Optical Character Recognition for R&D LLMs Learning Dataset

Seminar – Fall 2023

Min-kyun Ko

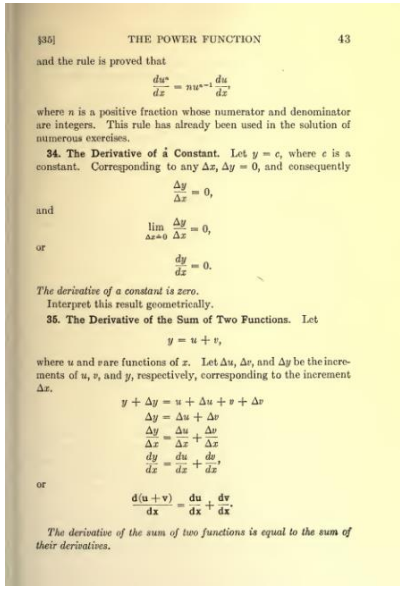
CONTENTS

01 Introduction

02 Related Studies

03 Motivation and Ongoing

- OCR is the process of converting a text image into **computer readable text**
- OCR = **Text Detection** + **Text Recognition**



§35] THE POWER FUNCTION 43

and the rule is proved that

$$\frac{du^n}{dx} = nu^{n-1}\frac{du}{dx},$$

where n is a positive fraction whose numerator and denominator are integers. This rule has already been used in the solution of numerous exercises.

34. The Derivative of a Constant. Let $y = c$, where c is a constant. Corresponding to any Δx, $\Delta y = 0$, and consequently

$$\frac{\Delta y}{\Delta x} = 0,$$

and

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = 0,$$

or

$$\frac{dy}{dx} = 0.$$

The derivative of a constant is zero.

Interpret this result geometrically.

35. The Derivative of the Sum of Two Functions. Let

$$y = u + v,$$

where u and v are functions of x. Let Δu, Δv, and Δy be the increments of u, v, and y, respectively, corresponding to the increment Δx.

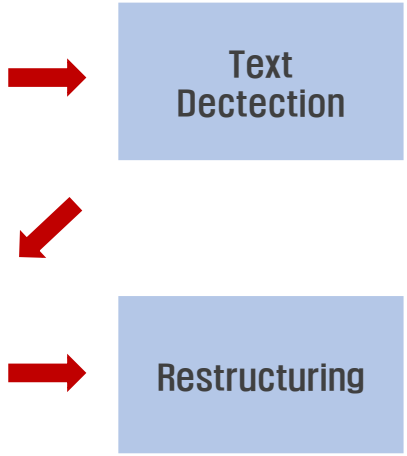
$$y + \Delta y = u + \Delta u + v + \Delta v$$
$$\Delta y = \Delta u + \Delta v$$
$$\frac{\Delta y}{\Delta x} = \frac{\Delta u}{\Delta x} + \frac{\Delta v}{\Delta x}$$
$$\frac{dy}{dx} = \frac{du}{dx} + \frac{dv}{dx},$$

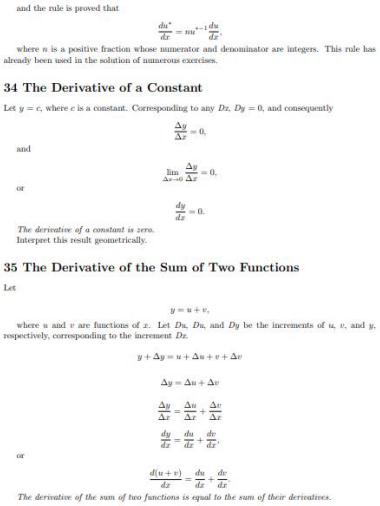
or

$$\frac{d(u+v)}{dx} = \frac{du}{dx} + \frac{dv}{dx}.$$

The derivative of the sum of two functions is equal to the sum of their derivatives.

Input image





and the rule is proved that

$$\frac{du^n}{dx} = nu^{n-1}\frac{du}{dx},$$

where n is a positive fraction whose numerator and denominator are integers. This rule has already been used in the solution of numerous exercises.

34 The Derivative of a Constant

Let $y = c$, where c is a constant. Corresponding to any Dx, $Dy = 0$, and consequently

$$\frac{\Delta y}{\Delta x} = 0,$$

and

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = 0,$$

or

$$\frac{dy}{dx} = 0.$$

The derivative of a constant is zero.

Interpret this result geometrically.

