



A(uto)ssignment

유현동 김현동
정성오 고현아



I. Introduction

II. Dataset & Preprocessing

III. Pipelines(YOLO & PORORO)

IV. Result

V. Limitation

Section 1

Introduction



페이지 넘기면서 문제 찾고
과제하기가 너무 힘든데...



문제를 자동으로 찾아서
문제집을 만들어주는
모델을 만들자!

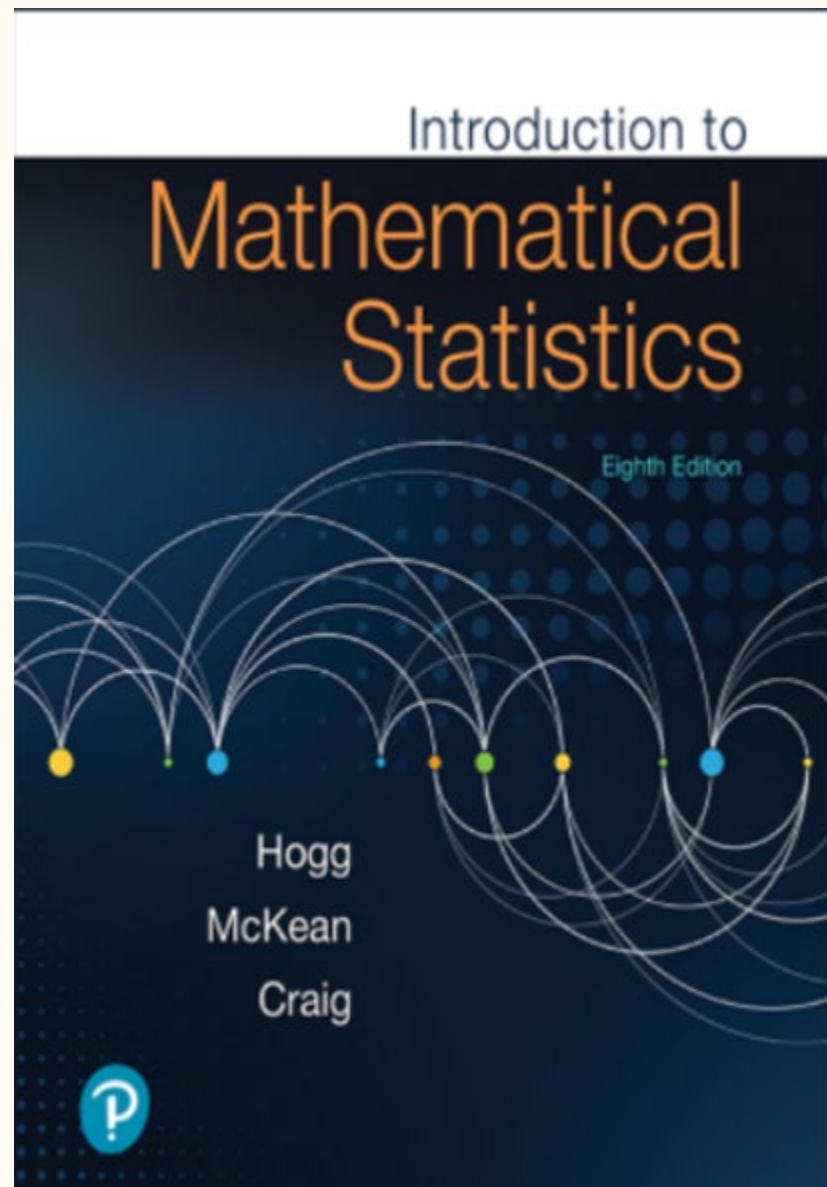


HW#3_Energy Fundamentals

Solve the problems from 4-33 to 4-44 at pages of 194 and 195 in the main textbook.

