Combining Active Learning and Data Augmentation for Image Classification

Ma, Y., Lu, S., Xu, E., Yu, T., & Zhou, L. (2020, September). Combining active learning and data augmentation for image classification. In *Proceedings of the 3rd International Conference on Big Data Technologies* (pp. 58-62).

김지원

jeewonkim@g.seoultech.ac.kr

IAI
Industrial Artificial Intelligence Lab

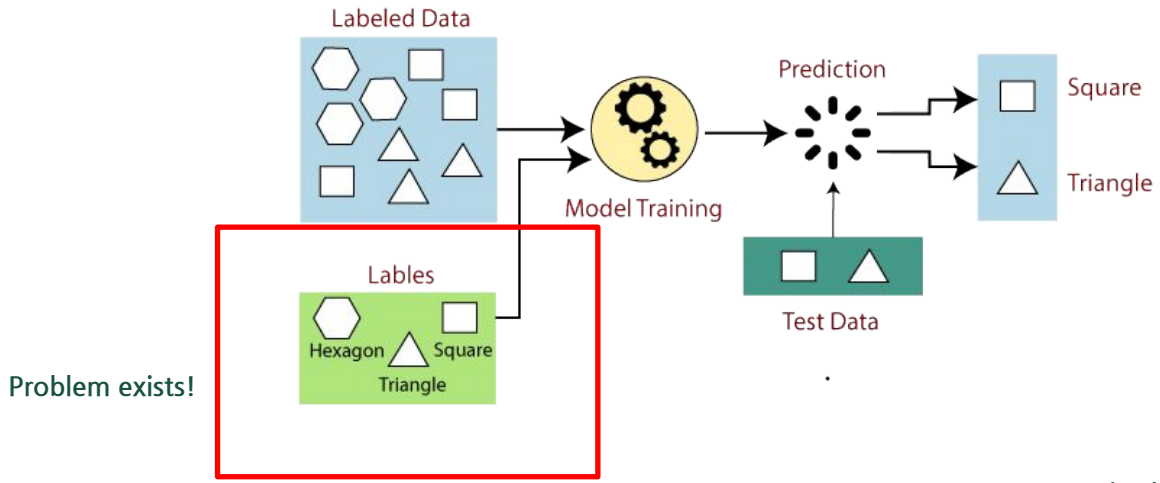
2023.03.0

Contents

Contents

Introduction

Introduction



Supervised Learning

1. Supervised Learning

- Image Classification에 많은 기여를 함
- Satisfactory Result를 얻어내기 위해선 많은 양의 Labeled Training Data 필요
- 문제점 : 실생활에선 Labeled Data를 구하기 어려울뿐더러, 다량의 Data에 대한 Manual annotation은 높은 Cost가 됨

2. Active Learning

- Unlabeled Data를 활용하여 Classification Model을 효율적으로 훈련할 수 있음

Introduction

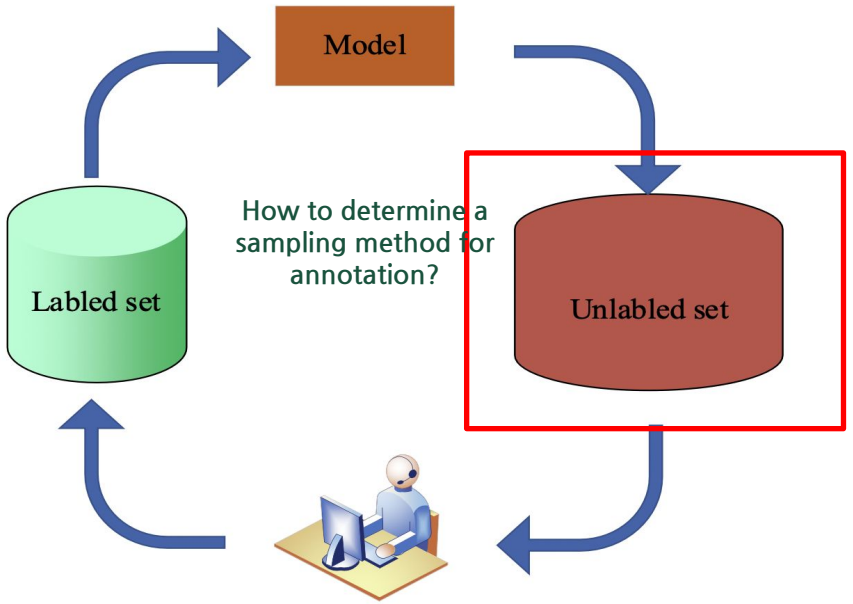


Figure 1. The specific process of active learning

1. Active Learning : Unlabeled Data를 효율적으로 활용하는 방법

- Iterative Process : Model Training, Model Testing, Sample Selection, Data Annotation이 반복 (Figure 1)
- Sampling Strategies이 핵심이며, 주로 Uncertainty Algorithm을 채택