35 The Derivative of the Sum of Two Functions

Let

$$y = u + v,$$

where u and v are functions of x. Let Du, Du, and Dy be the increments of u, v, and y, respectively, corresponding to the increment Dx.

$$y + \Delta y = u + \Delta u + v + \Delta v$$

$$\Delta y = \Delta u + \Delta v$$

$$\frac{\Delta y}{\Delta x} = \frac{\Delta u}{\Delta x} + \frac{\Delta v}{\Delta x}$$

$$\frac{dy}{dx} = \frac{du}{dx} + \frac{dv}{dx}$$

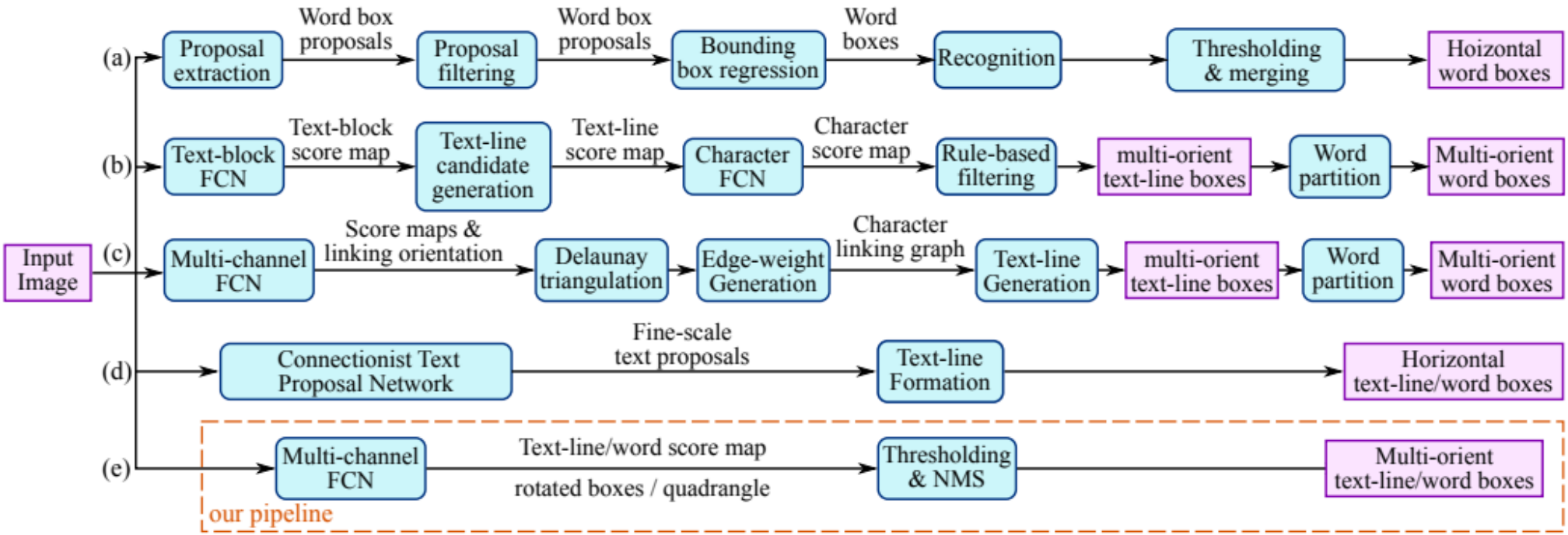
or

$$\frac{d(u+v)}{dx} = \frac{du}{dx} + \frac{dv}{dx}.$$

The derivative of the sum of two functions is equal to the sum of their derivatives.