Section 2

Dataset & Preprocessing



138 Multivariate Distributions

or independent random variables X_1 and X_2 becomes, for mutually independent random variables X_1, X_2, \dots, X_n ,

$$E[u_1(X_1)u_2(X_2) \cdots u_n(X_n)] = E[u_1(X_1)]E[u_2(X_2)] \cdots E[u_n(X_n)].$$

139

$$E\left[\prod_{i=1}^n u_i(X_i)\right] = \prod_{i=1}^n E[u_i(X_i)].$$

The moment-generating function (mgf) of the joint distribution of n random variables X_1, X_2, \dots, X_n is defined as follows. Suppose that

$$E[\exp(t_1 X_1 + t_2 X_2 + \cdots + t_n X_n)]$$

exists for $-h_i < t_i < h_i$, $i = 1, 2, \dots, n$, where each h_i is positive. This expectation is denoted by $M(t_1, t_2, \dots, t_n)$ and it is called the mgf of the joint distribution of X_1, \dots, X_n (or simply the mgf of X_1, \dots, X_n). As in the cases of one and two variables, this mgf is unique and uniquely determines the joint distribution of the n variables (and hence all marginal distributions). For example, the mgf of the marginal distributions of X_i is $M(0, \dots, 0, t_i, 0, \dots, 0)$, $i = 1, 2, \dots, n$; that of the marginal distribution of X_i and X_j is $M(0, \dots, 0, t_i, 0, \dots, 0, t_j, 0, \dots, 0)$; and so on. Theorem 2.4.5 of this chapter can be generalized, and the factorization

$$M(t_1, t_2, \dots, t_n) = \prod_{i=1}^n M(0, \dots, 0, t_i, 0, \dots, 0) \quad (2.6.6)$$

is a necessary and sufficient condition for the mutual independence of X_1, X_2, \dots, X_n . Note that we can write the joint mgf in vector notation as

$$M(\mathbf{t}) = E[\exp(\mathbf{t}'\mathbf{X})], \quad \text{for } \mathbf{t} \in B \subset R^n,$$

where $B = \{\mathbf{t} : -h_i < t_i < h_i, i = 1, \dots, n\}$.

The following is a theorem that proves useful in the sequel. It gives the mgf of a linear combination of independent random variables.

Theorem 2.6.1. Suppose X_1, X_2, \dots, X_n are n mutually independent random variables. Suppose, for all $i = 1, 2, \dots, n$, X_i has mgf $M_i(t_i)$, for $-h_i < t_i < h_i$, where $h_i > 0$. Let $T = \sum_{i=1}^n k_i X_i$, where k_1, k_2, \dots, k_n are constants. Then T has the mgf given by

$$M_T(t) = \prod_{i=1}^n M_i(k_i t), \quad -\min_i \{h_i\} < t < \min_i \{h_i\} \quad (2.6.7)$$

Proof. Assume t is in the interval $(-\min_i \{h_i\}, \min_i \{h_i\})$. Then, by independence,

$$\begin{aligned} M_T(t) &= E\left[e^{\sum_{i=1}^n k_i X_i}\right] = E\left[\prod_{i=1}^n e^{k_i X_i}\right] \\ &= \prod_{i=1}^n E[e^{k_i X_i}] = \prod_{i=1}^n M_i(k_i t), \end{aligned}$$

which completes the proof. ■

3.4. The Normal Distribution 195

3.4.8. Evaluate $\int_{-\infty}^{\infty} \exp[-2(x-3)^2] dx$.

3.4.9. Determine the 90th percentile of the distribution, which is $N(65, 25)$.

3.4.10. If $e^{it + it^2}$ is the mgf of the random variable X , find $P(-1 < X < 9)$.

3.4.11. Let the random variable X have the pdf

$$f(x) = \frac{2}{\sqrt{2\pi}} e^{-x^2/2}, \quad 0 < x < \infty, \quad \text{zero elsewhere.}$$

(a) Find the mean and the variance of X .

(b) Find the cdf and hazard function of X .

Hint for (a): Compute $E(X)$ directly and $E(X^2)$ by comparing the integral with the integral representing the variance of a random variable that is $N(0, 1)$.

3.4.12. Let X be $N(5, 10)$. Find $P[0.04 < (X - 5)^2 < 38.4]$.

3.4.13. If X is $N(1, 4)$, compute the probability $P(1 < X^2 < 9)$.

3.4.14. If X is $N(75, 25)$, find the conditional probability that X is greater than 80 given that X is greater than 77. See Exercise 2.3.12.

3.4.15. Let X be a random variable such that $E(X^{2m}) = (2m)!/(2^m m!)$, $m = 1, 2, 3, \dots$ and $E(X^{2m-1}) = 0$, $m = 1, 2, 3, \dots$. Find the mgf and the pdf of X .

3.4.16. Let the mutually independent random variables X_1, X_2 , and X_3 be $N(0, 1)$, $N(2, 4)$, and $N(-1, 1)$, respectively. Compute the probability that exactly two of these three variables are less than zero.

