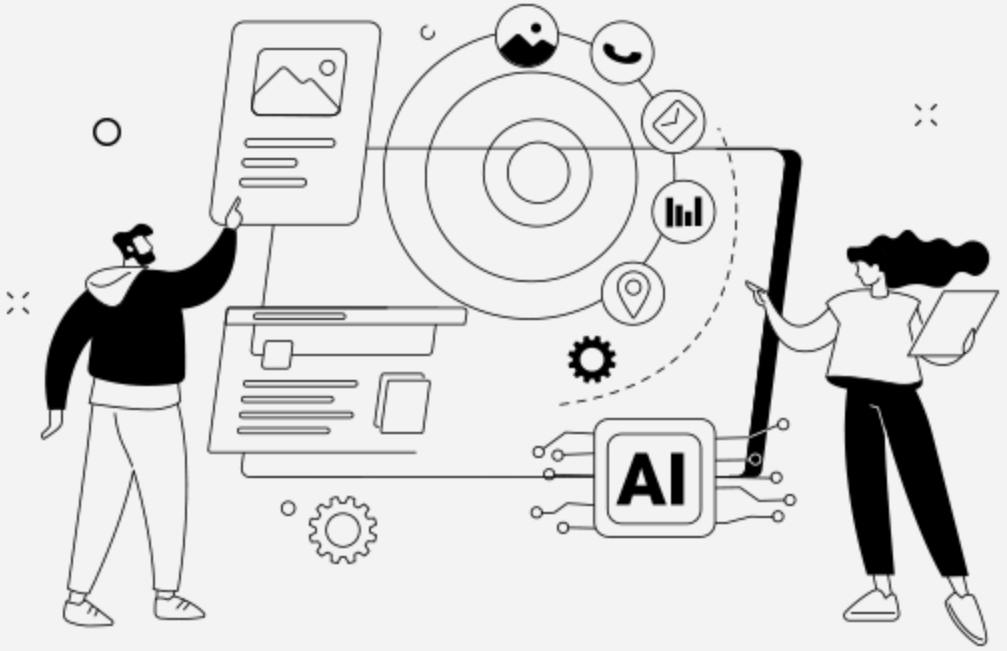
2022 데이터 크리에이터 캠프

Data Creator Camp



덕영고등학교 - 춘식이

과학기술정보통신부 NIA 한국지능정보사회진흥원

Mission 1

1-1. Training 데이터 셋의 데이터를 살펴보고 라벨 종류는 무엇이 있고, 각 라벨의 개수를 구하시오.

Answer:
라벨 종류는 cap_and_hat, outerwear, tops, bottoms, shoes로 총 5개이고,
라벨별 이미지의 개수는 다음과 같다.

cap_and_hat: 196개,
outerwear: 4606개,
tops: 18350개,
bottoms: 6424개,
shoes: 424개

Mission 2

2-1. 이미지 크기를 적절히 조절하거나, 해상도를 조절하여 학습 데이터 셋을 구축하시오.