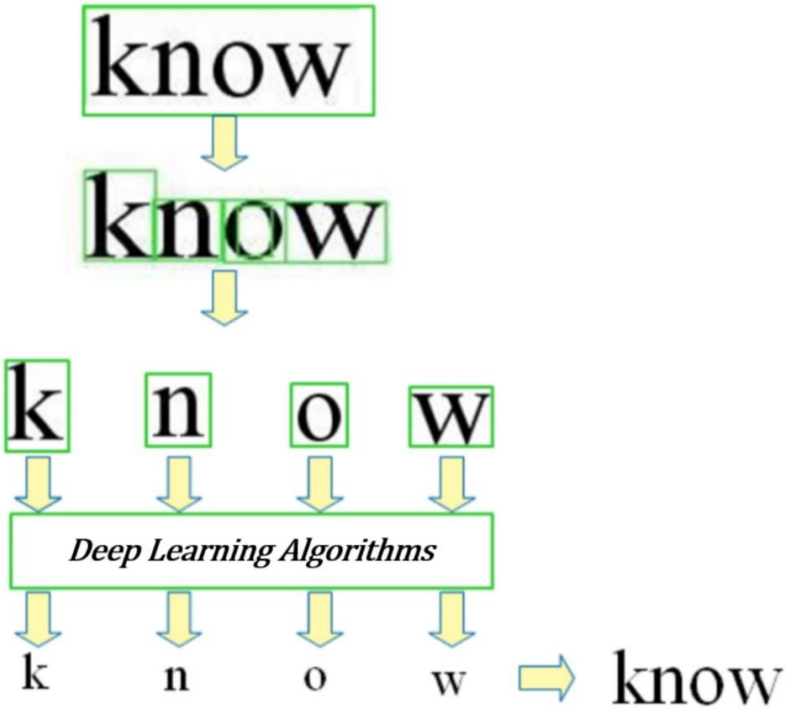
Output text

- Similar to techniques for Object detection or Segmentation
- A few characters make up a word or sentence, so you need to determine the minimum unit to detect it

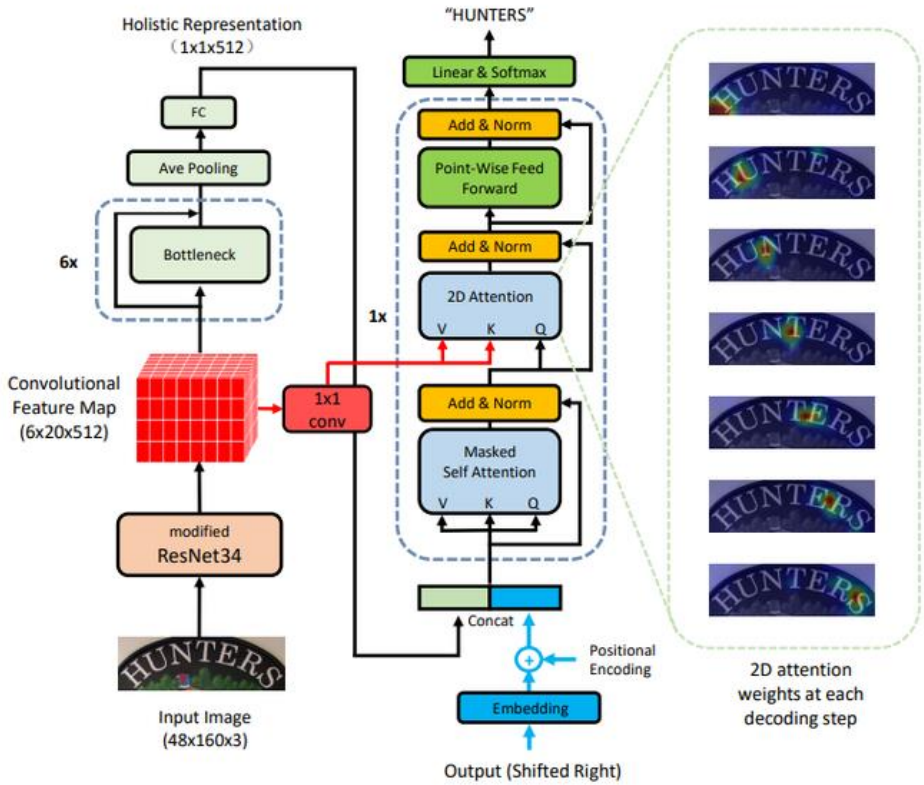


1. Introduction What is OCR: text recognition

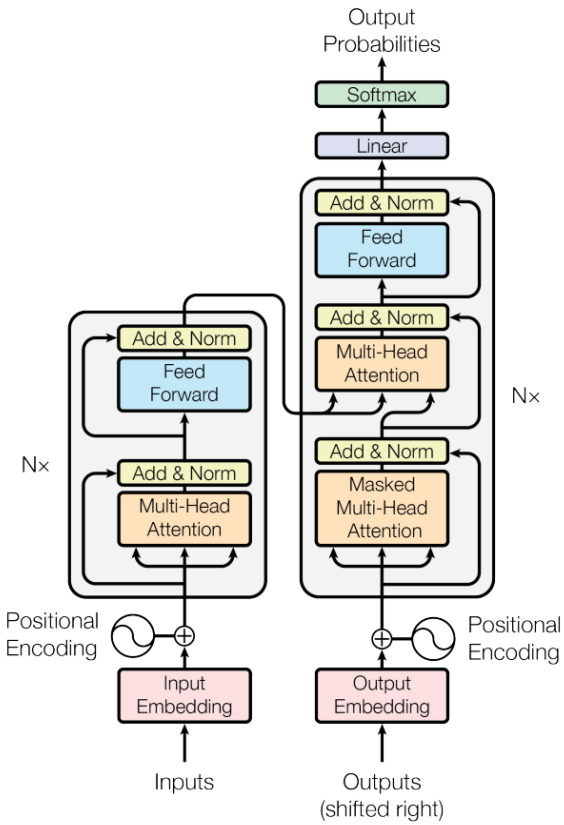
- Extracting feature, models are **learning several features** that distinguish letters.
- It find out what letters are in the input image.
- CNN + RNN = CRNN, CTC, TPS, Attention … etc.