3.4.17. Compute the measures of skewness and kurtosis of a distribution which is $N(\mu, \sigma^2)$. See Exercises 1.9.14 and 1.9.15 for the definitions of skewness and kurtosis, respectively.

3.4.18. Let the random variable X have a distribution that is $N(\mu, \sigma^2)$.

(a) Does the random variable $Y = X^2$ also have a normal distribution?

(b) Would the random variable $Y = aX + b$, a and b nonzero constants have a normal distribution?

Hint: In each case, first determine $P(Y \leq y)$.

3.4.19. Let the random variable X be $N(\mu, \sigma^2)$. What would this distribution be if $\sigma^2 = 0$?

Hint: Look at the mgf of X for $\sigma^2 > 0$ and investigate its limit as $\sigma^2 \rightarrow 0$.

3.4.20. Let Y have a truncated distribution with pdf $g(y) = \phi(y)/[\Phi(b) - \Phi(a)]$, for $a < y < b$, zero elsewhere, where $\phi(x)$ and $\Phi(x)$ are, respectively, the pdf and distribution function of a standard normal distribution. Show then that $E(Y)$ is equal to $[\phi(a) - \phi(b)]/[\Phi(b) - \Phi(a)]$.

Color	Class Name
	caption
	code
	image
	image_caption
	page_num
	question
	question_num
	table
	table_caption
	text
	title

Section 3

Pipelines



YOLO v7

PORORO

PyMuPDF를 활용해
PDF 파일을 PNG 파일로 변환



YOLOv7 모델을 활용해
문제 부분만(question class)
추출



PORORO 모델
->문제 번호 식별
->문제 + 여백 이미지 저장



사용자로부터
문제번호 입력받아
PDF 파일로 변환해 저장



Model: YOLO v7

You Only Look Once version 7

- 객체 감지(object detection) 및 분류(classification) 수행하는 딥러닝 모델
- Roboflow를 통해 Image Segmentation이 완료된 데이터를 입력으로 받아 classification 수행
- 이미지에서 question 부분을 탐지

Dataset for Training

Introduction to Mathematical Statistics(8th)

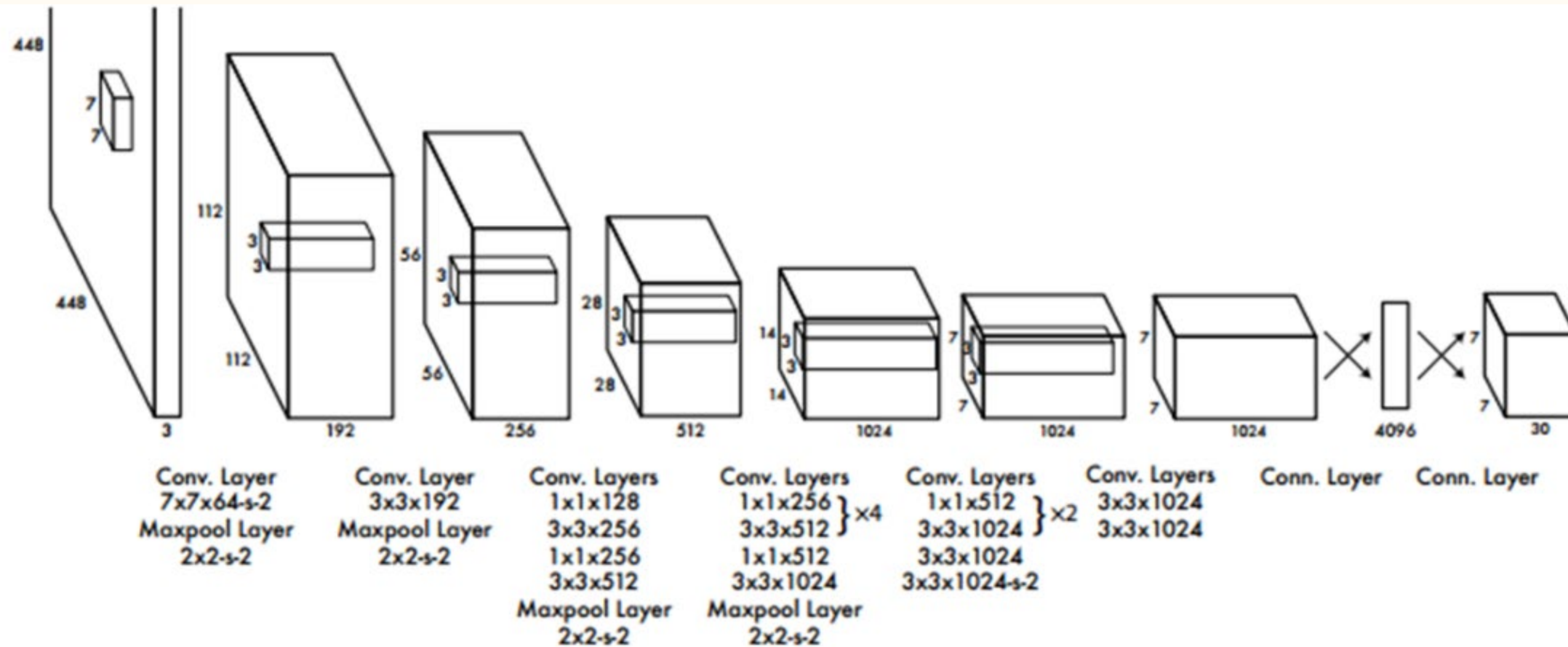
- pdf를 이미지로 바꾼 후 Roboflow를 통해 Image Segmentation 진행
- 11개의 클래스로 분류 후, training/validation/test 진행