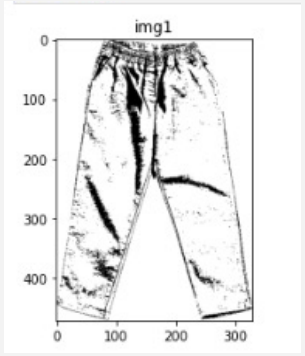
```python
def crop(img_num):
    file = f"{BASE_DIR}/dataset/Item-Image/img{img_num}.jpg"
    img1 = cv2.imread(file)

    # Image Shape
    h, w = img1.shape[:2]
    h1, h2 = int(h * 0.2), int(h * 0.7)
    w1, w2 = int(w * 0.05), int(w * 0.95)
    img = img1[h1: h2, w1: w2]

    # Resolution
    img = cv2.resize(img, None, fx=0.75, fy=0.75, interpolation=cv2.INTER_AREA)

    # Gray Scale
    gray = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)

    # Adaptive Threshold
    img2 = cv2.adaptiveThreshold(gray, 255, cv2.ADAPTIVE_THRESH_GAUSSIAN_C, cv2.THRESH_BINARY, 15, 2)

    #Blur
    blur = cv2.medianBlur(img2, 5, dst=None)

    # Add Kernel
    kernel = np.ones((3,3),np.uint8)

    # Morphology
    closing = cv2.morphologyEx(blur, cv2.MORPH_CLOSE, kernel)
    gradient = cv2.morphologyEx(closing, cv2.MORPH_GRADIENT, kernel)

    # 경계값 도출
    contours = cv2.findContours(gradient, cv2.RETR_TREE, cv2.CHAIN_APPROX_TC89_L1)[0]

    x1 = [] #x-min
    y1 = [] #y-min
    x2 = [] #x-max
    y2 = [] #y-max
    for i in range(1, len(contours)):
        ret = cv2.boundingRect(contours[i])
        x1.append(ret[0])
        y1.append(ret[1])
        x2.append(ret[0] + ret[2])
        y2.append(ret[1] + ret[3])

    x1_min = min(x1)
    y1_min = min(y1)
    x2_max = max(x2)
    y2_max = max(y2)
    cv2.rectangle(gradient, (x1_min, y1_min), (x2_max, y2_max), (0, 255, 0), 3)

    origial_img = gray[y1_min:y2_max, x1_min:x2_max]
    crop_img = img2[y1_min:y2_max, x1_min:x2_max]

    crop_img = cv2.resize(crop_img, (32, 32))

    #img_merge = np.hstack((origial_img, crop_img))

    #return img_merge

    return crop_img
```

1. 이미지의 위, 아래, 양 옆을 강제로 제거한다.
2. 이미지의 해상도를 낮추고, Gray Scale로 전환한다.
3. opencv의 AdaptiveThreshold를 이용해 이미지 이진화를 한다.
4. 노이즈를 낮추기 위해 1차적으로 Blur 처리를 한다.
5. 2차적으로 커널을 생성하여 Morphology 연산 중
   closing 연산과 gradient 연산을 거친다.
6. findContours 함수를 이용해 경계값을 도출한 후
   그 경계값으로 사각형을 생성해 이미지 trimming 좌표를 생성한다.
7. 적응형 이진화 된 이미지를 6번에서 구한 좌표로 자르고
   32 x 32 사이즈로 Resize 후 Return 한다.

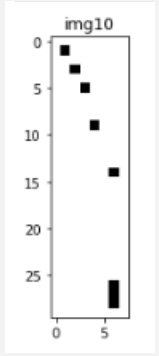
전처리 과정

전처리 코드를 완성하기 위해 많은 시행착오를 거쳤습니다.

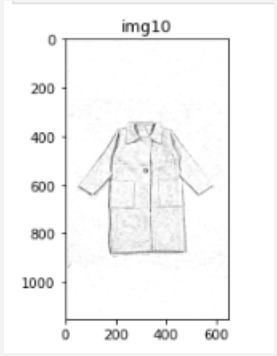
이미지를 이진화 후 모서리를 찾아내기 위해 Threshold 라는 함수를 사용했습니다.
하지만 배경이 흰색인 탓에 흰 옷 이미지 이진화 시 문제가 생기는 것을 확인할 수 있었고,
그래서 Adaptivethreshold 라는 적응형 이진화 함수를 사용하였습니다.
두 함수의 차이는 지역 이진화와 적응형 이진화입니다. 지역 이진화는 불균형한 조명에서 유리하지만
제공된 데이터셋은 그렇지 않기 때문에 적응형 이진화의 결과가 더 좋았던 것 같습니다.