- The development of the Transformer influenced text recognition
- Estimating the label in the first input character based on Attention
- Estimated label input again and estimate next label



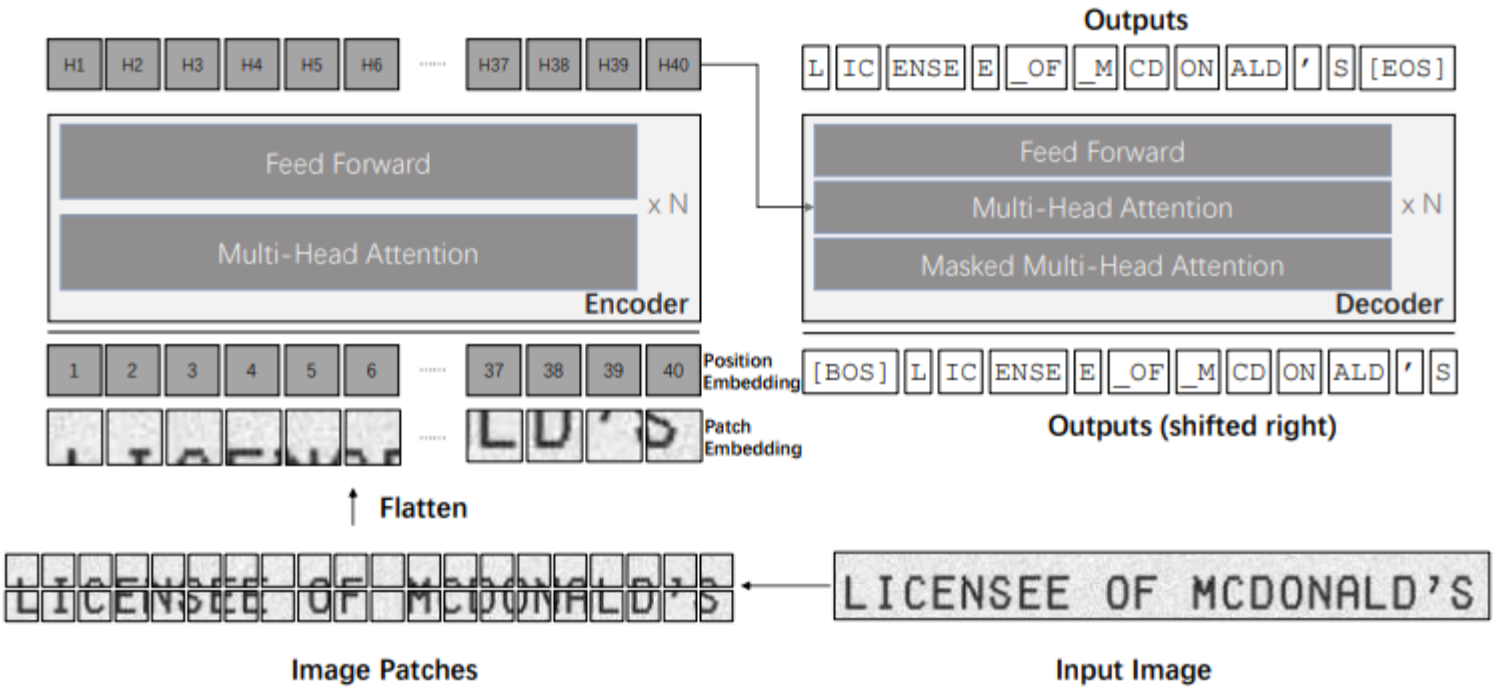
- Sequence to sequence model of **encoder-decoder** structure
- Consisting of **Attention**



