**사물데이터분류 경진대회**

1. 이미지 확인
2. 데이터 전처리

* Data augmentation으로 추가 데이터 생성

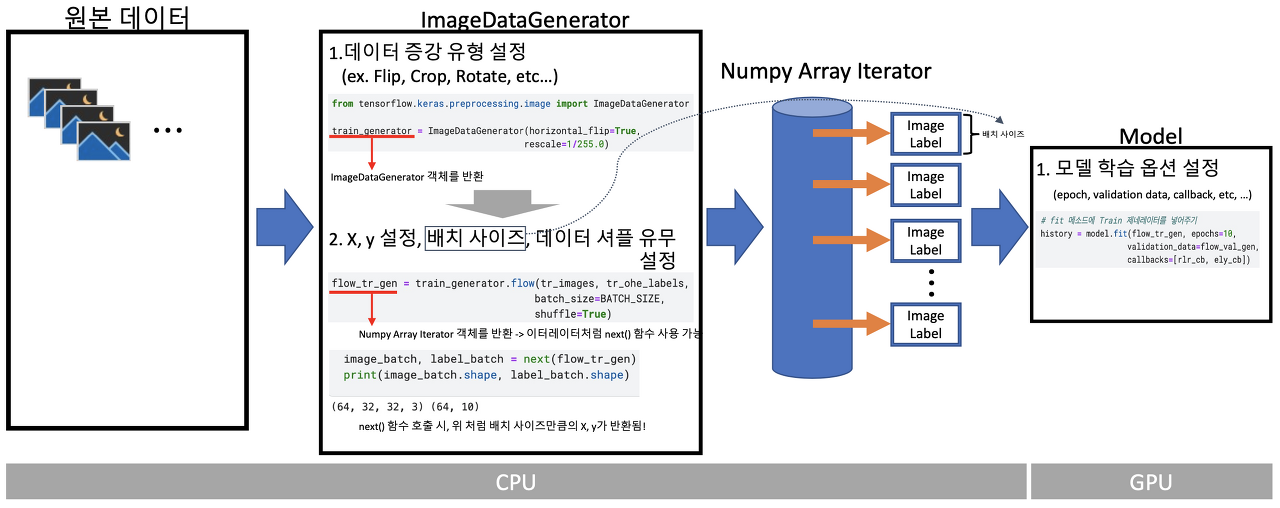
Data set 개수를 늘려서 overfitting을 해결하기 위해 augmentation 진행.

Keras의 ImageDataGenerator 함수 활용하여 augmentation함.