Threshold 함수 사용
(갈색 옷)



Threshold 함수 사용
(흰 옷)

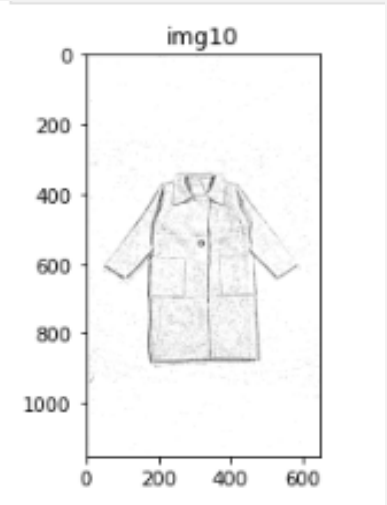


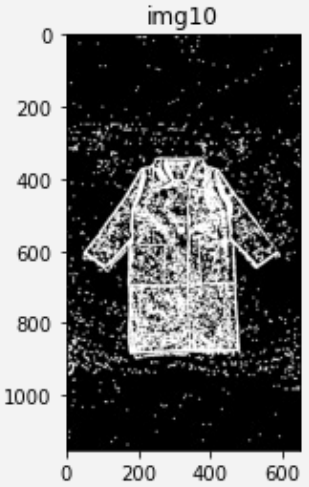
AdaptiveThreshold 함수 사용
(흰 옷)

전처리 과정

다만 이미지가 제대로 crop 되지 않은 것을 확인할 수 있습니다.
이미지를 확인해보니 노이즈가 많아
경계값을 제대로 도출하지 못한 것으로 보입니다.

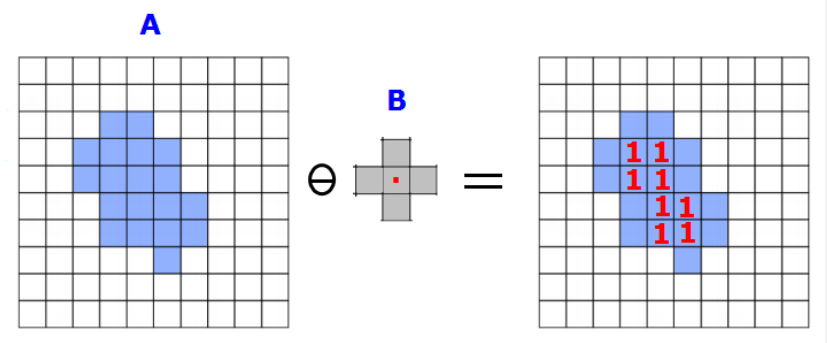
그래서 이미지의 노이즈를 줄이는 방법을 생각해보게 되었습니다.



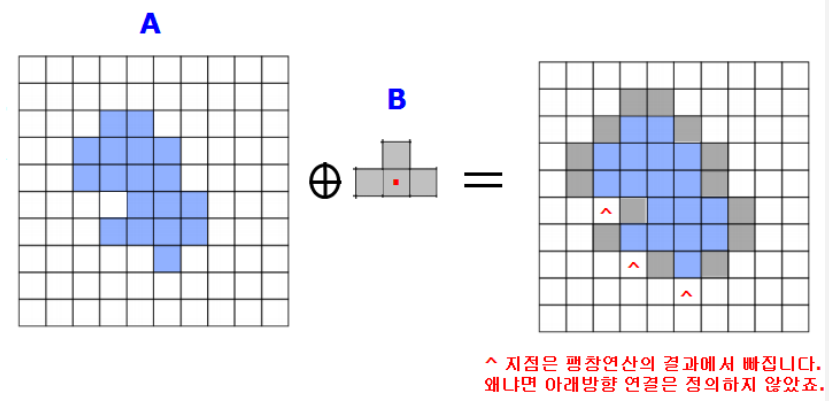


전처리 과정

처음은 Morphology 연산으로 시작했습니다.
Morphology 연산이란 '형태학' 이라는 뜻인데,
사진, 영상 분야에서의 노이즈 제거 등에 쓰이는 형태학적 연산을 말합니다.
대표적인 연산으로는 침식 연산과 팽창 연산이 있습니다.



침식은 말 그대로 이미지를 깎아 냅니다.
이미지를 0과 1로 전환하고, 커널을 생성하여
이 커널 안에 들어오지 못하는 이미지는 삭제해버리는 것입니다.



팽창은 침식의 반대입니다.
커널에 픽셀이 걸치기만 해도 1로 바꿔버리는 것입니다.