2. Related Studies

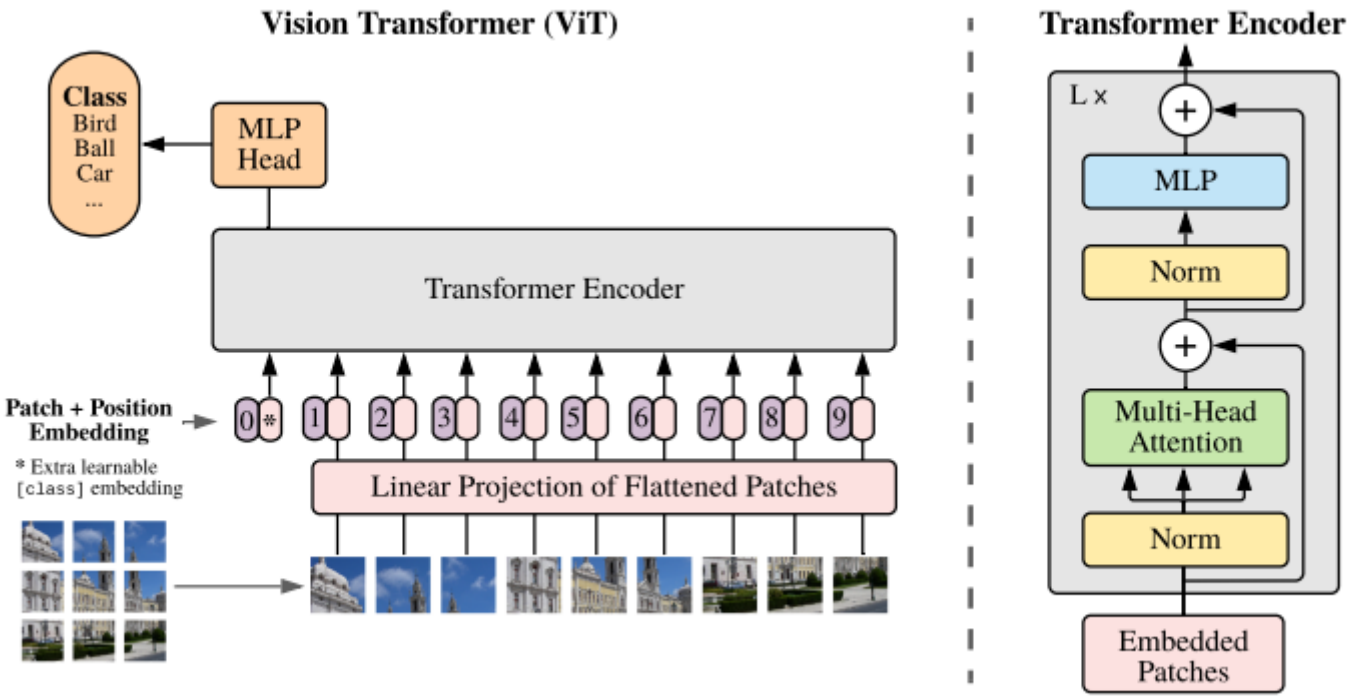
- ❖ It is many Transformer-related studies
- ❖ TrOCR, Donut, Nougat, … etc.

- End-to-end Transformer En-Decoder model
- It use well-trained image models and pre-trained language models