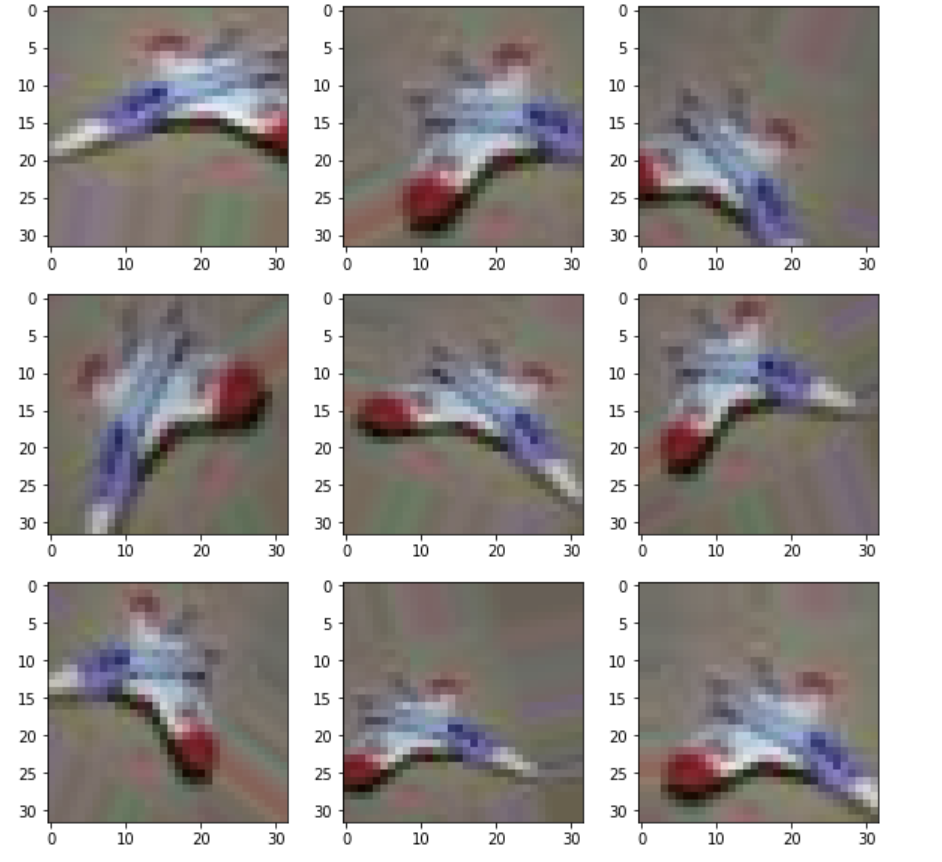
텍스트이(가) 표시된 사진

자동 생성된 설명

* rotation\_range : 이미지를 랜덤한 각도로 돌리는 정도
* brightness\_range : 이미지의 밝기를 랜덤하게 다르게 주는 정도
* zoom\_range : 사진을 확 대하는 정도
* width\_shift\_range : 사진을 왼쪽 오른쪽으로 움직이는 정도
* height\_shift\_range : 사진을 위 아래로 움직이는 정도
* horizontal\_flip : y축을 기준으로 반전 (오른쪽 왼쪽 뒤집기)
* vertical\_flip : x축을 기준으로 반전 (위 아래 뒤집기)



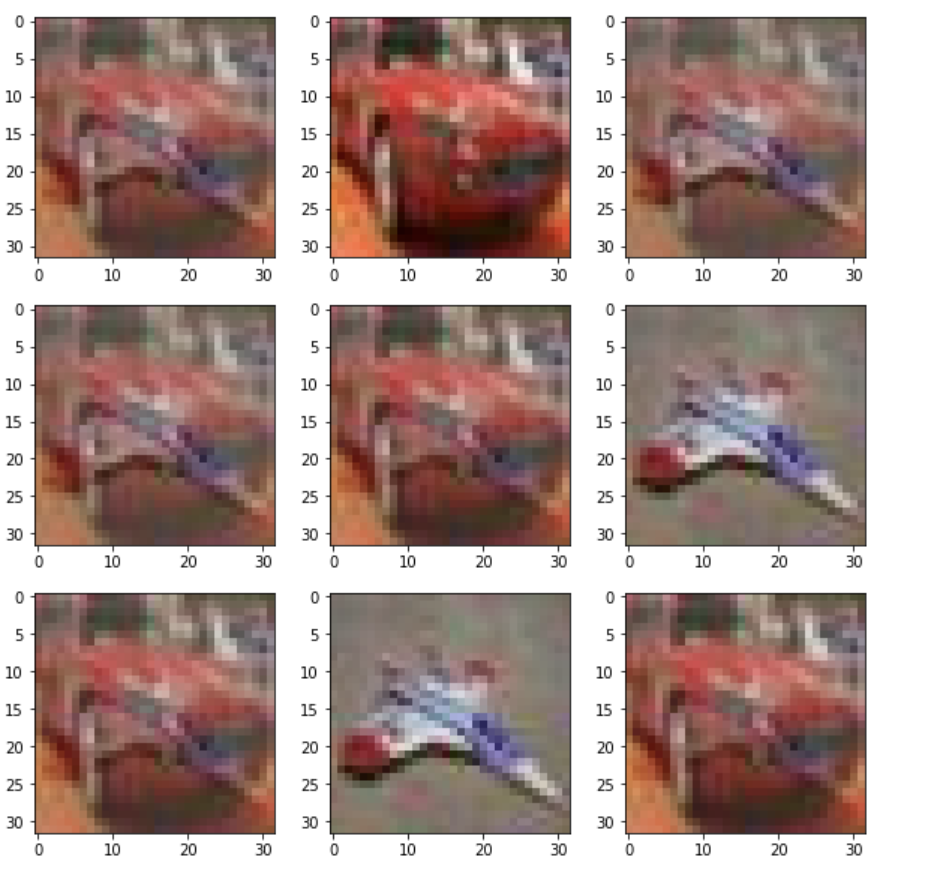
<ImageDataGenerator도식>



<data augmentation 예시>

* Mixup augmentation으로 추가 데이터 생성

Mixup의 경우 두 데이터를 람다라는 파라미터를 활용하여 적절하게 섞은 후 모델의 학습에 활용하는 방법으로, ERM에 비하여 클래스 분류 시에 더 좋은 효과를 보이는 것으로 알려져 있습니다. Mixup 알고리즘은 라벨이 잘못 분류된 라벨 오류에도 더 낮은 테스트 에러를 보이기 때문에, 이 특성을 활용할 수 있는 Multi-label 문제에 적합합니다.