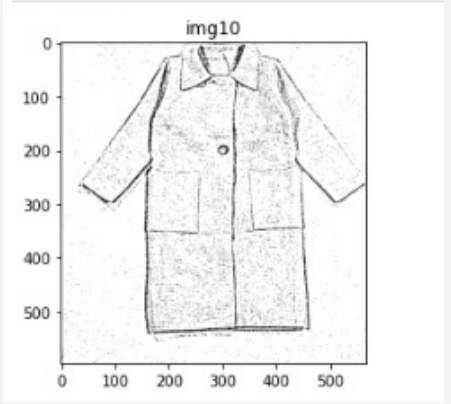
이미지 출처: https://bkshin.tistory.com

전처리 과정

저희는 침식 연산과 팽창 연결을 사용하는 닫힘 연산과 그레디언트 연산을 사용하였습니다.

닫힘 연산은 팽창 연산 후 침식 연산을 적용하여 주변보다 어두운 노이즈를 제거하는데 효과적입니다.

그레디언트 연산이란
팽창 연산을 적용한 이미지에서 침식 연산을 적용한 이미지를 빼면 경계 픽셀만 얻게 되는데
이 연산을 그레디언트 연산이라고 합니다.

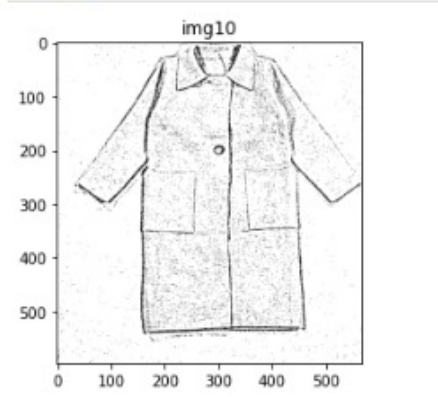


저희는 닫힘 연산을 사용하여 1차적으로 노이즈를 없애고
그레디언트 연산을 사용하여 2차적으로 노이즈를 없애며 옷의 경계를 뚜렷하게 했습니다.

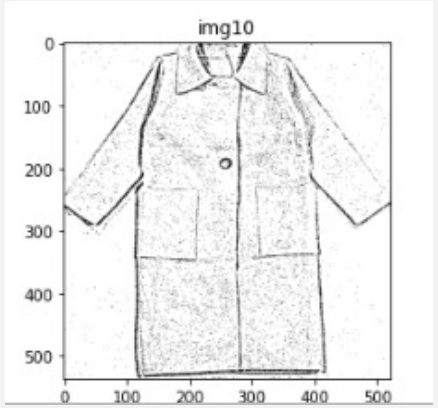
전처리 과정

처음보단 많이 좋아졌지만 아직 여백이 있는 걸 확인할 수 있습니다.

그래서 저희는 이미지에 블러 처리를 하게 되었습니다.
결과를 보니, 딱 옷만 크롭된 것을 확인할 수 있습니다.



블러 처리 전



블러 처리 후

원본 이미지

Mission 2

2-2. Color는 자세한 정보지만, 데이터가 크고, Gray는 덜 자세하지만 데이터가 작아 학습에 유리하다. 어떤 데이터셋이 분류문제에서 더 좋은 결과를 보이는가?

Answer:

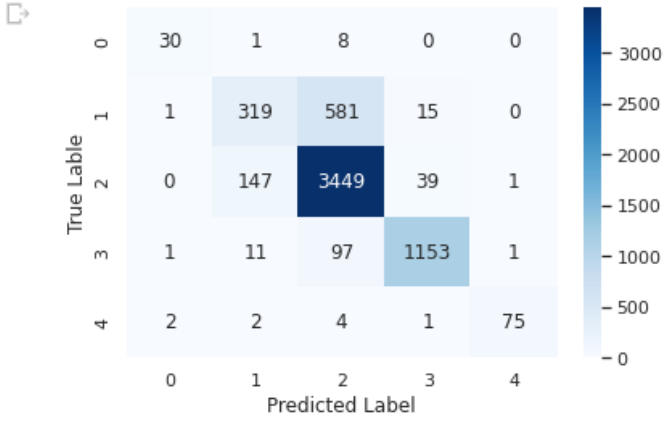
Color는 3차원 데이터로 정보가 너무 많아 정확도가 떨어질 수 있습니다.
저희가 만드는 패션 상품 분류기는 5개의 라벨로 구분만 하면 되는데,
구분은 색이 아닌 형태로 하기 때문에 색상 정보는 필요 없다고 생각되어
Gray Scale로 학습하였습니다.

