

Evaluation Metrics

2024.07.12

주현빈

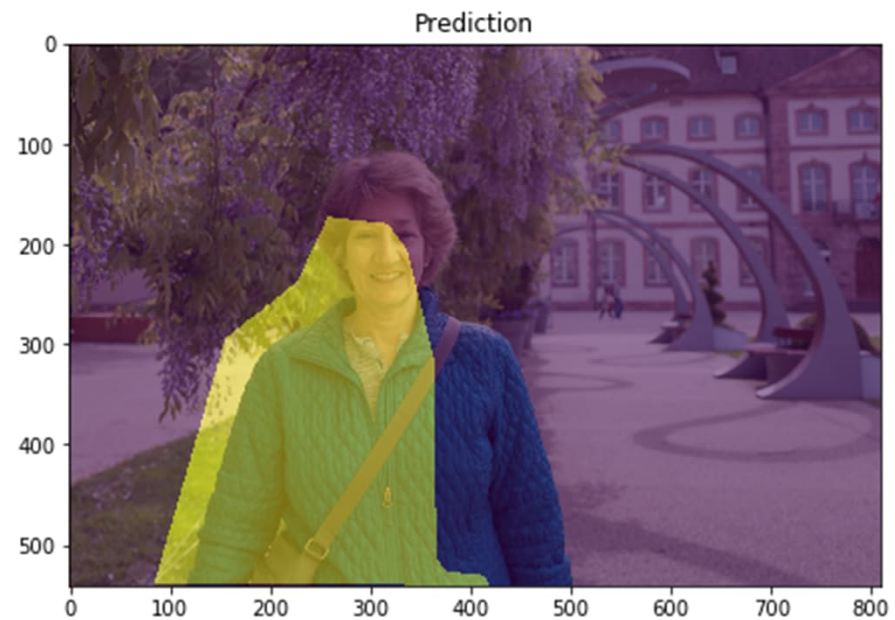
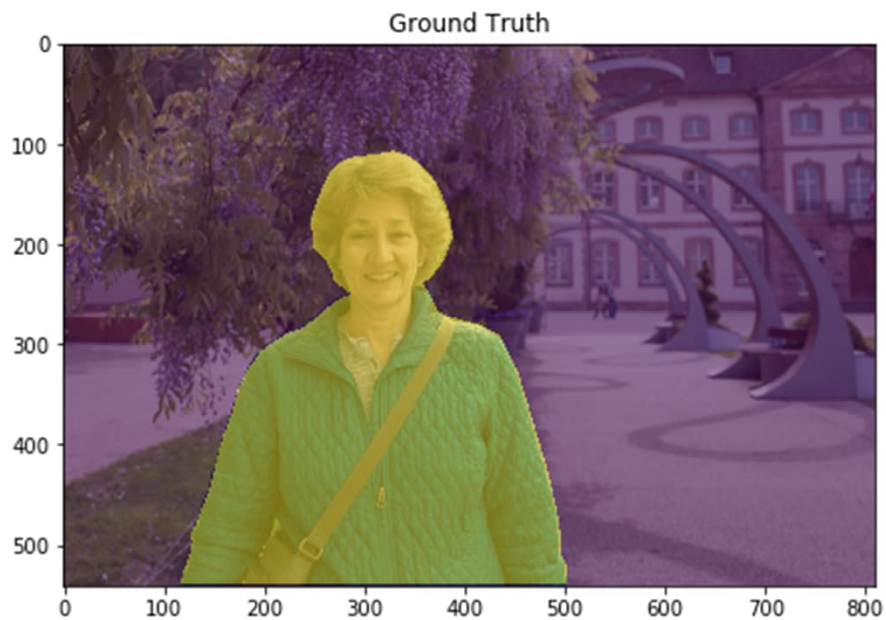


CONTENTS

- Tasks in ML / DL
- What is Loss function and Evaluation Metrics?
- Evaluation Metrics