03

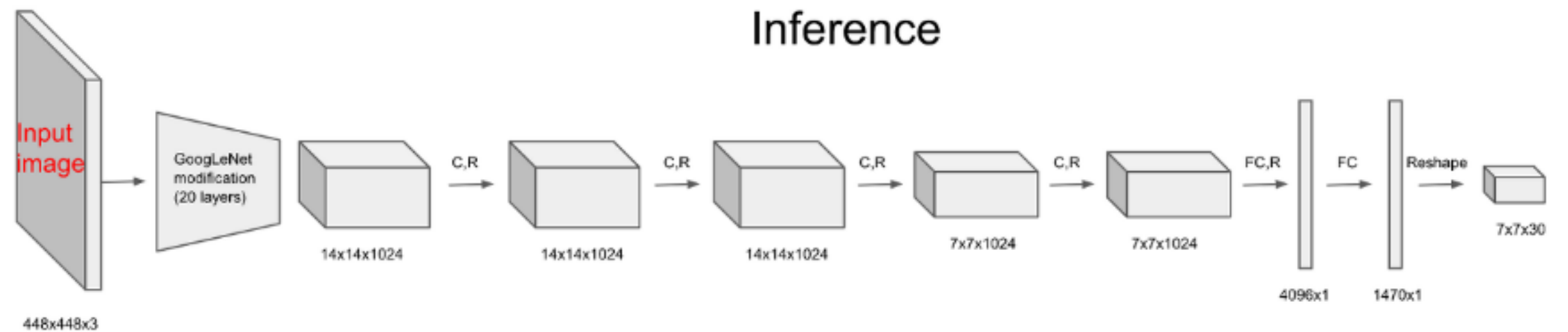
YOLO란?