Mission 3

3-1. 분류 문제를 수행하여 Validation 데이터의 라벨 별 정확도를 제시하시오.

Answer:

```
[86] cm = confusion_matrix(np.argmax(yValidation, axis = -1), np.argmax(pred, axis=-1))
     sns.heatmap(cm, annot=True, fmt='d', cmap='Blues')
     plt.xlabel('Predicted Label')
     plt.ylabel('True Lable')
     plt.show()
```




```
print(classification_report(np.argmax(yValidation, axis = -1), np.argmax(pred, axis=-1)))
```

	precision	recall	f1-score	support
0	0.88	0.77	0.82	39
1	0.66	0.35	0.46	916
2	0.83	0.95	0.89	3636
3	0.95	0.91	0.93	1263
4	0.97	0.89	0.93	84
accuracy			0.85	5938
macro avg	0.86	0.77	0.81	5938
weighted avg	0.84	0.85	0.83	5938

Mission 3

3-2. 정확도를 올리는 작업을 수행하고 작업 수행과정을 설명하시오.

Answer:

1. 데이터 전처리:

정확도를 올리려면 모델 학습의 기초인 전처리가 잘 되어있어야 합니다.

전처리 과정은 위에서 설명했으니 생략하겠습니다.

2. 모델 학습:

모델 학습 시 레이어 설정과 최적화 함수, 하이퍼파라미터 등을 조작하며 모델의 성능을 늘렸습니다.

Mission 3

3-3. 오류가 나온 이미지에 대해 왜 오류가 나왔는지
그동안 미션 수행에서 얻은 경험과 지식을 통해 설명하시오.

Answer:

오류가 나온 이미지는 대부분 라벨링이 잘못 매칭된 경우였습니다.
이 부분 제외시키고 학습하였더니 정확도가 많이 증가했습니다.

```
====] - 2s 8ms/step - loss: 0.4313 - accuracy: 0.8272 - val_loss: 0.4623 - val_accuracy: 0.8242
====] - 2s 8ms/step - loss: 0.4273 - accuracy: 0.8302 - val_loss: 0.4594 - val_accuracy: 0.8255
====] - 2s 8ms/step - loss: 0.4242 - accuracy: 0.8301 - val_loss: 0.4485 - val_accuracy: 0.8302
====] - 2s 8ms/step - loss: 0.4188 - accuracy: 0.8325 - val_loss: 0.4504 - val_accuracy: 0.8352
```