2. Uncertainty Algorithm

- **문제점** : Current Classifier에 대하여 어떤 Samples이 가장 Uncertain한지 결정하며 Unlabeled Sample의 정보를 무시 => 다량의 Unlabeled Sample이 있는 경우 매우 비효율적
- **(해결책1)**Xin Li et al. : Candidate Sample과 Unlabeled Sample의 상호 정보를 측정할 수 있도록 Uncertainty Algorithm과 Information Density정보를 결합하여 좋은 결과를 냄(Novel Adaptive Active Learning)
- **(해결책2)**Ksenia K et al. : Unlabeled Data의 expected errors 예측 (Data-Driven Active Learning)
- 위의 해결책에서는 Initial Training Set(Labeled Data)의 양이 작다는 것을 고려하지 않음 => Low Accuracy 초래

IAI
Industrial Artificial Intelligence Lab

Data Augmentation

Data Augmentation

Data Augmentation

- 위의 문제를 해결하기 위해 **'Active Learning + Data Augmentation'** 방법이 제시
- 장점 : Labeled Data의 양을 늘릴 수 있음, 적은 양의 Labeled Data의 충분한 사용 가능, Classifier의 나은 결과

(방법1) Flip Augmentation

- 같은 Category에 속하게 하며, 다른 Category의 다른 Sample 간의 존재하는 정보를 모델링하지 않음

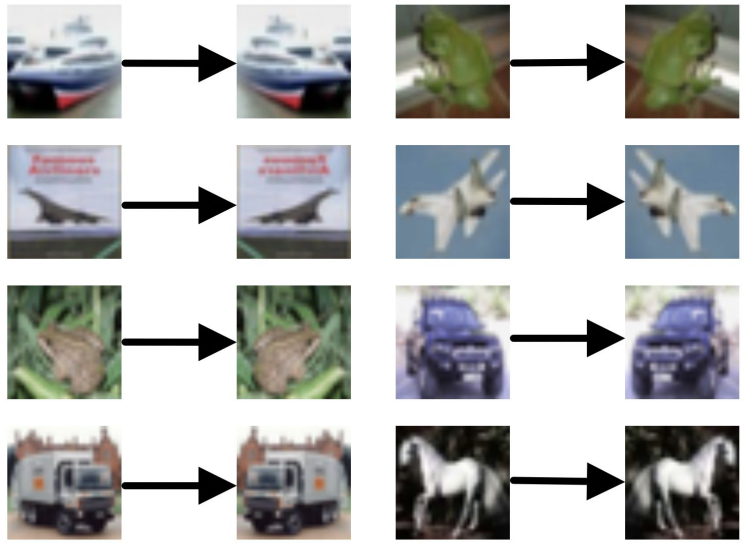


Figure 2. Flip augmentation examples.

Data Augmentation

(방법2) Mixup

- Linear Interpolation을 통해 Training Sample과 해당 Label을 구성하는 방법
- 무작위로 Training Dataset에서 두 개의 Sample을 고른 후 해당 Sample 쌍의 특징 벡터에 대해 Linear Interpolation 수행
- Mixup 이후 두 Sample의 Label은 새 Sample의 범주 확률 분포
- 장점 : 사전 지식에 의해 Training Data의 분포를 확장하며 새로운 Sample-Label 쌍을 얻음

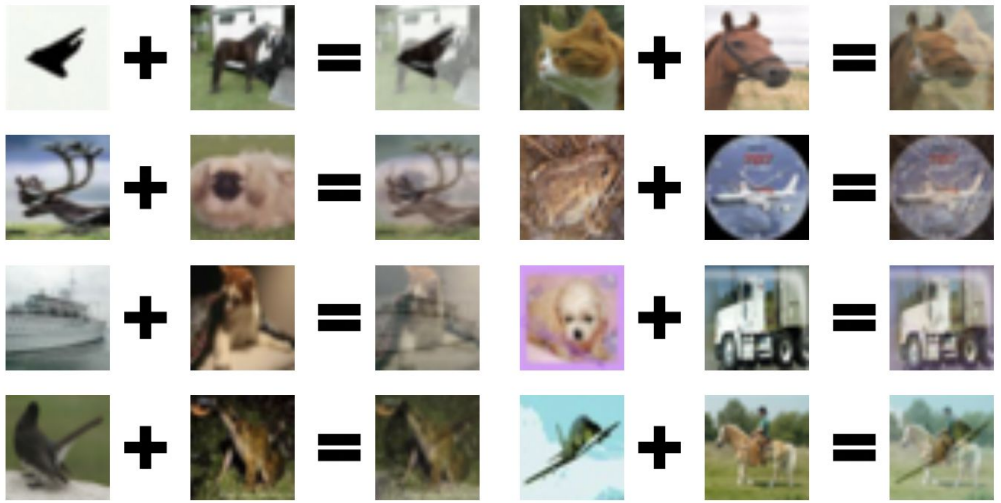


Figure 3. Mixup augmentation examples.