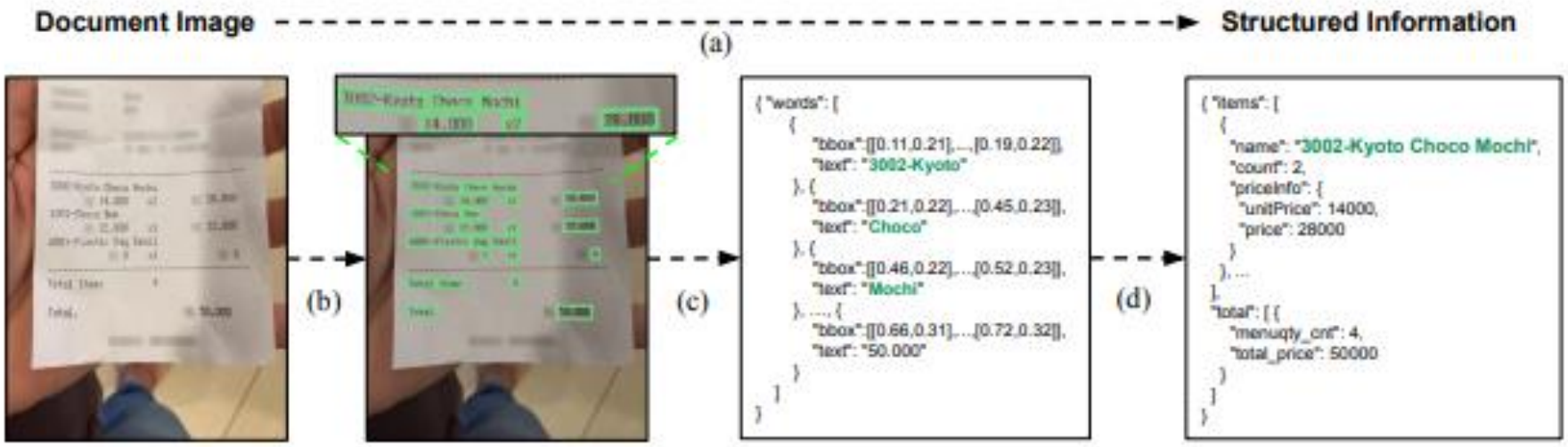
2. Related Studies TrOCR

- It used ViT as an encoder
- ViT, Vision Transformer, splits the input images into Patches (16x16 splited image)
- Stride and Pooling effect is weak due to lack of window sliding
- As image size gets bigger, so Patches increase more, self-Attention is slower

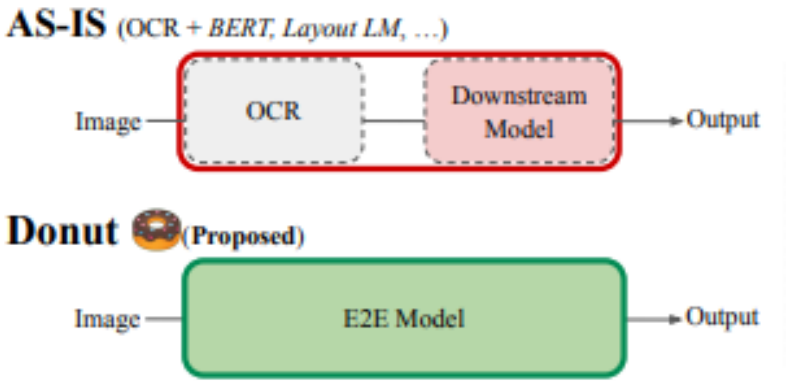


❖ Existing VDU System

- ❖ Architecture that relies on isolated OCR modules to extract text
- ❖ OCR is expensive and it is not always available
- ❖ OCR errors negatively influence subsequent processes



- End-to-end Transformer En-Decoder model
- It is not dependent on other OCR modules
- End-to-end structure that maps directly from raw input images to outputs



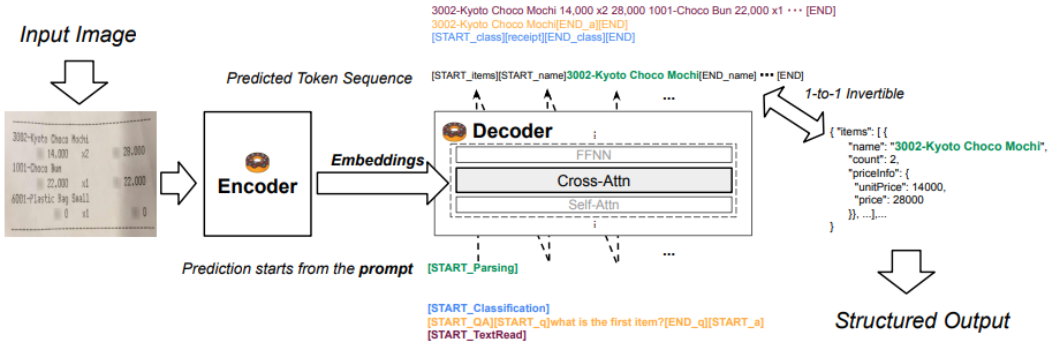
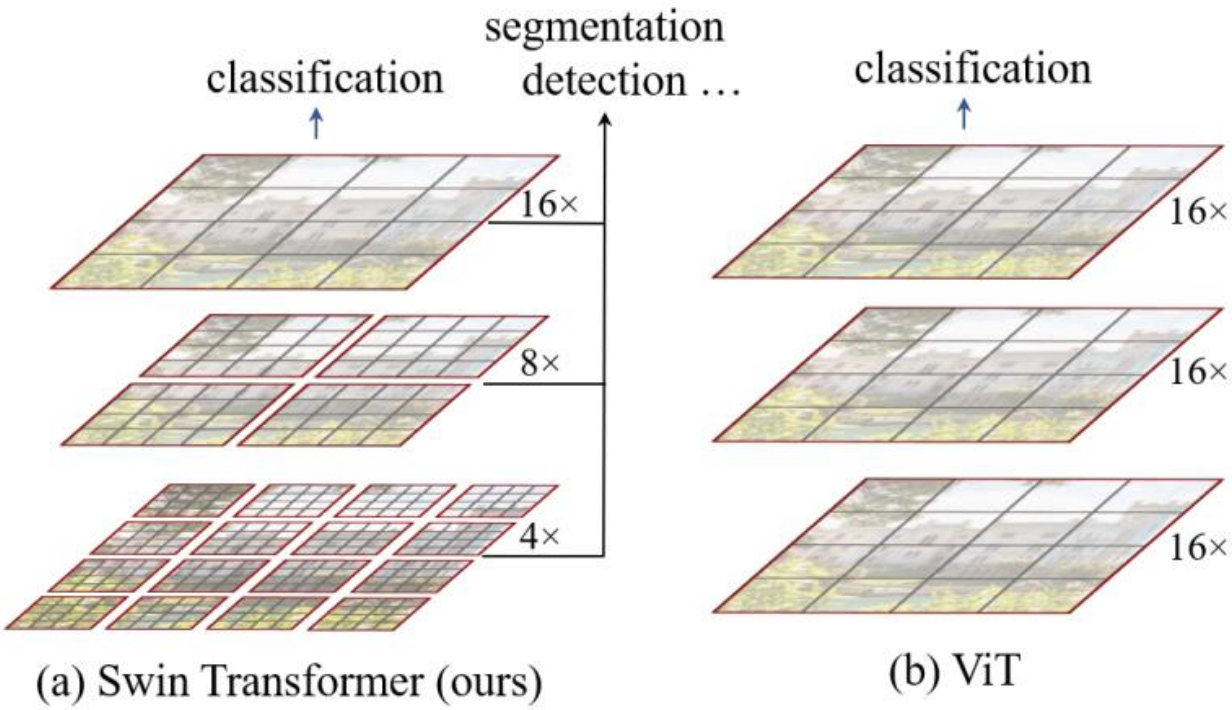


