"Решение задачи VQA посредством нейросетевой модели"

Дипломная работа Базовая кафедра распознавания изображений и обработки текста ABBYY

Студент: Даниял Алиев
Руководитель: Артур Бегаев

Соревнование Receipt-VQA-2023

Входные данные:

- Изображение чека
- Вопрос
- Категория вопроса ("amount", "count", "ratio")
- Валюта, используемая в чеке
- Список операция необходимый для вычисления результата

Выходные данные:

Ответы на вопросы в формате float

Метрика:

$$\mathrm{MASE} = \mathrm{mean}\left(\frac{|e_j|}{\frac{1}{J}\sum_{j=1}^{J}|Y_j - \bar{Y}|}\right) = \frac{\frac{1}{J}\sum_j |e_j|}{\frac{1}{J}\sum_j |Y_j - \bar{Y}|}$$

Соревнование Receipt-VQA-2023

Input image

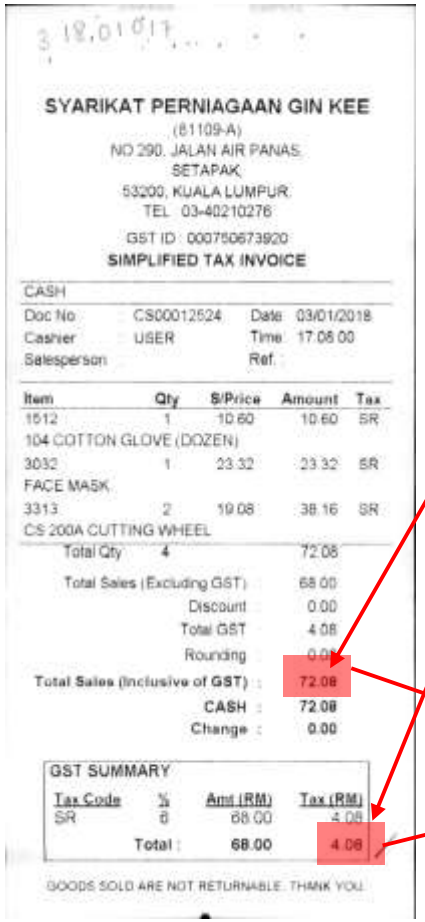
Input question:
How much should be paid?

Output:
6.2

receipt_currency: Malaysian ringgit
question_category: amount
operations: amount

Соревнование Receipt-VQA-2023

Input image



Input answer:

How much is tax to total amount ratio?,

Receipt_currency: Malaysian ringgit

question_category: ratio

Operations: division

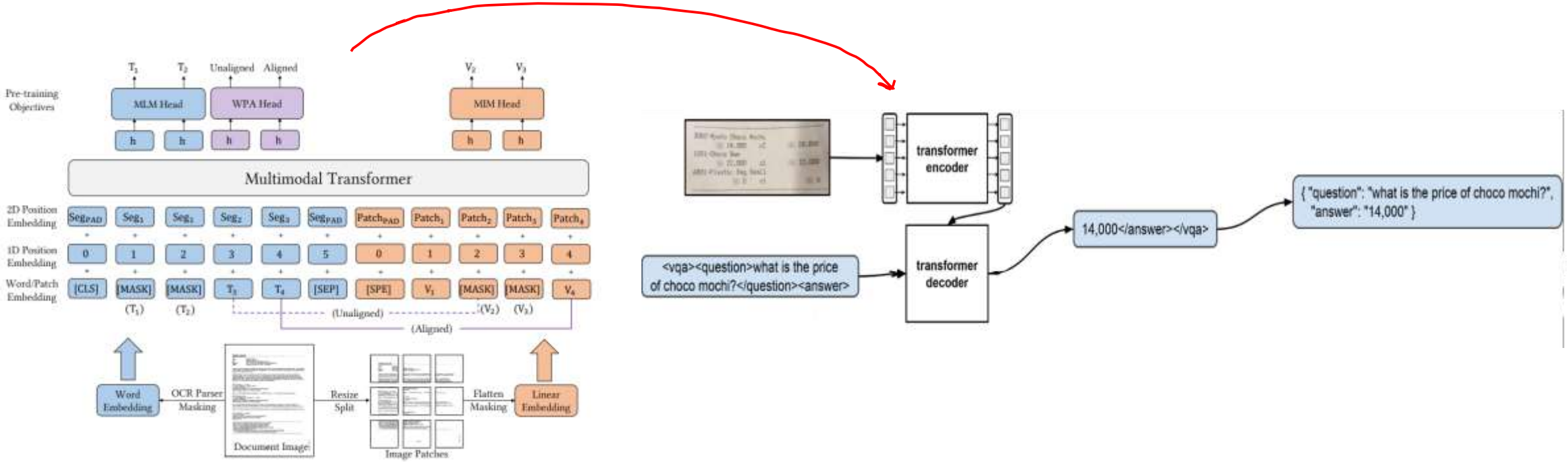
Output:

4.08/72.08=0.0566

Работа модели на категории “amount”

Категория amount включает в себя вопросы, ответы на которые содержатся в самом изображении – не требуется никаких дополнительных вопросов.