Quiz

사람을 배경으로부터 분리하는 DL model을 구축했다고 가정

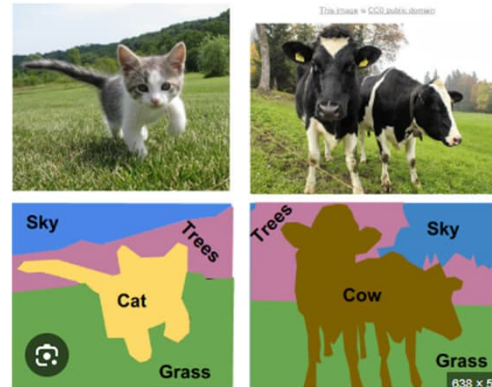
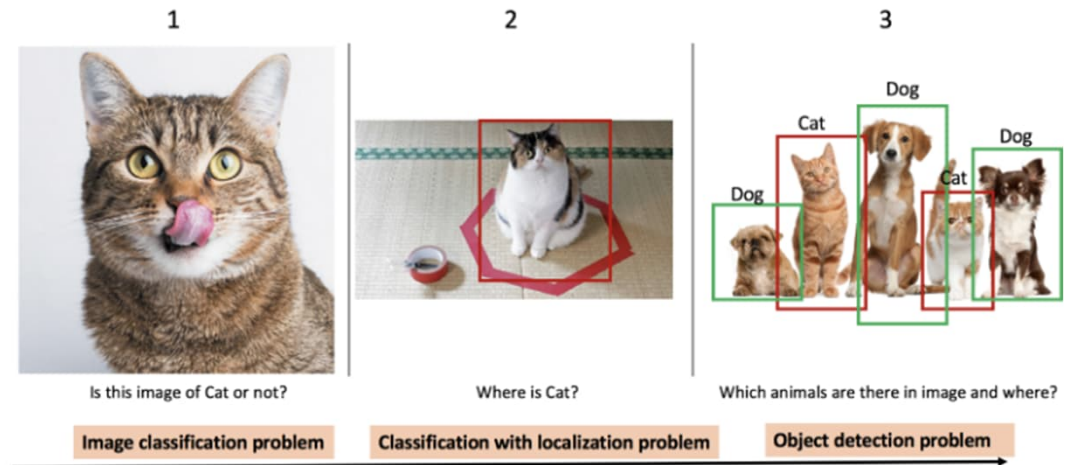
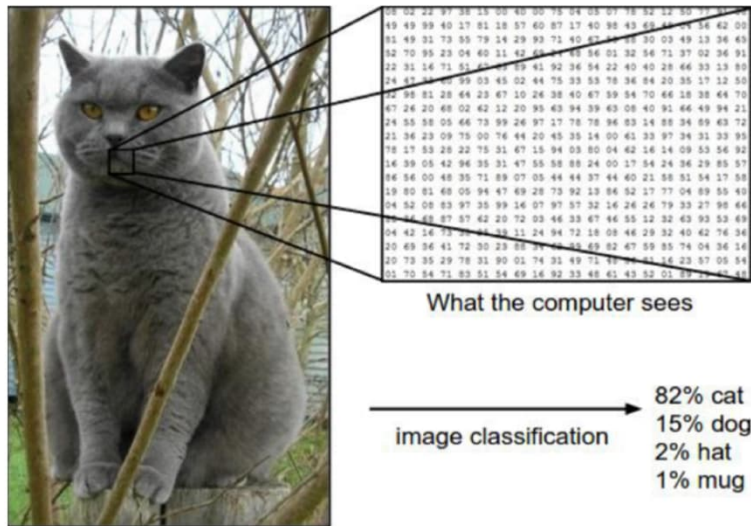