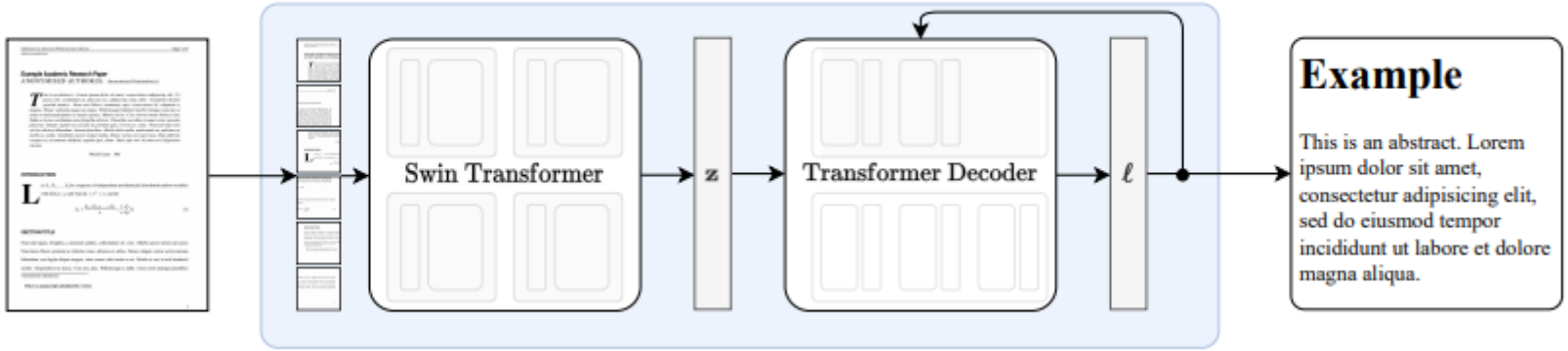
Figure 3: The overview of **Donut**. The encoder maps a given document image into embeddings. With the encoded embeddings, the decoder generates a sequence of tokens that can be converted into a target type of information in a structured form.

