**USING VISION LANGUAGE MODELS FOR OPTIMIZING LIGHT INDUSTRY TASKS**

Сидельников Никита, ИТМО

**Лёгкая промышленность сегодня**

Прежде чем перейти к описанию проделанной работы, кратко опишем область и существующие в ней сегодня проблемы.

Лёгкая промышленность – это совокупность отраслей промышленности, осуществляющих первичную обработку различных видов сырья и производство из них предметов массового потребления.

Сегодня данная отрасль в нашей стране значительно отстает от многих других. Это происходит из-за достаточно большого списка проблем. Одними из основных являются:

* Кадровый дефицит;
* Технологическая и техническая отсталость;
* Низкий уровень научной и инновационной деятельности;
* Отсутствие сырьевой базы и комплектующих.

Для решения данных проблем можно разрабатывать новые технологии и находить способы применения уже существующих. Например, автоматизация производства может значительно упростить процессы на предприятиях, а благодаря быстрому развитию технологий искусственного интеллекта её можно внедрить на новые этапы производственного цикла, где этого автоматизация не применялась. Моя работа посвящена именно этой теме.

**Постановка задачи**

В рамках магистерской диссертации я занимался разработкой MVP системы автоматизированного контроля производственного потока швейного предприятия. Основная задача системы – формирование статистики на разных этапах производственного потока в полностью автоматическом режиме.

Все производство можно разделить на этапы, представленные на картинке:



Рисунок 1 – Схема производственного потока швейного предприятия

В рамках MVP работа была сфокусирована на этапах *раскроя* и *ВТО*. Некоторые задачи, которые было необходимо решить:

* Классификация изделий по типу и цвету на этапе ВТО
* Сегментация и подсчет занимаемой лекалами площади

*Задача на школу*: проверить, возможно ли решить поставленные задачи в zero-shot режиме с помощью Vision Language Models.

**Данные**

Для решения задач были собраны несколько наборов данных:

Таблица 1 – Числовое описание данных

|  |  |  |
| --- | --- | --- |
| **Задача** | **Кол-во изображений** | **Кол-во классов** |
| Детектирование одежды | ~2800 | 2 |
| Классификация одежды по цвету | ~2800 | 5 |
| Сегментация лекал и полотна ткани | ~3200 | 2 |

Важно подчеркнуть, что для разметки данных была применена автоматизация. Сервис Roboflow предлагает инструменты для автоматической разметки данных в задачах сегментации с использованием моделей семейства SAM.

Для оценки качества работы будем использовать не все данные, а только 1000 примеров. Распределение в задаче детектирования составило 40/60 между классами футболка/толстовка. 20/15/20/30/15 между классами черный/белый/бежевый/коричневый/розовый.

**Гипотезы**

Сформулируем гипотезы, которые проверим в рамках проекта:

* VLM способна решить задачу классификации изделия в кадре по типу с точностью не меньше 0.9 (accuracy)
* VLM способна решить задачу классификации изделия в кадре по цвету с точностью не меньше 0.9 (accuracy)
* VLM способна решить задачу детектирования изделия в кадре с точностью не меньше 0.85 (IoU)
* VLM способна решить задачу сегментирования лекал в кадре с точностью не меньше 0.8 (IoU)
* VLM способна решить несколько поставленных выше задач за одно обращение к ней

**Список выбранных моделей**

Ниже представлен список использованных для проверки гипотез моделей. Статьи, описывающие архитектуры и другие особенности, можно найти по ссылкам.

* [MINICPM-LLAMA3-V-2\_5](https://huggingface.co/openbmb/MiniCPM-Llama3-V-2_5)
* [VILT-B32-FINETUNED-VQA](https://huggingface.co/dandelin/vilt-b32-finetuned-vqa)
* [BLIP-VQA-CAPFILT-LARGE](https://huggingface.co/Salesforce/blip-vqa-capfilt-large)
* [FLORENCE-2-LARGE](https://huggingface.co/microsoft/Florence-2-large)
* [INTERNVL-2](https://huggingface.co/OpenGVLab/InternVL2-8B)
* [PALIGEMMA](https://huggingface.co/docs/transformers/main/en/model_doc/paligemma)

Далее дадим обобщенные результаты качества работы моделей для каждой из задач, а затем представим суммированную таблицу с числовыми метриками.

**Решение нескольких задач за один запрос**

В рамках данной гипотезы было интересно проверить сможет ли какая-то модель решить сразу несколько описанных выше задач, и одна такая модель нашлась! Ей оказалась – minicpm-llava3-v2.5. Однако, стоит отметить, что при таком формате взаимодействия модель не соблюдает просьбу о форматировании ответа, что усложняет дальнейшую пост-обработку результатов.

