# 메타정보 기반 증식 데이터 검증 기술 연구

박종빈*, 정종진, 김경원
*한국전자기술연구원
*jpark@keti.re.kr

# A Study on the Augmentation Data Validation

Jongbin Park*, Jong-Jin Jung, Kyung-Won Kim
*Korea Electronics Technology Institute

요 약

본 논문은 기계학습 및 데이터 분석을 위한 증식데이터의 검증 방법에 관한 것이다. 텍스트나 이미지와 같은 서로 다른 형식의 데이터 간 비교를 지원하는 것이 주요 특징이다. 증식된 데이터의 검증은 태그 정보를 활용하거나 사전에 수행된 결합 임베딩 메타정보를 기반으로 수행한다. 본 논문에서는 제안 기능들을 포함하는 통합 소프트웨어를 구현했으며, 이를 이용하여 임계 범위를 벗어나는 아웃라이어를 제거하거나, 소정의 조건에 부합하는 데이터를 추출하여 새로운 기계학습용 데이터 셋을 구성할 수 있도록 했다.

## Ⅰ. 서 론

본 논문은 증식된 데이터의 품질을 개선하거나 소정의 데이터 분석 환경에 맞게 데이터를 가공하는 방법에 관한 것이다. 통상적으로 데이터 증식은 인공지능 모델 학습이나 데이터 분석 과정에서 부족한 데이터를 보충하거나, 불균일한 데이터 분포를 보상하기 위해 사용한다. 그러나 증식된 데이터가 오히려 아웃라이어(Outlier)로 기능하거나 학습 성능을 저해할 가능성도 존재한다.

본 논문은 이러한 문제에 대응하기 위한 것으로써, 증식된 데이터 상호간 특징을 분석하여 태그 정보나 임베딩 벡터와 같은 메타정보를 생성하고, 이를 통해 데이터 정제와 검증, 그리고 새로운 데이터 셋 구성을 수행하는 것을 주요 목적으로 한다.

## Ⅱ. 본론

### 1. 데이터 증식

데이터 증식을 위해서는 표 1과 같이 먼저 시드 데이터 $S$를 입력받는다. 이후 시드 데이터 $S$에서 이미지와 텍스트와 같은 부분 데이터를 추출한다. 추출된 부분 데이터를 기반으로 변이형 오토인코더 (VAE, Variational Auto Encoder)나, GAN(Generative Adversarial Network) 기술을 활용하여 새로운 데이터를 생성한다. 여기서 시드 데이터의 일례로는 비디오 파일이나 웹페이지가 가능하다.

표 1. 데이터 증식 과정

- 입력 시드 데이터 : $S$
- 추출된 부분 데이터 집합 :
  $\{A_1, A_2, \dots, A_n\}$, $\{n : n \in \mathbb{N}\}$ ($n$은 자연수)
- 증식 데이터 집합 :

$$\boldsymbol{A}=\begin{Bmatrix} A_{11}\ A_{12} \dots A_{1m} \\ A_{21}\ A_{22} \dots A_{2m} \\ \dots\ \dots\ \dots\ \dots \\ A_{n1}\ A_{n2} \dots A_{nm} \end{Bmatrix}$$

, 상기 식에서 $\{n, m : n \in \mathbb{N}, m \in \mathbb{N}\}$ ($n$, $m$은 자연수)이며 $A_{nm}$은 증식된 데이터 원소로써, 집합 $\{A_1, A_2, \dots, A_n\}$에 속하는 원소 $A_n$을 $m$번째로 생성한 것을 의미한다. 식에서는 간결한 표현을 위해 각각의 $n$개 원소들에 대해 동일하게 $m$개의 증식이 이뤄지는 것으로 나타냈으나 원소별로 생성되는 데이터의 개수는 달라질 수 있다.

### 2. 이종 증식 데이터 검증의 어려움

본 논문에서 데이터에 대한 검증은 증식된 데이터가 소정의 학습과정이나 데이터 분석 과정에 효과적으로 활용할 수 있도록 하기 위한 처리 과정으로 한정한다. 이를 위해 증식 데이터의 정량적 특징을 얻어오거나, 증식 데이터 상호간 유사도의 계산이 필요하다. 표 1에 표시한 바에 따르면 $A_i$에 대해서 $A_{ij}$와의 관련성을 계산하는 과정에 해당한다. 그러면 이 값을 기준으로 사용자가 설정한 임계값을 기준으로 수락(Acceptance)이나 거절(Rejection), 계층적 군집화(Hierarchical Clustering)를 기반으로 유사 데이터 그룹을 부분적으로 추출할 수 있게 된다.

그러나 데이터 상호간 비교는 다양한 현실적인 문제들을 가지고 있다. 일례로 미디어 타입이 이미지로 동일할지라도 가로와 세로 크기가 다르거나, 색을 표현하는 채널수, 색 표현 공간 등이 다르다면 온전한 비교를 할 수 없다. 그리고 MSE (Mean Squared Error) 혹은 PSNR (Peak to Signal to Noise Ratio)과 같이 영상에만 적용할 수 있는 유사도 메트릭(Metric)을 이종 형식을 갖는 데이터 셋에 적용할 수는 없다.