Q. 이 모델의 성능은 몇점인가요?

Tasks in ML / DL

Computer Vision

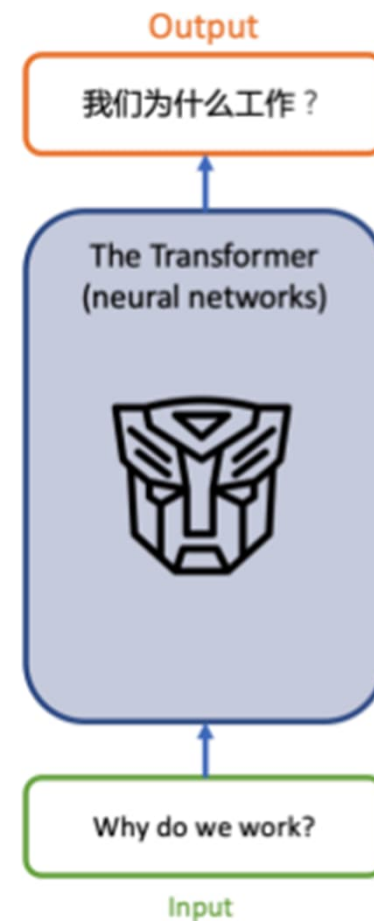
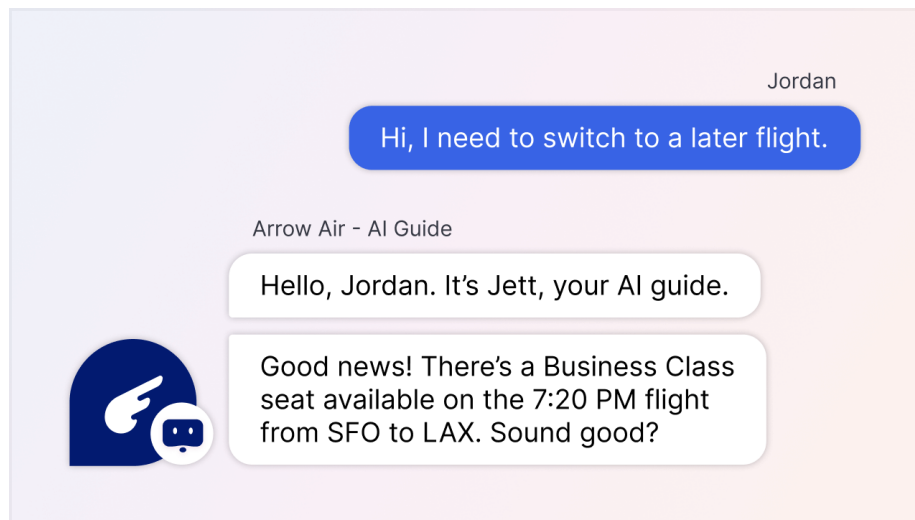
1. Classification
2. Detection
3. Segmentation
4. Image Generation



Tasks in ML / DL

Natural Language Processing

1. Machine Translation
2. Summarization
3. Sentiment Analysis
4. Chatbot

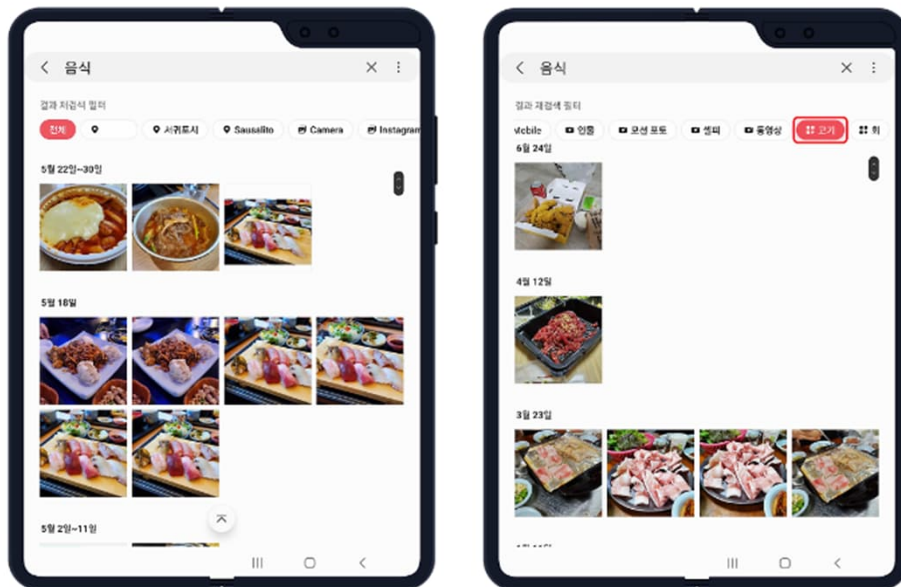


Tasks in ML / DL

Vision Language Model

1. Text-Image Retrieval
2. Image Captioning
3. Image Generation by text prompt

진료를 보는 판다 이미지 생성좀 해줘



What is Loss Function and Evaluation Metrics?