YOLOv1 기준

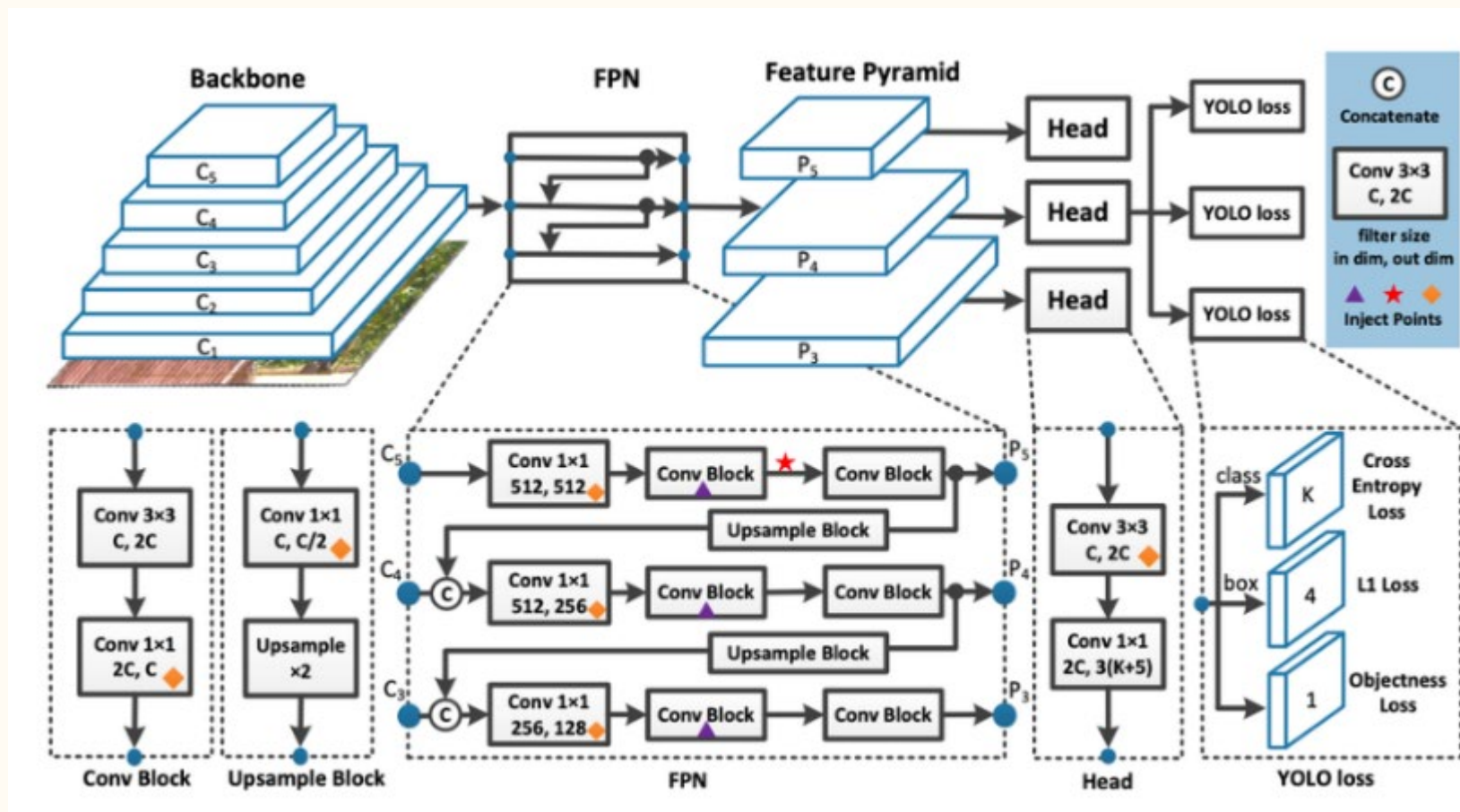


Inference



03

YOLO v7



- YOLO 모델들은 single network 하나만을 사용하여 속도가 빠름
- YOLO v7은 모델 학습 시, memory cost, 계산량을 줄이기 위해 E-ELAN을 baseline 구조 적용
- 기존의 SOTA 모델보다 파라미터 수와 계산량을 50% 감소시키고 더 빠른 inference time과 더 높은 정확도를 달성



1.4.3. Suppose we are playing draw poker. We are dealt (from a well-shuffled deck) five cards, which contain four spades and another card of a different suit. We decide to discard the card of a different suit and draw one card from the remaining cards to complete a flush in spades (all five cards spades). Determine the probability of completing the flush.



1.5.8. Suppose the random variable X has the cdf

$$F(x) = \begin{cases} 0 & x < -1 \\ \frac{x+2}{4} & -1 \leq x < 1 \\ 1 & 1 \leq x. \end{cases}$$