$$\lambda \sim Beta(\alpha, \alpha)$$

$$\tilde{x} = \lambda x_i + (1-\lambda) x_j$$

$$\tilde{y} = \lambda y_i + (1-\lambda) y_j$$

두개의 Random Sample : (xi,yi), (xj,yj)
X : original feature vector of sample
Y : one-hot label encoding
Lambda : beta distribution을 따르는 parameter
Alpha : hyperparameter

$$\alpha \in [0, +\infty)$$

Sampling Strategy

Sampling Strategy

Uncertainty Sampling이 Sampling Strategy 으로 선택됨

(방법1) Random Sampling

- Unlabeled Seet에서 랜덤하게 선택됨

(방법2) Uncertainty Sampling

- Information Entropy를 계산하여 Uncertainty가 큰 Sample을 찾아냄
- 각 Sample에 대한 Entropy 계산 후, Entropy가 내림차순으로 정렬
- 큰 Entropy가 선택됨

$$H(x) = -\sum_{i=1}^{m} p(y_i \mid x, L) \log_2 p(y_i \mid x, L)$$

X : A sample in the unlabeled set
M : The number of categories
Yi : i-th category
P(yi|x,L) : x가 i-th category에 속할 확률

Method

Method

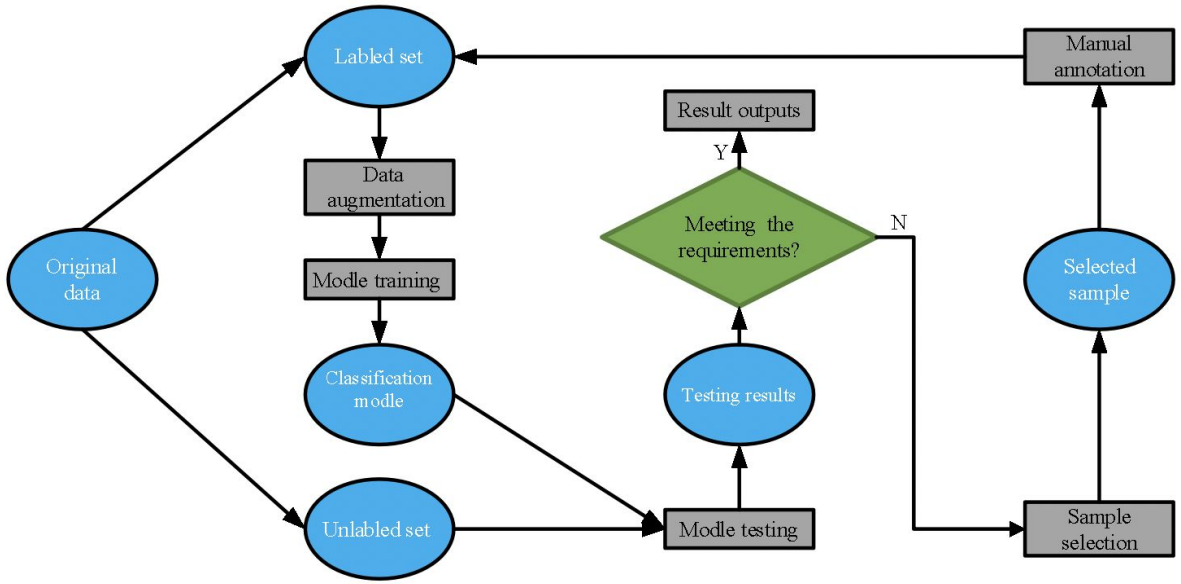


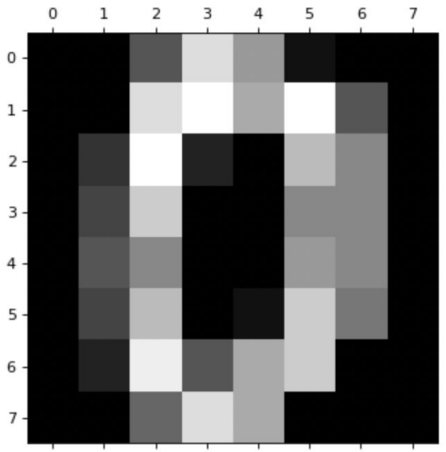
Figure 4. Method flowchart.