ML, DL은 오직 데이터만을 활용하여 Train된다

Training : Loss function을 최소화하기

Loss Function : 머신러닝의 train 단계에서, 미니 배치에 대해 얼마나 잘 풀었는지를 정량적으로 평가 -> 연습문제를 얼마나 잘 풀었는지 -> **Loss Function에 따라 얼마나 가중치를 update할지 정해짐**

Evaluation Metric : 학습이 끝난 model에서, test dataset에 대해 얼마나 일반화 성능이 높은지를 평가 -> 수능 성적이 몇점인지

NN1

	Dog	1	0.9
	Cat	0	0.1
	Dog	0	0.1
	Cat	1	0.9
	Dog	1	0.4
	Cat	0	0.6

Dog Image를 보고,
Dog : 0.9, Cat : 0.1을 출력했다면, 얼마나
잘 출력했다고 말할 수 있는가?

$$\begin{aligned}\text{CE Loss} &= -\sum_{c=0}^1 y_c \log(\hat{y}_c) \\ \text{CE Loss} &= -(y_0 \log(\hat{y}_0) + y_1 \log(\hat{y}_1)) \\ \text{CE Loss} &= -(1 \cdot \log(0.9) + 0 \cdot \log(0.1)) \\ \text{CE Loss} &= -\log(0.9) \\ \text{CE Loss} &= -(-0.10536) = 0.10536\end{aligned}$$

Dog Image를 보고,
Dog : 0.4, Cat : 0.6을 출력했다면, 얼마나
잘 출력했다고 말할 수 있는가?

$$\begin{aligned}\text{CE Loss} &= -(1 \cdot \log(0.6) + 0 \cdot \log(0.4)) \\ \text{CE Loss} &= -\log(0.6) \\ \text{CE Loss} &= -(-0.51083) = 0.51083\end{aligned}$$

What is Loss Function and Evaluation Metrics?

이에 반해, Evaluation Metrics는 학습이 끝난 모델이 얼마나 Test Dataset에 대해 일반화 성능이 높은지 평가한다

Dog와 Cat을 Classification하는 문제라고 가정
Test Dataset : 100장 (Dog : 70장, Cat : 30장)

Actual Class	Cat	Dog
	23	7
Dog	10	60
Predicted Class		

왼쪽과 같은 Confusion Matrix가 있을 때,
모델의 성능이 어떻다고 평가할 수 있을까?

A1. 100개 중에, 83개나 맞췄네 !

-> 단순히 정확도로만 평가할 수 있을까?

What is Loss Function and Evaluation Metrics?

이에 반해, Evaluation Metrics는 학습이 끝난 모델이 얼마나 Test Dataset에 대해 일반화 성능이 높은지 평가한다

Cancer Positive, Negative를 Classification하는 문제라고 가정

Test Dataset : 1000장 (Cancer Negative : 990장, Cancer Positive : 10장)

Confusion Matrix with Enhanced Contrast

Actual	Actual Negative	989	1
	Actual Positive	9	1
		Predicted Negative	Predicted Positive
		Predicted	

A1. 1000장 중, 990장을 맞췄네 !

A2. 암 환자가 실제로 10명인데, 1명밖에 못 맞췄네..