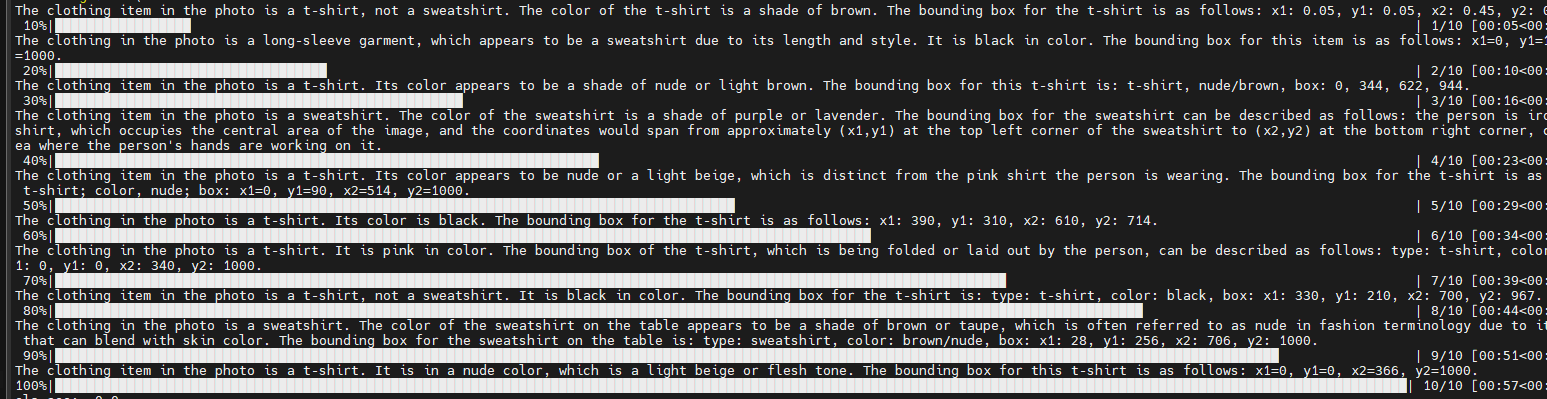


Рисунок 12 – Пример ответа модели minicpm на запрос: 'There is person working with piece of clothing on the photo. What type of clothing is it, tshirt or sweatshirt? What is it color, black, brown, nude, pink or white? What is bounding box of it? Give answer in format: type, color, box: x1, y1, x2, y2.'

**Классификация изделий по типу**

Данная задача состоит в том, чтобы определить изделие какого типа находится в руках у работника – футболка или толстовка. С решением данной задачи справилась только часть моделей. Качество сильно зависит от входного промпта, важно отметить, что много времени уходит на его подбор. Необходимой точности, в конечном итоге, добиться не удалось.

Примеры ответы нескольких моделей представлены ниже на рисунках.

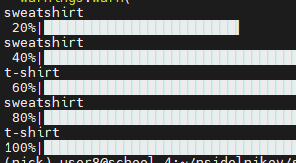


Рисунок 1 – Пример ответа модели minicpm на запрос: 'There is piece of clothing in the hands of person. What is it? tshirt or sweatshirt. Give short answer, type only.'

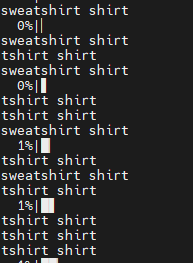


Рисунок 2 – Пример ответа модели vilt на запрос: 'What is the biggest piece of clothing on the picrure, tshirt or sweatshirt?'. В левом столбце содержится истинное значение, в правом предсказанное.

**Классификация изделий по цвету**

Данная задача состоит в том, чтобы сказать какого цвета изделие находится в руках у сотрудника. Набор данных содержит в себя изделия черного, коричневого, бежевого, белого и розового цветов. С данной задачей справились те же модели, что и с предыдущей. Достичь высокого качества метрики не удалось. Несколько примеров ниже.

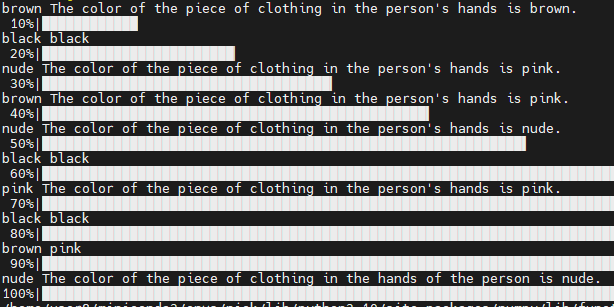


Рисунок 3 – пример ответа модели minicpm на запрос: 'There is piece of clothing in the hands of person. What is its color? Choose one from the list: black, white, pink, brown, nude. Answer shortly, just color name.'



