는데, 이런 데이터가 불확실성이 높다고 하는 연구도 있다.[19][20]. 불확실성은 결국 그림 5 에서와 같이 각 class에 따른 경계선을 확정하는 과정이라고 할 수 있다., Classifier가 가장 어려워하는 부분이 바로 각 class간 경계선 부분인데, 이 부분에 해당하는 데이터 샘플은 oracle에게 판별을 요청하는 것이다.

3) 불확실성와 유사성을 이용한 informative sample 선택

불확실성과 유사성 모두 중요한 특성이고, 두 가지 특성을 동시에 고려하여 labeling이 꼭 주어져야 하는 데이터의 수를 가능한 한 줄이려고 하는 많은 연구들이 있다[21-30]. .전체 데이터 수에 비해 훨씬 적은 수의 데이터에 대해서만 labeling 하더라도, 전체 데이터를 이용할 때와 비교하여 약간 낮거나 비슷한 수준의 학습 성능 이라면, labeling 부담을 경감시켜 실생활에 적용하기 매우 유리해진다. 동일 class에 속하는 데이터끼리는 상호간 거리가 까깝도록, 다른 class에 속하는 데이터간에는 상호간 거리가 더 멀어지도록 특징벡터를 구성하도록 cost function을 구성하는 것이 핵심 개념이 된다.

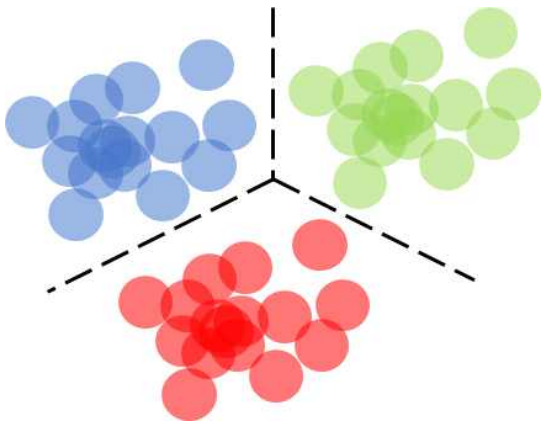

그림 5. 불확실성의 개념. class간 경계선에 가까운 데이터는 불확실성이 높다고 할 수 있다.
Figure 5. The concept of uncertainty. If a sample is close to the decision plane between classes, its uncertainty is considered to be high.

III. 분석 및 한계점

1. Few shot learning과 self-supervised learning

Few shot learning과 self-supervised learning의 경우, 일부 주어진 label을 근거로, label이 없는 데이터의 특징 벡터가 서로 많이 차이가 나도록 학습한다는 공통점이 있다. 주로 연구된 데이터도 Siamese 데이터셋 또는

ImageNet 데이터셋과 같이, 상호간 모양이나 특성이 다른 데이터를 다루고 있다. 현재까지 좋은 결과들이 있으나, 의료영상에서 폐 영상간 비교, 또는 신장 영상간의 비교 등, 서로 비슷한 특성을 지니는 영상에서 미세한 차이를 찾아내기 위한 연구에는 적합하지 않을 수 있다는 것이 한계라고 할 수 있다.

2. Active learning과 weak&semi-supervised learning

유사성, 불확실성를 혼합하여 사용한는 방법은 일반적으로 expert에 의해 주어진 metric이 필요하다. 이것을 학습시 cost function의 형태로 이용하나, 이 metric을 자체를 고안하거나 선택해야 한다는 점은 한계점이다. 데이터마다, 상황마다 적합한 metric이 다를 가능성이 크다. 더불어, 데이터베이스에 속한 모든 데이터간의 상호 metric을 계산해야하는데, 데이터량이 선형적으로 증가하더라도, 상호 비교에 의한 계산시간은 지수적으로 증가하는 문제가 있다. 물론 metric 연산을 다량으로 수행하는 것은, expert의 annotation을 수행하는 것 보다 편의성 및 필요한 성능을 얻기 위한 training 시간을 예측할 수 있다는 면에서 유리해지는 것은 사실이나, data가 늘어날수록 필요한 계산시간이 기하급수적으로 늘어나는 것은 실용성 측면에서 문제가 될 수 있다. 따라서, 추후 연구 방향은 유사성와 불확실성을 대체할 더 좋은 metric을 구현한 cost function의 구성, 그리고 데이터 상호간 유사성와 불확실성 등을 비교하는, 데이터 수에 따른 지수적 계산시간 증가 문제를 회피하거나 완화할 수 있는 방향이라고 할 수 있다[31].

Ⅳ. 결 론

Deep learning을 이용한 연구를 위해서는, 대용량 data에 대한 labeling이 필요하고, 이 부담을 완화하기 위한 여러 가지 방법 연구되어왔다. 본고에서는, 현재까지 labeling 부담을 감소하기 위한 기존 연구들의 방식을 정리하고, 앞으로 새로운 연구들이 진행되어야 할 방향을 정리해보았다. 앞으로의 연구는, 유사성와 불확실성을 포함하는 metric을 cost function으로 어떻게 구현하느냐의 방향과, 각 데이터간 상호 비교하는 과정에서 야기되는 지수적 계산량 증가 문제를 완화하는 방향으로 진행될 것을 제시하였다. 제시된 부분에 대한 연구가 더 발전된다면, 대용량 데이터이용한 deep learning method를 더 실용적으로 이용할 수 있을 것으로 예상한다.