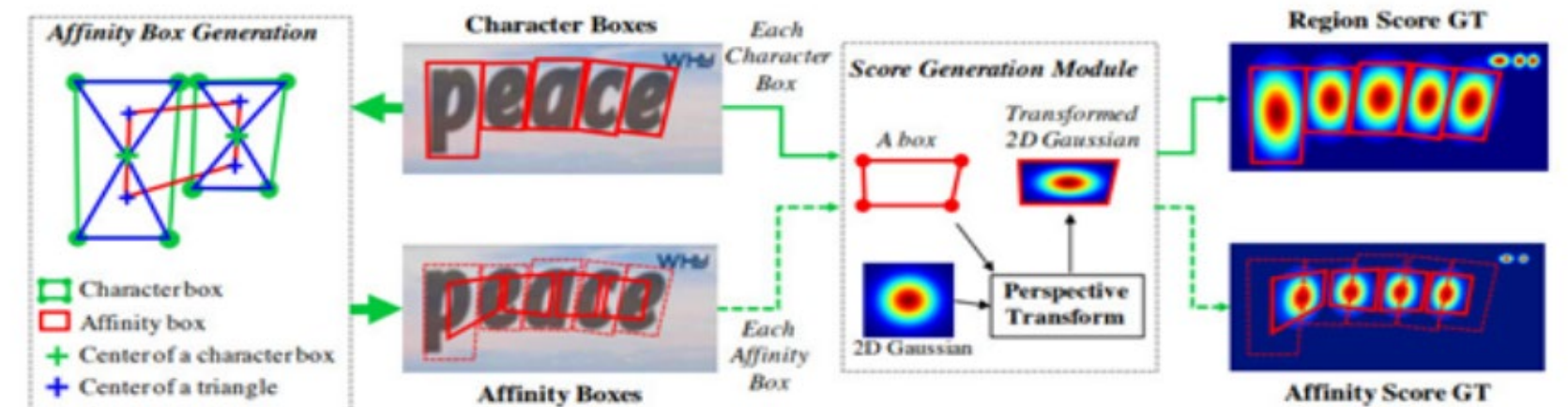
Write an R function to sketch the graph of $F(x)$. Use your graph to obtain the probabilities: (a) $P(-\frac{1}{2} < X \leq \frac{1}{2})$; (b) $P(X = 0)$; (c) $P(X = 1)$; (d) $P(2 < X \leq 3)$.



Model: PORORO

Platform Of neuRal mOdelS for natuRal
language prOcessing

- text classification, sequence tagging 등의 다양한 task 수행
- PORORO ocr을 사용하여 question으로 분류된 이미지들의 문제 번호를 인식
- YOLO v7을 통해 분류한 question class를 이미지로 저장하여 input으로 제공