Рисунок 4 – Пример ответа модели internvl2 на запрос: '<image>\nWhat color is piece of clothing person workin with? Choose from black, brown, nude, white and pink.'

**Детектирование**

Данная задача состоит в том, чтобы получить ограничивающую рамку изделия, с которым работает сотрудник. Классификация изделия была опущена, и выполнена отдельно, так как на первом этапе экспериментов оказалось, что модель плохо отвечает на сложный вопрос.

С детектированием справились немного другие модели. Необходимого качества все же достигнуто не было. Примеры запросов и ответов, а также визуализации результатов представлены ниже.

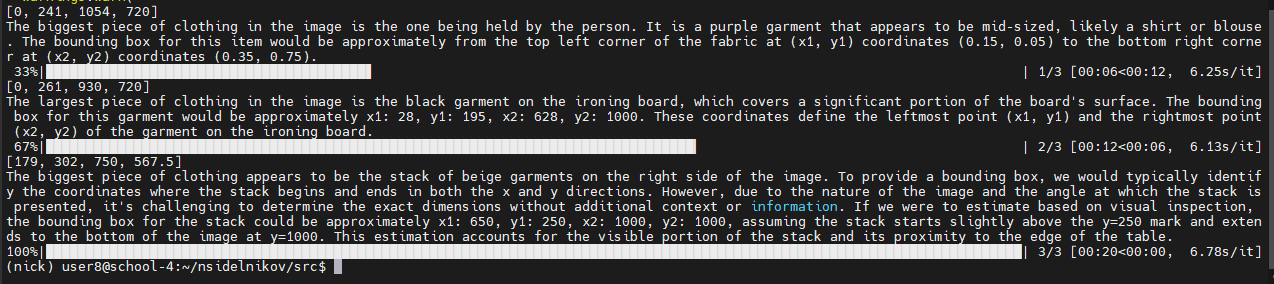


Рисунок 5 – Пример ответа модели minicpm на запрос: Detect the biggest piece of clothing, give bounding box in format x1, y1, x2, y2

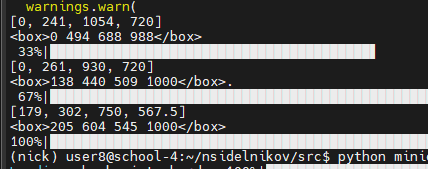


Рисунок 6 – Пример ответа модели minicpm на запрос: Detect the biggest piece of clothing on the picture and give its bounding box in format x1, y1, x2, y2. Answer shoud be short, just coordinates.