References

[1] Phil-Joo Moon, “Analysis of Deep Learning Methods for Classification and Detection of Malware,” International Journal of Advanced Culture Technology Vol.9 No.3 291-297 (2021). https://doi.org/10.17703/IJACT.2021.9.3.291

[2] Weol-Youg. Kim, Seung-Jung. Shin, “Large orchard apple classification system,”The Journal of the Convergence on Culture Technology (JCCT), Vol. 4, No. 4, pp.393-399, November 30, 2018. https://doi.org/10.17703/JCCT.2018.4.4.393

[3] M. Abe, K. Nakagawa,“Cross-sectional Stock Price Prediction using Deep Learning for Actual Investment Management,”Proceedings of the 2020 Asia Service Sciences and Software Engineering Conference, May 2020, https://doi.org/10.1145/3399871.3399889

[4] G. Koch R. Zemel, R. Salakhutdinov, “Siamese Neural Networks for One-shot Image Recognition,” Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 2015.

[5] O. Vinyals, et al, “Matching Networks for One Shot Learning,” 30th Conference on Neural Information Processing Systems (NIPS 2016), 2016.

[6] J. Snell, K. Swersky, R.S. Zemel, “Jake Snell, Kevin Swersky, Richard S. Zemel,”Conference on Neural Information Processing Systems (NIPS 2017),2017.

[7] A. Dosovitskiy, et al, “Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks,” Conference on Neural Information Processing Systems (NIPS 2014), 2014.

[8] I. Misra, L. Maaten, “Self-Supervised Learning of Pretext-Invariant Representations,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[9] S. Gidaris, P. Singh, N. Komodakis, “Unsupervised Representation Learning by Predicting Image Rotations,” The proceedings of International Conference on Learning Representations(ICLR 2018), 2018.

[10]T. Chen, S. Kornblith, Mo. Norouzi, G. Hinton, “A Simple Framework for Contrastive Learning of Visual Representations,” The proceedings of

International Conference on Machine Learning (ICML 2020), 2020.
[11]K. He, et al, “Momentum Contrast for Unsupervised Visual Representation Learning,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
[12]B. Bozorgtabar. et al, “Informative sample generation using class aware generative adversarial networks for classification of chest Xrays,” Computer Vision and Image Understanding 184(1), 2019. https://doi.org/10.1016/j.cviu.2019.04.007
[13]H. Zhen, et al,, Biomedical “Image Segmentation via Representative Annotation,” Proceedings of the AAAI Conference on Artificial Intelligence , vol. 33, pp.5901-5908, 2019.
[14]S. Mo, et al, “Mining GOLD Samples for Conditional GANs,” Proceedings of the 33rd International Conference on Neural Information Processing Systems, pp.6170 - 6181, 2019.
[15]D. Mahapatra, et al, “Efficient Active Learning for Image Classification and Segmentation using a Sample Selection and Conditional Generative Adversarial Network,” MICCAI 2018. Lecture Notes in Computer Science, vol 11071. 2018.
[16]C. Mayer,, R Timofte., “Adversarial Sampling for Active Learning,” 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 2020.
[17]Y. Gal, R. Islam, N., Z. Ghahramani, “Deep Bayesian Active Learning with Image Data,” Proceedings of the 34th International Conference on Machine Learning, PMLR 70:1183-1192, 2017.
[18]M. Gorriz, et al, “Cost-Effective Active Learning for Melanoma Segmentation,” the proceedings of NIPS 2017 workshop, 2017.
[19]J. Dolz, et al, “Deep CNN ensembles and suggestive annotations for infant brain MRI segmentation,” Computerized Medical Imaging and Graphics, Vol 79, January 2020. https://doi.org/10.1016/j.compmedimag.2019.101660
[20]R. Shen, et al, “Breast mass detection from the digitized X-ray mammograms based on the combination of deep active learning and self-paced learning,” Future Generation Computer Systems, 2019. https://doi.org/10.1016/j.future.2019.07.013
[21]C. Dai, et al, “Suggestive Annotation of Brain Tumour Images with Gradient-guided Sampling,” The proceeding of MICCAI 2020, 2020, pp.156-165.
[22]L. Yang, et al, “Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation,” Lecture Notes in Computer Science, 2017. https://doi.org/10.1007/978-3-319-66179-7_46
[23]Y. Geifman, and R.Elyaniv, “Deep Active Learning over the Long Tail,” arXiv: Learning, 2017.
[24]C. Yin, et al, “Deep Similarity-Based Batch Mode Active Learning with Exploration-Exploitation,” The proceeding of 2017 IEEE International Conference on Data Mining (ICDM), pp.575 - 584, 2017.
[25]J., T. Ash, et al, “Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds,” The proceedings of ICLR 2020 Conference, 2020.
[26]X. Du, ,et al, “Building an Active Palmprint Recognition System,” The proceedings of The International Conference on Image Processing, ICIP 2019, Taipei, Taiwan, pp.1685 - 1689, 2019.
[27]D. Gissin, S. Shalevshwartz., “Discriminative Active Learning,” arXiv: Learning, 2018.
[28]A. Kirsch, et al, “BatchBALD: Efficient and Diverse Batch Acquisition for Deep Bayesian Active Learning,” Proceedings of the 33rd International Conference on Neural Information Processing Systems, pp.7026 - 7037, 2019.
[29]F. Zhdanov, “Diverse mini-batch Active Learning,” arXiv: Learning, 2019.
[30]R. Pinsler, et al, “Bayesian Batch Active Learning as Sparse Subset Approximation,” Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019.
[31] S. Han, “Analysis of Suggestive Learning in Deep Learning,” The papers of Korea National University of Transportation, 2021.

※ 이 논문은 2021년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2021R1F1A1049467)