В качестве модели для fine-tuning берем модель **LayoutLMv3** (LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking)



Работа модели на категории “ratio”

Ratio определяется как отношение двух чисел, каждое из которых можно найти, используя уже обученную LayoutLMv3 модель

What is net total and total amount ratio?

RESTORAN WAN SHENG
002043319-W
No.2, Jalan Temenggung 19/9,
Seksyen 9, Bandar Mahkota Cheras,
43200 Cheras, Selangor
GST REG NO: 001335787520

Tax Invoice

INV No.: 1057981 Cashier: Thandar
Date : 23-03-2018 13:28:55

Description	Qty	U.price	Total	TAX
Teh (B)	1 x	2.20	2.20	SR
Cham (B)	1 x	2.20	2.20	SR
Bunga Kekwa	1 x	1.70	1.70	SR
Take Away	3 x	0.20	0.60	SR

Total QTY: 6
Total (Excluding GST): 6.33
GST payable (6%): 0.37
Total (Inclusive of GST): 6.70
TOTAL : 6.70

CASH : 6.70

GST Summary		Amount(RM)	Tax(RM)
SR	(@ 6%)	6.33	0.37

lmV3

lmV3

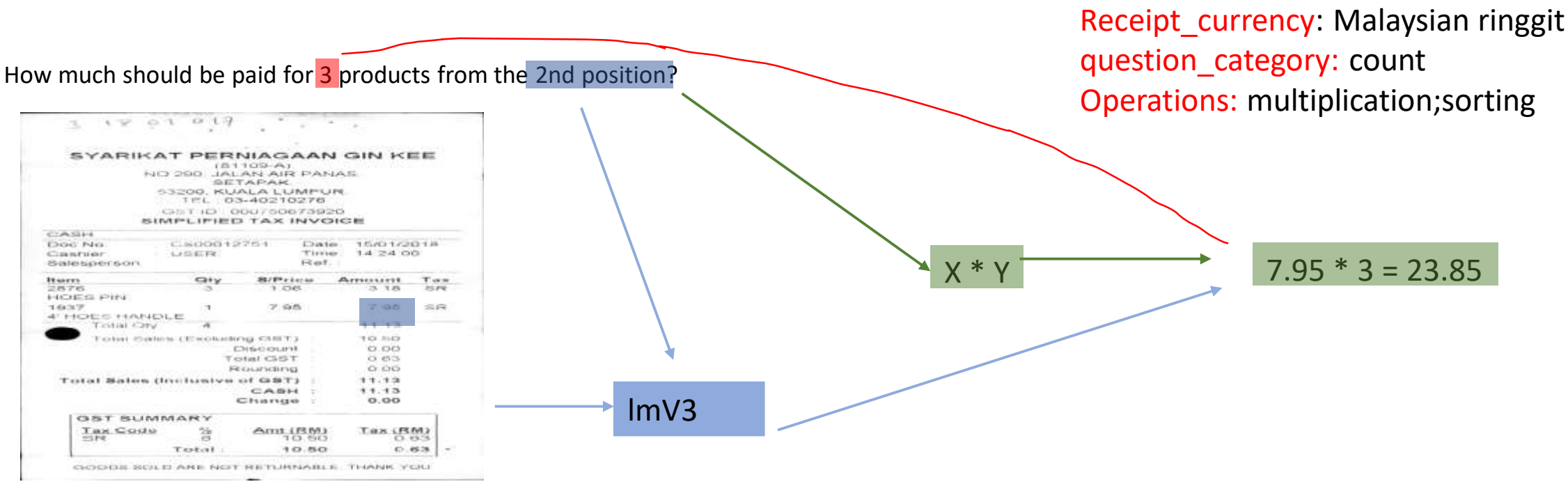
output

6.33/6.9 = 0.944

Работа модели на категории “count”

Категория достаточно объемная – рассмотрим для начала вопросы, которые подразумевают некоторые арифметические операции. Тогда алгоритм будет следующим:

1) Определить шаблон выражения, которое даст ответ на вопрос (с помощью operations)
2) Находим неизвестные с помощью предобученной LayoutLMv3
3) Подставляем числа в выражение и получаем результат