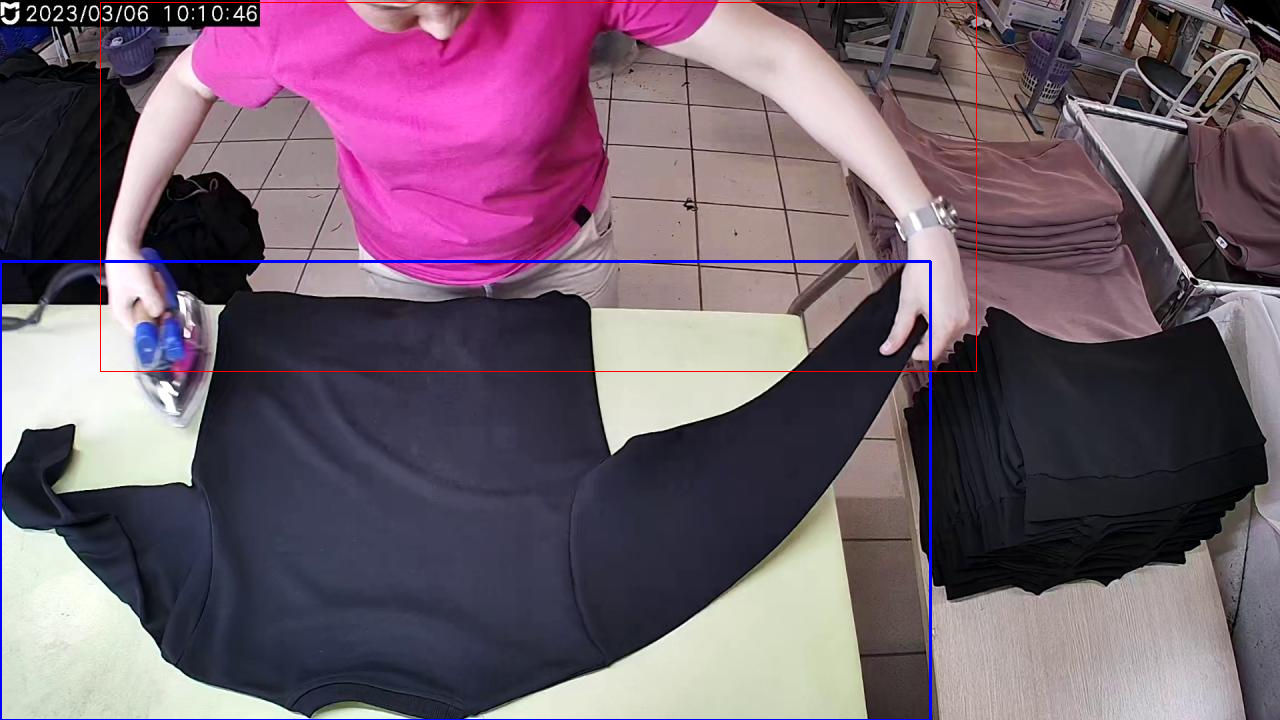


Рисунок 7 – пример плохого ответа florence2



Рисунок 8 – Пример хорошего ответа florence2

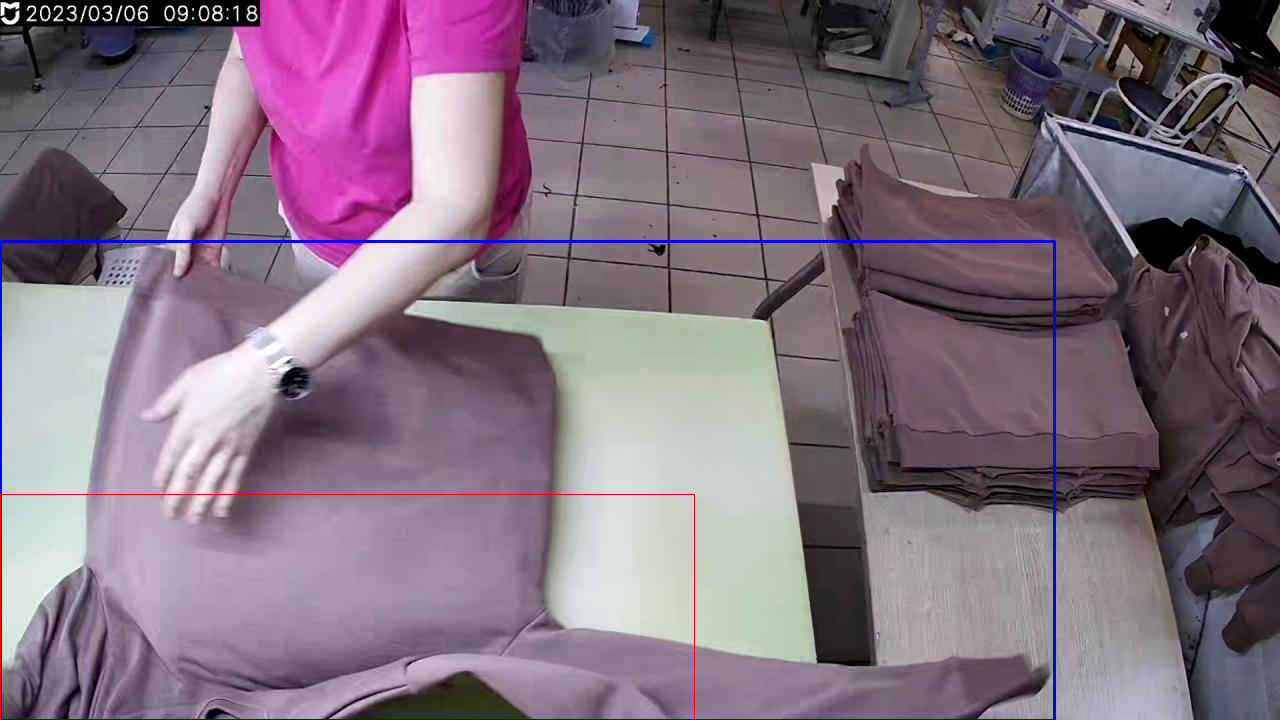


Рисунок 9 – Пример плохого ответа minicpm

**Сегментация**

В рамках данной задачи было необходимо сегментировать картонные лекала, разложенные на поверхности ткани. Ни одна из выбранных моделей не справилась с задачей сегментации. Большинство выдавало просто описательный текст картинки, инструкции для решения задачи сегментации или отвечало, что не может решить данную задачу. Florence2, имеющая функционал для сегментации, не смогла понять, что такое лекала, и где они расположены. Получить хоть как-то правильных масок не удалось, поэтому метрики сегментации для всех моделей будем считать нулевыми и не вынесем их в финальную таблицу. Примеры ответов моделей, а также изображения представлены ниже.