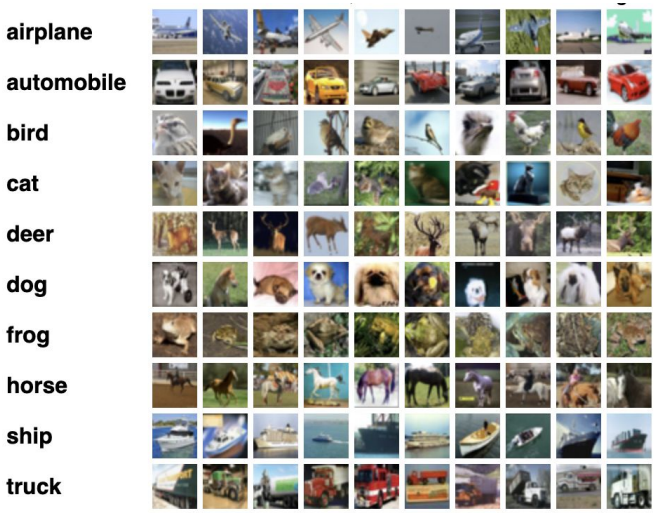
1. Origical Data을 Labeled/Unlabeled로 나눔
2. Labeled Data에 한해 Augmentation 수행
3. Model Training with Labeled Data
4. Unlabeled Data으로 Model Test
5. Sampling Strategy를 사용하여 Sample 선택
6. 선택된 Sample의 Annotation 후 Labeled Data 업데이트
7. Stop Requirement에 도달할때까지 반복

Experiments and Results

Experiments and Results

Datasets and Model Selection





Digits Dataset in the Scikit-Learn Library

- 8*8
- 1797 images
- Logistic Regression Model

Cifar10

- 32*32
- 60,000 images
- Resnet20

The number of initial dataset은 Unlabeled Data의 수보다 작거나 같을 예정

IAI
Industrial Artificial Intelligence Lab

Experiments and Results

Discussion of Results

Table 1. Classification accuracy on the digits dataset (%)

Ratio of labeled set	None		Flip		Mixup	
	Random	Uncertainty	Random	Uncertainty	Random	Uncertainty
0.1	85.0	89.9	90.2	91.8	92.2	92.4
0.2	87.3	91.3	91.2	92.4	92.1	92.6
0.3	91.4	91.9	92.1	92.8	92.5	93.0
0.4	91.3	92.5	92.5	93.2	93.1	93.6
0.5	92.4	93.0	93.2	93.7	93.6	93.8

Table 2. Classification accuracy on the Cifar10 set (%)

Ratio of labeled set	None		Flip		Mixup	
	Random	Uncertainty	Random	Uncertainty	Random	Uncertainty
0.1	67.2	68.2	68.3	69.0	68.9	68.9
0.2	74.5	74.7	76.2	74.8	76.8	75.5
0.3	78.4	79.8	80.0	80.6	80.6	81.2
0.4	80.7	81.3	82.6	82.8	81.4	83.3
0.5	82.3	83.7	83.6	84.5	83.9	84.7

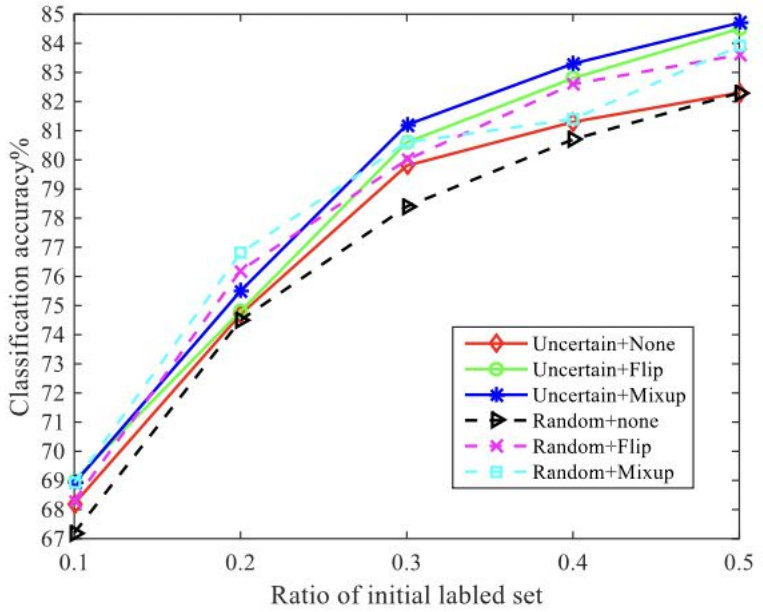


Figure 5. A comparison of classification accuracy on the Cifar10 set.

Augmentation을 사용했을 때 성능이 더 좋음

Random sampling보단 Uncertainty Sampling에서 더 나은 성능

IAI
Industrial Artificial Intelligence Lab

Conclusion

Conclusion

Conclusion

- Data Augmentation을 통해 Labeled Data의 수를 늘릴 수 있었고, 결론적으로 더 나은 결과(정확도)를 얻을 수 있었음