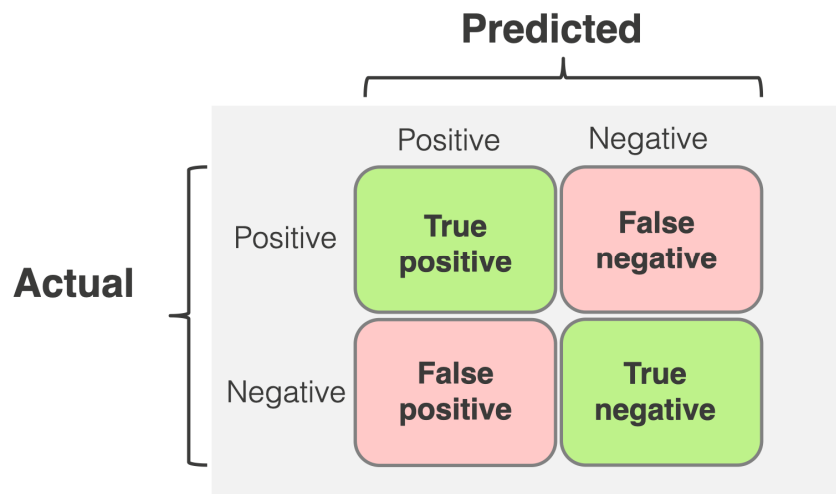
-> 의료 데이터셋에서는 data imbalance가 심하다

-> Loss function과 Metric에 대한 이해가 필요

Evaluation Metrics

1. Classification Task 1.1. Confusion Matrix

-> 모델의 학습이 완료된 뒤, Threshold가 정해졌을 때 test dataset에 대해 통계적으로 분석하는 방법



1. Accuracy -> $TP + TN / TP + TN + FP + FN$

2. Precision -> $TP / TP + FP$ (Positive라고 예측한 것 중 정답)

3. Recall -> $TP / TP + FN$ (실제 Positive 중 정답)

4. F1 score -> Precision과 Recall의 조화평균

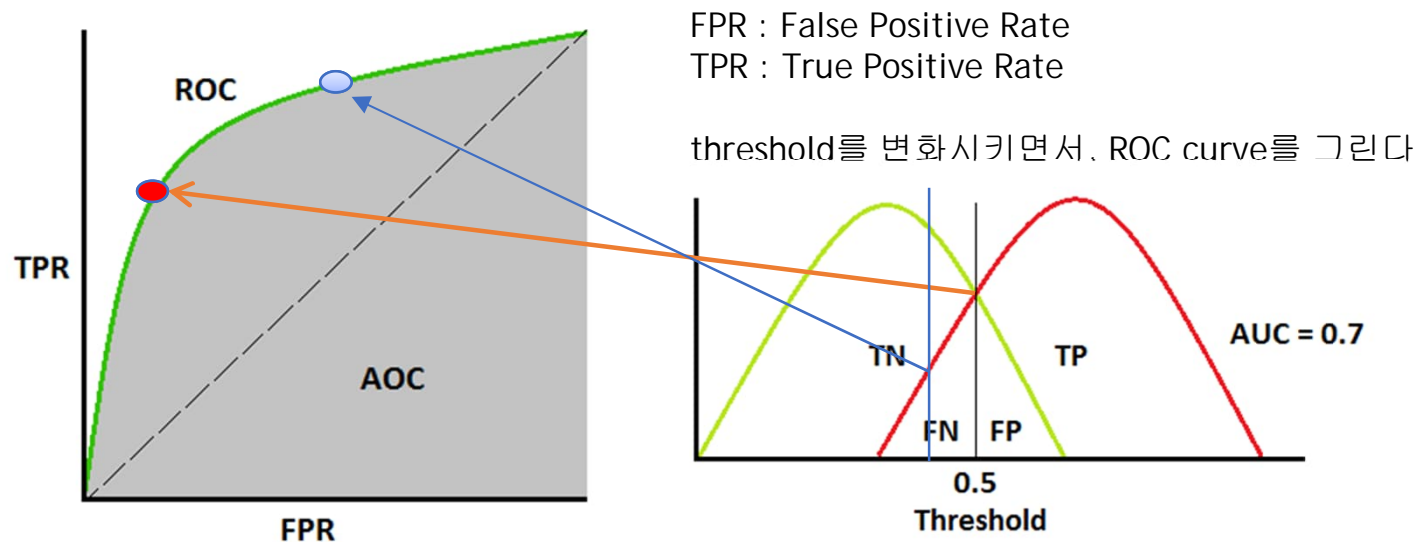
Method	Accuracy	Precision	Recall	F1-Score
CCAIE	0.7375	0.73	0.72	0.72
DenseNet	0.7054	0.72	0.71	0.71
ResNet	0.6129	0.67	0.63	0.65
Inception	0.5895	0.61	0.60	0.60
Xception	0.6988	0.73	0.70	0.71

1.2. AUC (Area under the ROC(Receiver Operating Characteristic))

앞의 경우, Test dataset에 대한 통계적인 성능을 평가했다

-> AUC의 경우 모델 그 자체의 분류 능력을 함께 평가함 ! (Threshold를 변화시켜 가면서)

-> 모델이 얼마나 특징을 잘 파악해서, 두 class를 구분할 수 있는 형태로 만들었는가?