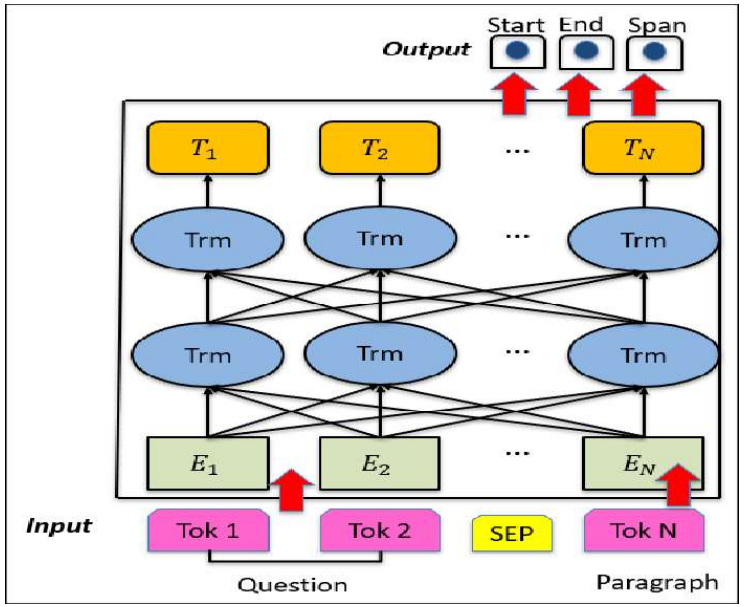
Извлечение информации из текста

Для ответа на вопросы, требующие арифметические операции, мы извлекали ключевые слова из вопроса и использовали их для подачи в LayoutLmv3.

Для выполнения данной задачи нужно решить задачу question answering – для этого нужно предварительно подготовить ключевые слова для каждого из вопросов, используя регулярные выражения (все вопросы одной категории очень похожи)

Далее снова воспользуемся трансформером с Hugging Face – BertForQuestionAnswering - для fine-tuning на наших данных.

How many goods in the 1st position can be purchased for 22?



1st position

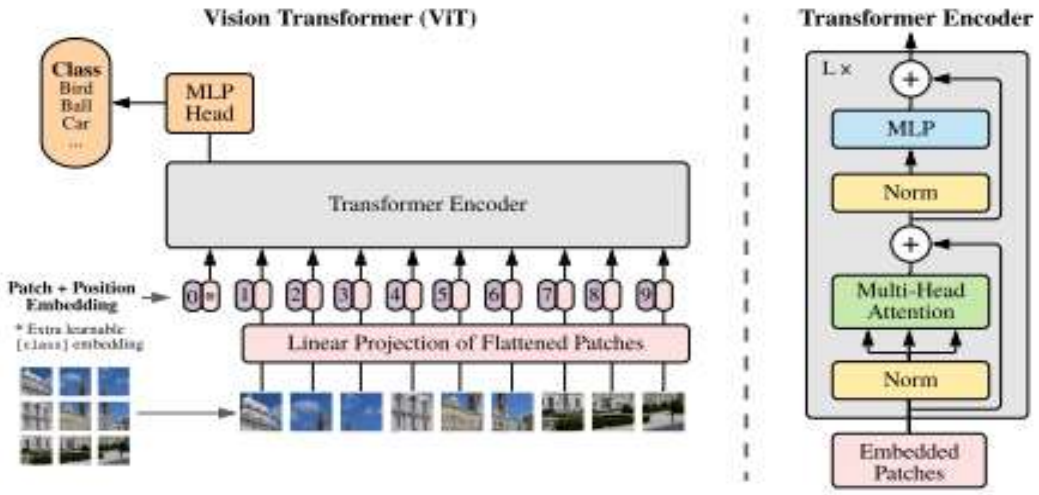
22

Работа модели на категории “count”

В данной категории присутствуют так же вопросы, ответы на которые не получаются вычислением арифметический выражений, пример: How many positions are free of cost?

Ответы на такие вопросы дать достаточно тяжело – по итогу лучший вариантом оказалось использование ViTModel (https://huggingface.co/docs/transformers/model_doc/vit#transformers.ViTModel)

Используем last_hidden_state, который подается на линейный слой и Softmax – MLP Head (используем классификацию на 8 элементов – на train выборке этот диапазон покрывается 90% ответов)



Результаты

VQA Track

Participant	Best Submission Name	MASE Total	MASE Amount	MASE Count	MASE Ratio
daniyallaiev	submission0	0.7874	0.7979	0.5718	0.9926
BASELINE (abegaev)	answer	0.8786	0.8068	0.8291	1.0000

Анализ полученных результатов:

- На amount результаты практически идентичны, потому что baseline модель так же использовала **LayoutLMv3** на этой категории
- Видим существенное улучшение на категории count за счет грамотного применения Bert и предобученной на прошлой категории модели
- На ratio результат так же превосходит baseline

Проблема модели:

- На многих изображениях датасета OCR работает некорректно и зачастую “не видит” ответ на вопрос, после чего дальнейшие действия уже бесполезны

Возможные решения:

- Самостоятельно обучение OCR на исходном датасете (требует дополнительной разметки)
- Предварительная обработка изображений – применение фильтров и т.д.

Дальнейшие действия

- Улучшение работы модели – корректировка работы OCR, применение предварительной обработки изображений, применение новых идей архитектуры модели
- Проверить работу модели на **TAT-DQA** датасете (https://github.com/NExTplusplus/TAT-DQA)
- Проанализировать полученные результаты, найти уязвимости модели
- Снова постараться улучшить модель

Спасибо за внимание!