이상치 제외 전

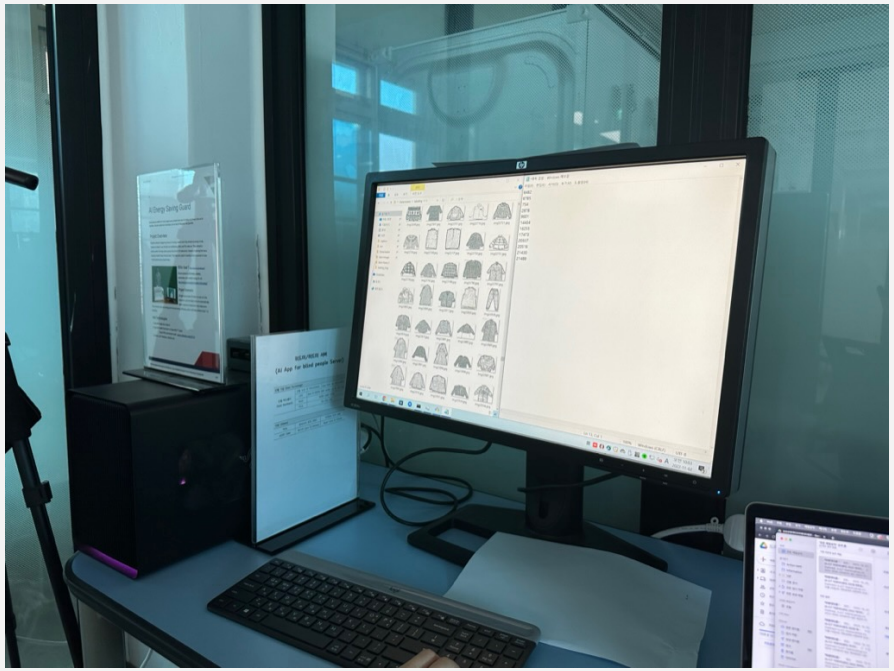
```
- 2s 8ms/step - loss: 0.2554 - accuracy: 0.9006 - val_loss: 0.5014 - val_accuracy: 0.8327
- 2s 8ms/step - loss: 0.2613 - accuracy: 0.8977 - val_loss: 0.4945 - val_accuracy: 0.8297
- 2s 8ms/step - loss: 0.2567 - accuracy: 0.8998 - val_loss: 0.4899 - val_accuracy: 0.8292
- 2s 8ms/step - loss: 0.2523 - accuracy: 0.9035 - val_loss: 0.5044 - val_accuracy: 0.8300
```

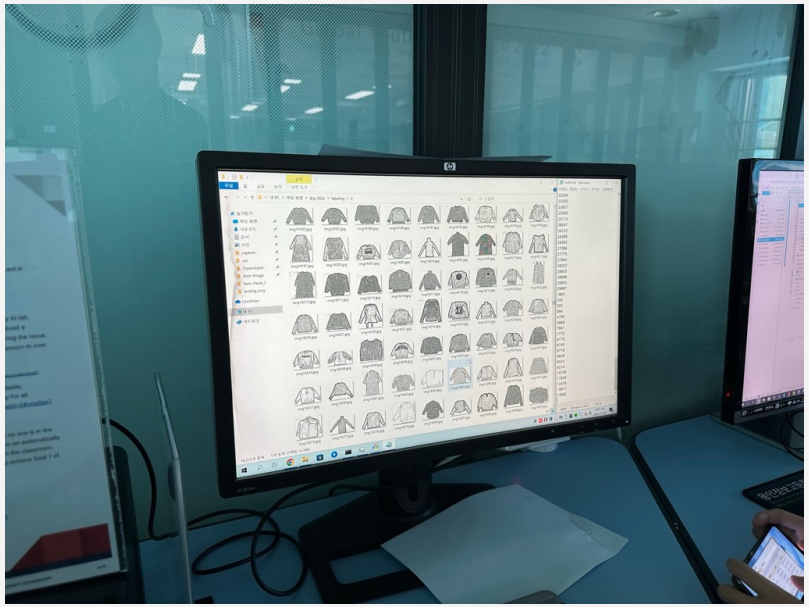
이상치 제외 후

이상치 제외 작업

라벨이 잘못 설정되어 있는 이슈가 있었습니다.

이상치 제거 알고리즘을 만들 수도 있었지만,
시간도 촉박하고 라벨링이 잘못된 데이터가 많지 않아서 직접 제외 작업을 진행했습니다.





클래스 불균형 이슈

5개의 클래스는 서로 다른 개수로 구성되어 있습니다.

특히, tops의 개수가 몇배 더 많이 구성되어 있는데
이를 고려하지 않고 학습 시 모델의 성능이 하락할 수 있습니다.

그렇기에 Weighted Cross Entropy, Focal Loss와 같은 함수를 사용하여
부족한 클래스에 가중치를 곱해 Loss 값을 높혀주거나
정규화 과정을 통한다면 클래스 불균형이 어느 정도 해결될 것이라고 생각합니다.

그러나 시간 관계상 클래스 불균형 문제를 생각하지 못했습니다.

감사합니다

2022 DATA CREATOR CAMP

과학기술정보통신부 NIA 한국지능정보사회진흥원