3. 태깅 정보 및 결합 임베딩 기반 처리

이러한 문제를 해결하고자 본 논문에서는 태깅과 결합 임베딩(Joint Embedding)[1] 기법을 핵심요소로 사용한다. 태그 정보는 HTML페이지, 이미지, 텍스트, 오디오 등 다양한 형식의 미디어 타입을 상위 수준에서 기술하는데 유용하다. 이렇게 기술된 태그 정보는 기본적으로 단어이므로 LSA(Latent Semantic Analysis)[2]나 LDA(Latent Dirichlet Allocation)[3]와 같은 주제모델링(Topic Modeling)기법을 적용하거나, 태그들에 대해서 워드임베딩(Word Embedding)을 수행함으로써 태그 상호 간 유사도 연산을 수행할 수 있다. 워드임베딩과 유사하게 영상이나 오디오에 대해서도 고차원 벡터 공간상의 하나의 벡터로 매핑하는 임베딩 처리가 가능하다. 그러나 이러한 임베딩 과정은 동일한 형식의 미디어 타입에 적용하는 것이 통상적이었다. 이를 이종의 데이터 형식들을 포괄하는 데이터 셋에도 확장하는 방법이 결합 임베딩[1]으로써, 본 논문에서는 이를 증식된 데이터들에 적용한다.
이런 과정을 통해 임계 범위에 해당하는 데이터의 선택, 대부분의 데이터 셋과는 특징이 현저하게 다른 아웃라이어 검출, 유사도 기반 계층적 분류와 같은 처리가 가능하다.

4. 개발 내용

본 논문에서 개발한 증식된 데이터에 대한 검증 도구는 그림 1과 같이 파이썬 언어를 사용하여 독립적인 소프트웨어로 구현하였다. 구체적인 기능으로는 시드 데이터와 증식된 데이터 간 비교, 증식된 데이터 상호 간 비교를 수행한다.
입력데이터가 영상 형식으로 동일한 경우에 대응하여 PSNR(Peak Signal to Noise Ratio), MSE(Mean Square Error), NRMSE(Normalized Root MSE, SSIM(Structural Similarity Index Measure), JS(The Jensen-Shannon distance between p and q), KLD(Kullback Leibler Divergence)와 같은 저수준 메트릭을 제공한다. 입력데이터가 텍스트 형식으로 동일한 경우에 대응하여 LSA, LDA, 워드임베딩 기법을 제공한다. 입력데이터가 텍스트와 이미지 형식이 혼합된 경우이면서 태그 정보가 존재하는 경우에는 태그 정보를 워드임베딩 처리하여 유사도 비교를 수행한다. 입력데이터가 텍스트와 이미지 형식이 혼합된 경우이면서 태그 정보가 존재하지 않는 경우에는 각각의 증식 데이터에 대해서 결합 임베딩을 수행하여 임베딩 벡터를 메타정보로 추가한다. 이를 위해 구글의 Inception 딥러닝 구조[4]를 수정하여 적용했다.
계산된 유사도 결과를 기초로 조건에 해당하는 데이터만을 추출할 수 있으며 추출된 데이터 집합은 기계학습 및 분석을 위한 데이터 셋으로 활용할 수 있다. 계산된 메타정보와 상호 유사도 값은 CSV(Comma Separated Values) 형식으로 내보내기 할 수 있으며, 사전에 연결된 데이터베이스 시스템에 연계 등록할 수 있도록 했다.

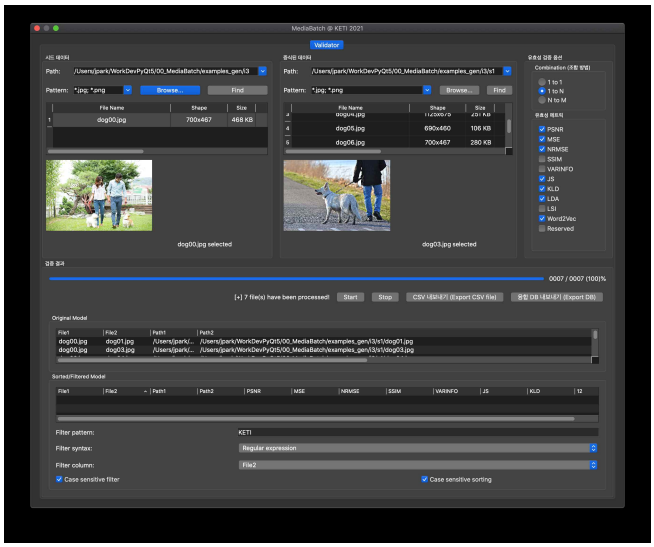

그림 1. 개발한 데이터 검증 도구

## Ⅲ. 결론

본 논문에서는 기계학습 및 데이터 분석을 위한 증식 데이터의 검증 기술을 소개했다. 텍스트와 이미지와 같은 서로 다른 형식을 갖는 데이터 간 상호 비교를 지원하는 것이 주요 특징이며, 이를 위해 태깅 정보와 사전에 생성한 결합 임베딩 정보를 사용했다. 비교 및 검증을 위한 소프트웨어를 개발했으며, 이를 통해 임계 범위를 벗어나는 아웃라이어를 제거하거나, 소정의 조건에 부합하는 데이터 셋을 새롭게 구성할 수 있도록 했다. 이는 기계학습을 위한 학습데이터 확보 단계나 데이터 분석 과정에서 활용이 기대된다.

## ACKNOWLEDGMENT

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No. 2020-0-0062, 이종 정보 활용 및 데이터 융합을 통한 데이터 증식 기술 개발)

## 참 고 문 헌

[1] Liwei Wang, Yin Li, Svetlana Lazebnik, "Learning deep structure-preserving image-text embeddings", Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5005-5013, 2016

[2] Deerwester, Scott C., et al. "Indexing by latent semantic analysis." JAsIs 41.6 pp.391-407, 1990

[3] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." the Journal of machine Learning research 3, pp. 993-1022, 2003

[4] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, Zbigniew Wojna, , "Rethinking the inception architecture for computer vision, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818--2826, 2016