<mixup augmentation 예시>

1. Train set, valid\_set으로 데이터셋 나누기( stratify 추가하여 label에 데이터 골고루 분포되도록)
2. CNN 모델 구현
3. 모델 학습
4. 제출

**주차수요예측경진대회**

1. 데이터 업로드
2. 데이터 탐색(타입, 결측치 확인)
3. 전처리
   1. 결측치 대체

결측 부분과 유사한 데이터를 찾아 대체(자격유형)

실제 임대 모집 공고를 확인하여 평균값으로 대체

나머지 결측 데이터 0으로 대체

* 1. 오류 데이터 제거
  2. 파생변수 생성

지역, 공급유형의 피처 수가 비교적 많아 이를 n개의 그룹으로 통합

* + 1. 지역 : 등록 차량수의 중앙값을 기준으로 지역을 5개 그룹으로 통합

텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트, 영수증, 스크린샷이(가) 표시된 사진

자동 생성된 설명

* + 1. 공급유형 : ‘등록차량수 - 단지내주차면수’ 차이를 기준으로 공급유형을 5개 그룹으로 통합

텍스트이(가) 표시된 사진

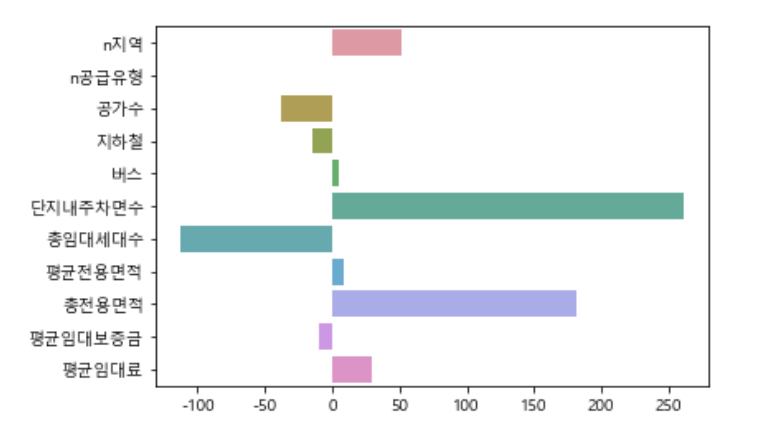
자동 생성된 설명

* + 1. 전용면적 : 5단위로 변환
  1. 다중공선성을 고려하여 기존 피처 일부를 제외하고 새 피처 생성.

텍스트이(가) 표시된 사진

자동 생성된 설명

1. 모형 학습
   1. 피처와 타겟 간 상관 확인
   2. 초기모형 학습(linear regression)
   3. 회귀계수 확인



* 1. Permutation importance를 통한 피처 중요도 검토
     + Permutation importance

모델 예측에 가장 큰 영향을 미치는 feature를 파악하는 방법 중 하나로 특정 컬럼의 데이터의 순서를 뒤바꿔 노이즈를 준 다음 스코어를 확인하는 방법이다. 특성 중요도의 다른 방법인 Drop- Column과는 달리 재학습이 필요없다는 장점이 있다.

sklearn의 permutation\_importance 이용함.

텍스트이(가) 표시된 사진

자동 생성된 설명

* 1. 다항회귀 적용

텍스트이(가) 표시된 사진

자동 생성된 설명

1. 모형 개선
   1. 가중치 생성

Cook’s distance를 활용하여 가중치 부여

* + - Cook’s D : 각 관측값이 전체 최소제곱추정량에 미치는 영향력을 보여주는 통계적 지표. 값이 클수록 회귀모형에 큰 영향을 주는 샘플이다.

Cook’s D 값이 큰 샘플에 작은 가중치, Cook’s D 값이 작은 샘플에 큰 가중치를 부여하여 이상치 통제하고 잘 예측하지 못하는 샘플을 강조하는 용도로 활용.

* 1. Stratified KFold 사용

1. 제출