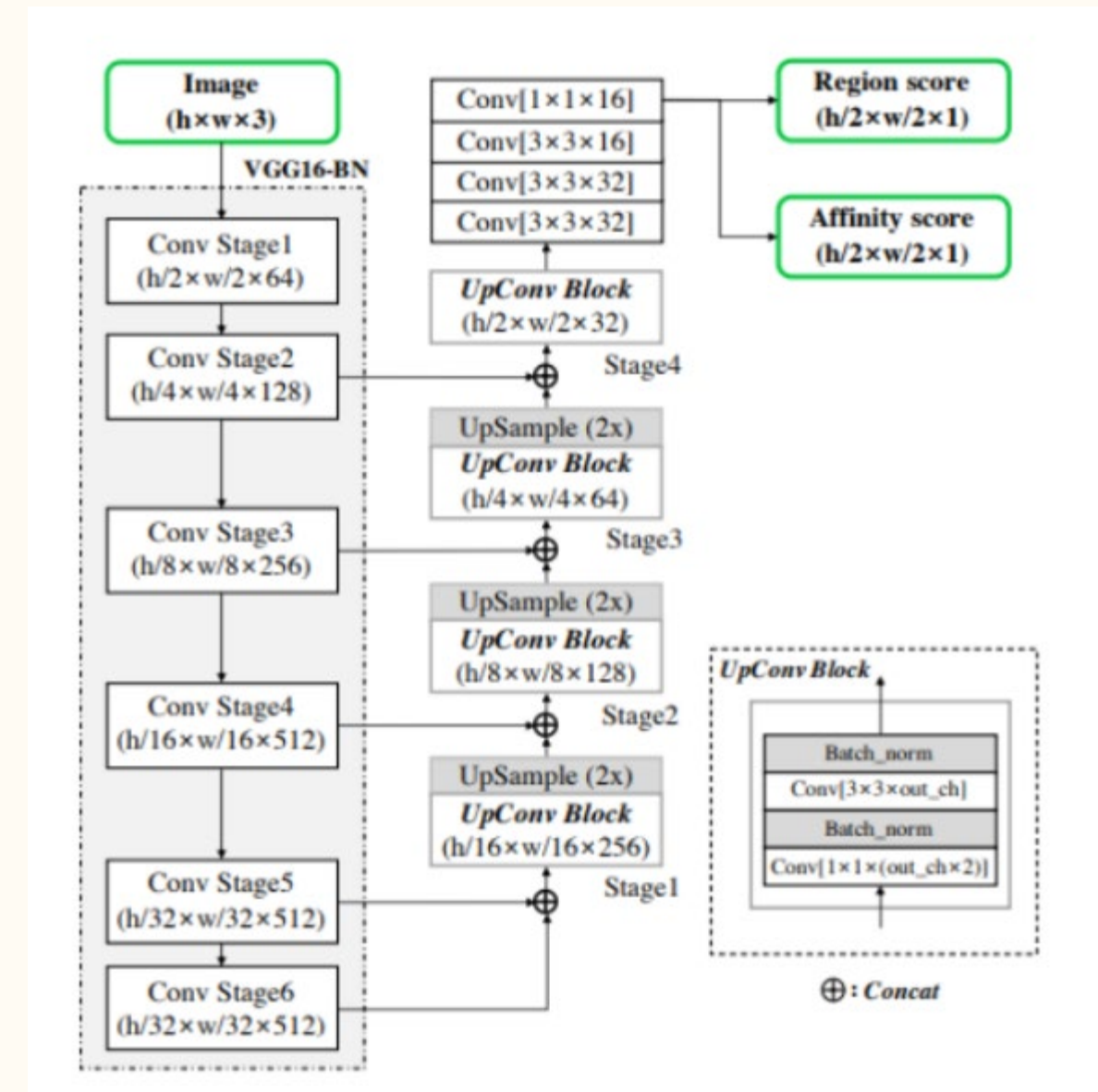




Model: CRAFT by NAVER

Character-Region Awareness For Text detection

- 입력값으로 이미지를 받고, label 로 region, affinity map 을 갖는 형태
- VGG16 과 U-Net 를 합친 것과 유사한 형태





Loss function of CRAFT

$$L = \sum_p S_c(p) \cdot (\|S_r(p) - S_r^*(p)\|_2^2 + \|S_a(p) - S_a^*(p)\|_2^2)$$

$S_c(p)$: pixel-wise confidence, 각 픽셀에서 word 부분이 맞는지 confidence를 사용하여 계산
나머지 부분 : 각 픽셀에서 regional score, affinity score의 정답과의 유클리디안 유사도를 계산



TPS+VGG(or ResNet)+BiLSTM+CTC(or Attn, Transformer)

TPS Spatial Trasformation Network

- Thin-Plate Spline(박막 스플라인)
- 입력 이미지를 변환하여 네트워크가 더 잘 처리하도록 도와줌
- 주어진 입력 점들 간의 최소 에너지를 사용하여 곡면을 매핑하는 비선형 함수



2. VGG or ResNet

- 사용자의 입력에 따라 VGG 또는 ResNet 사용
- feature extractor로써 사용됨



3. BiLSTM

- bi-directional LSTM, sequence modeling
- 시계열 또는 시퀀스 데이터의 time step에서 양방향 장기 종속성을 학습하는 RNN 계층
- 기존 LSTM에서 역방향으로 실행되는 다른 LSTM을 추가
- 각 시점에서 hidden state가 이전 시점과 미래 시점의 정보를 모두 갖는 효과가 있기 때문에 모델을 전체 sequence 데이터로부터 학습할 수 있도록 하려는 경우 유용



4. CTC or Attention or Transformer

- Connectionist Temporal Classification
- 시계열 데이터를 처리하는 데 사용되는 딥러닝 모델, OCR에서는 text detection을 통해 입력된 데이터를 처리하기 위해 사용
- 음성 인식이나 필기 인식 등의 분야에서 많이 사용됨
- 입력 데이터의 시간적 순서를 유지하면서, 출력 시퀀스의 길이가 입력 시퀀스와 다를 수 있는 상황을 효과적으로 처리

Section 4

Result

**Input**

요청한 문제가 데이터베이스 내에 모두 존재하는 경우

Output

```
몇 개의 문제가 필요하시나요? 3
원하는 문제 번호를 입력하세요: 5.1.1.
Q1: 5.1.1.
원하는 문제 번호를 입력하세요: 5.3.11.
Q2: 5.3.11.
원하는 문제 번호를 입력하세요: 6.1.12.
Q3: 6.1.12.
입력받은 문제 번호들: ['5.1.1.', '5.3.11.', '6.1.12.']
PDF 파일이 생성되었습니다: /content/drive/MyDrive/outputs/output_7.pdf
```