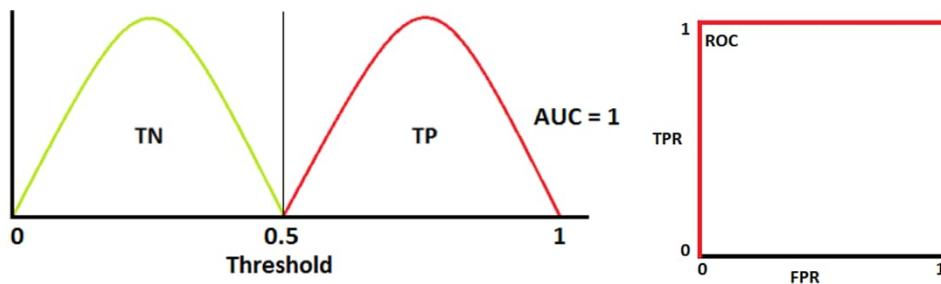
Evaluation Metrics

1.2. AUC (Area under the ROC(Receiver Operating Characteristic))

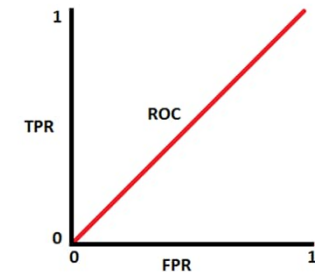
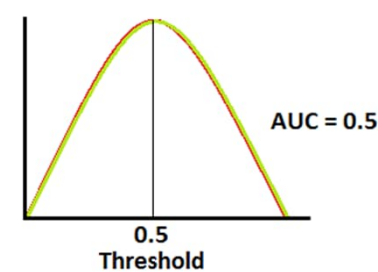
앞의 경우, Test dataset에 대한 통계적인 성능을 평가했다

-> AUC의 경우 모델 그 자체의 분류 능력을 함께 평가함 ! (Threshold를 변화시켜 가면서)

-> 모델이 얼마나 특징을 잘 파악해서, 두 class를 구분할 수 있는 형태로 만들었는가?



Class의 구분이 굉장히 뚜렷해서,
어떤 threshold에서도 완벽히 구분할 수 있는 상태 -> AUC = 1

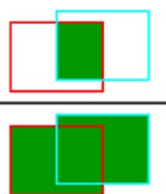


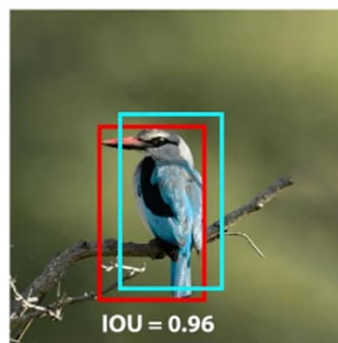
Class의 구분이 전혀 되지 않아서,
어떤 Threshold에서도 FPR과 TPR이 동일한 상태 -> AUC = 0.5

Evaluation Metrics

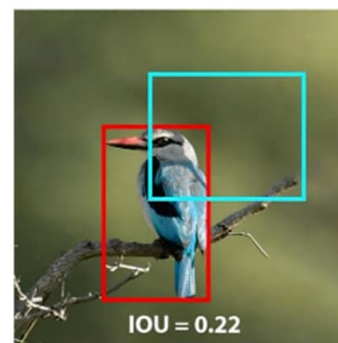
2. Segmentation, Bounding Box Creation Task

2.1. IoU (Intersection over Union)

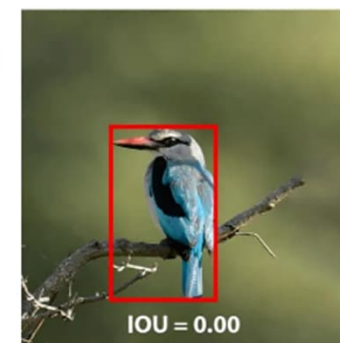
$$IOU = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$




True Positive



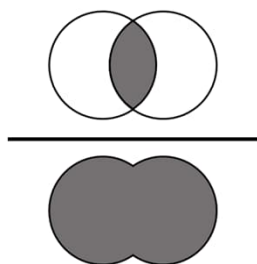
False Positive



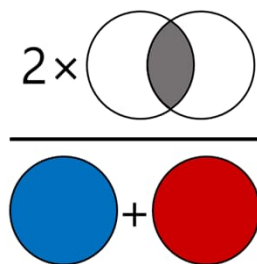
False Negative

2.2. Dice Coefficient

IoU



Dice Coefficient



Dice Coef가 IoU에 비해 상대적으로 더 민감하다

Evaluation Metrics

3. 생성형 모델에서의 성능 평가는..?



Text Prompt : On a peaceful afternoon, a dog and cat are basking in the warm sunlight

어떤 모델의 성능이 더 높다고 말할 수 있는가?

이해가 안가시는 부분이 있다면, 질문 부탁드립니다 !