- In this study, Swin-Transformer is used
- Swin-transformer is one of ViT
- It has both stride and pooling effects through the window
- Previously Patches that were squared by the image size
- The amount of Patches is very small because it limited in the window



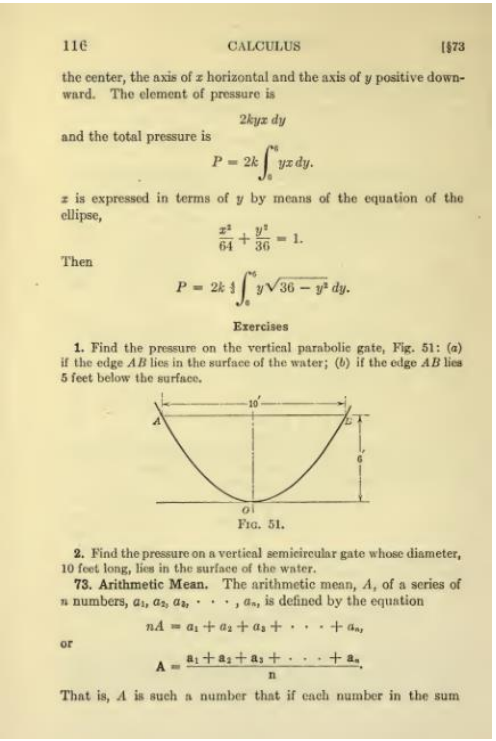
2. Related Studies Nougat

- Neural Optical UnderstandinG for Academic documenTs
- It use swin-transformer model
- It processes scientific documents into a markup language
- It is built on the Donut architecture



2. Related Studies Nougat

- In a scientific research article, there are three distinct types of text
  - Plain text
  - Mathematical expressions
  - Tables
- Numbers and punctuation have **ambiguity** where text begins and ends



116 CALCULUS [§73

the center, the axis of x horizontal and the axis of y positive downward. The element of pressure is

$$2kyx\,dy$$

and the total pressure is

$$P = 2k\int_0^6 yx\,dy.$$

x is expressed in terms of y by means of the equation of the ellipse,

$$\frac{x^2}{64} + \frac{y^2}{36} = 1.$$

Then

$$P = 2k\,3\int_0^6 y\sqrt{36 - y^2}\,dy.$$

Exercises

1. Find the pressure on the vertical parabolic gate, Fig. 51: (*a*) if the edge AB lies in the surface of the water; (*b*) if the edge AB lies 5 feet below the surface.

FIG. 51.

2. Find the pressure on a vertical semicircular gate whose diameter, 10 feet long, lies in the surface of the water.

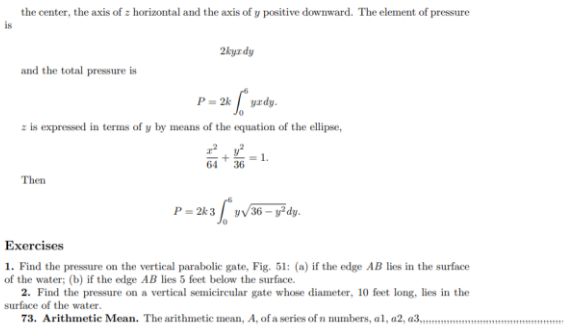
73. Arithmetic Mean. The arithmetic mean, A, of a series of n numbers, $a_1, a_2, a_3, \cdots, a_n$, is defined by the equation

$$nA = a_1 + a_2 + a_3 + \cdots + a_n,$$

or

$$A = \frac{a_1 + a_2 + a_3 + \cdots + a_n}{n}.$$

That is, A is such a number that if each number in the sum



the center, the axis of z horizontal and the axis of y positive downward. The element of pressure is

$$2kyx\,dy$$

and the total pressure is

$$P = 2k\int_0^6 yx\,dy.$$

z is expressed in terms of y by means of the equation of the ellipse,

$$\frac{x^2}{64} + \frac{y^2}{36} = 1.$$

Then

$$P = 2k\,3\int_0^6 y\sqrt{36 - y^2}\,dy.$$

Exercises

1. Find the pressure on the vertical parabolic gate, Fig. 51: (a) if the edge AB lies in the surface of the water; (b) if the edge AB lies 5 feet below the surface.

2. Find the pressure on a vertical semicircular gate whose diameter, 10 feet long, lies in the surface of the water.

73. Arithmetic Mean. The arithmetic mean, A, of a series of n numbers, $a1$, $a2$, $a3$,.....

3. Motivation and Ongoing

❖ Motivation

- ❖ It is not easy to create R&D corpora dataset
- ❖ Words and formulas used in other fields are different each other.
- ❖ It is very difficult to create dataset by hand

❖ Ongoing works

- ❖ Need to create dataset by extracting computer readable text from PDF related to R&D
- ❖ First, extract only text data to create corpora dataset
- ❖ Then, extract tables, formula, images and etc. for non-text dataset
- ❖ Collect R&D PDF and create only text dataset based on Nougat model

Thank you