**Input**

요청한 문제들이 일부 데이터베이스에 존재하는 경우

Output

```
몇 개의 문제가 필요하시나요? 5
원하는 문제 번호를 입력하세요: 1.2.1.
Q1: 1.2.1.
원하는 문제 번호를 입력하세요: 2.3.2.
Q2: 2.3.2.
원하는 문제 번호를 입력하세요: 1.10.6.
Q3: 1.10.6.
원하는 문제 번호를 입력하세요: 4.1.2.
Q4: 4.1.2.
원하는 문제 번호를 입력하세요: 6.5.15.
Q5: 6.5.15.
입력받은 문제 번호들: ['1.2.1.', '2.3.2.', '1.10.6.', '4.1.2.', '6.5.15.']
문제번호 1.2.1.가 데이터베이스에 존재하지 않습니다.
문제번호 4.1.2.가 데이터베이스에 존재하지 않습니다.
PDF 파일이 생성되었습니다: /content/drive/MyDrive/outputs/output_8.pdf
```



Input

요청한 문제가 데이터베이스 내에 존재하지 않는 경우

Output

```
몇 개의 문제가 필요하시나요? 3
원하는 문제 번호를 입력하세요: 6.6.3.
Q1: 6.6.3.
원하는 문제 번호를 입력하세요: 4.10.2.
Q2: 4.10.2.
원하는 문제 번호를 입력하세요: 10.5.3.
Q3: 10.5.3.
입력받은 문제 번호들: ['6.6.3.', '4.10.2.', '10.5.3.']
문제번호 6.6.3.가 데이터베이스에 존재하지 않습니다.
문제번호 4.10.2.가 데이터베이스에 존재하지 않습니다.
문제번호 10.5.3.가 데이터베이스에 존재하지 않습니다.
Traceback (most recent call last):
  File "/content/drive/MyDrive/유현동/problem_query.py", line 66, in <module>
    images_to_pdf(image_paths, output_pdf)
  File "/content/drive/MyDrive/유현동/problem_query.py", line 46, in images_to_pdf
    images[0].save(output_pdf, save_all=True, append_images=images[1:])
IndexError: list index out of range
```




1.10.6. The mgf of X exists for all real values of t and is given by

$$M(t) = \frac{e^t - e^{-t}}{2t}, \quad t \neq 0, \quad M(0) = 1.$$

Use the results of the preceding exercise to show that $P(X \geq 1) = 0$ and $P(X \leq -1) = 0$. Note that here h is infinite.

2.3.2. Let $f_{12}(x_1|x_2) = c_1x_1/x_2^2$, $0 < x_1 < x_2$, $0 < x_2 < 1$, zero elsewhere, and $f_2(x_2) = c_2x_2^4$, $0 < x_2 < 1$, zero elsewhere, denote, respectively, the conditional pdf of X_1 , given $X_2 = x_2$, and the marginal pdf of X_2 . Determine:

- (a) The constants c_1 and c_2 .
- (b) The joint pdf of X_1 and X_2 .
- (c) $P(\frac{1}{4} < X_1 < \frac{1}{2} | X_2 = \frac{1}{3})$.
- (d) $P(\frac{1}{4} < X_1 < \frac{1}{2})$.

6.5.15. A machine shop that manufactures toggle levers has both a day and a night shift. A toggle lever is defective if a standard nut cannot be screwed onto the threads. Let p_1 and p_2 be the proportion of defective levers among those manufactured by the day and night shifts, respectively. We shall test the null hypothesis, $H_0 : p_1 = p_2$, against a two-sided alternative hypothesis based on two random samples, each of 1000 levers taken from the production of the respective shifts. Use the test statistic Z^* given in Example 6.5.3.

- (a) Sketch a standard normal pdf illustrating the critical region having $\alpha = 0.05$.
- (b) If $y_1 = 37$ and $y_2 = 53$ defectives were observed for the day and night shifts, respectively, calculate the value of the test statistic and the approximate p -value (note that this is a two-sided test). Locate the calculated test statistic on your figure in part (a) and state your conclusion. Obtain the approximate p -value of the test.

Section 5

Limitation



- Overfitting : 다른 교재에도 훈련시킨 YOLO 모델을 적용시켰으나 question을 제대로 인식하지 못하는 문제 발생
- OCR model : OCR 모델의 한계점으로 question별로 추출된 이미지의 text를 모두 인식하지 못함(question number를 인식하지 못하는 경우가 생김)
- 데이터 전처리 : 교재의 모든 페이지를 segmentation 해야 하기 때문에 많은 시간이 걸림

Q & A

Thank You!