Рисунок 10 – Пример изображения с зоны раскроя

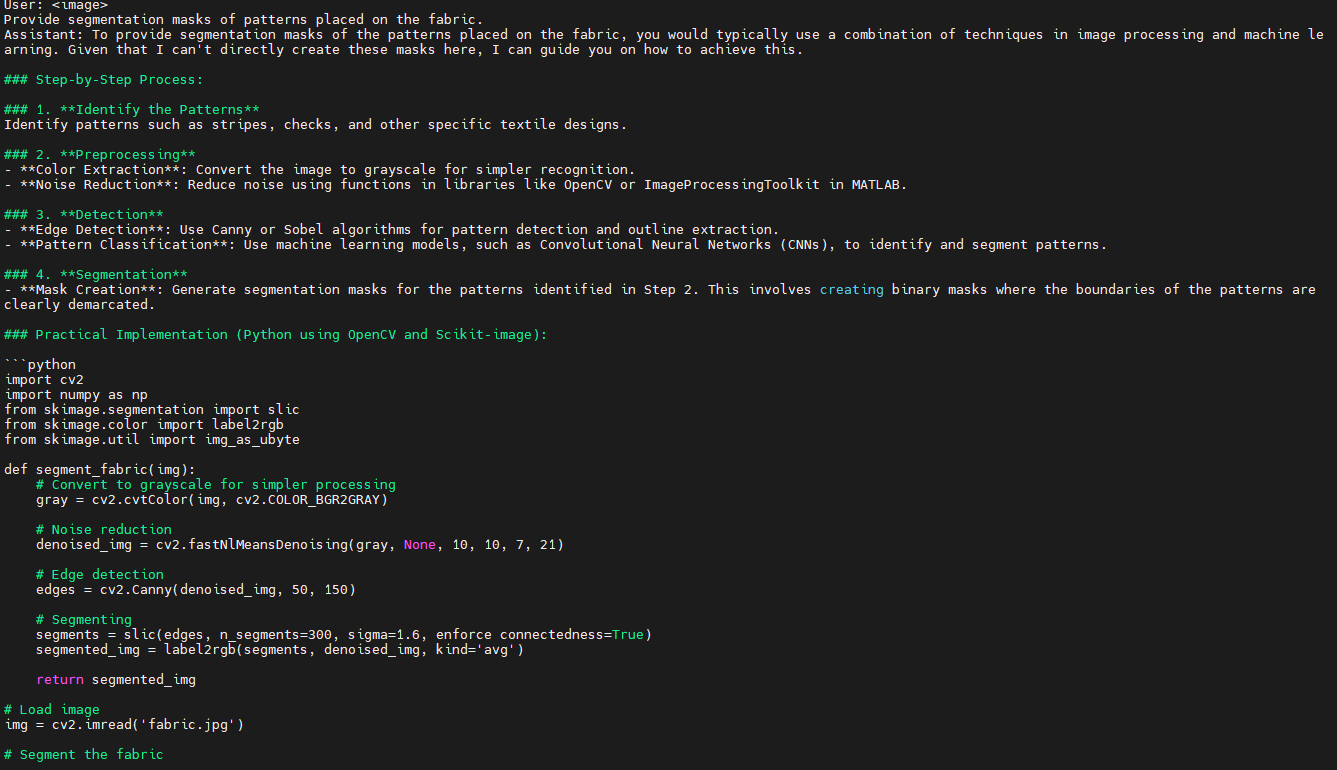


Рисунок 11 – Пример ответа модели intervl2

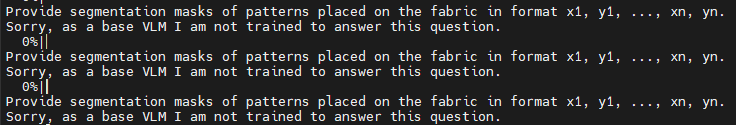


Рисунок 12 – Пример ответа модели paligemma

**Числовые результаты**

В таблице ниже представлены метрики

Таблица 2 – Метрики

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Наименование** | **Количество параметров** | **Классификация цвета (accuracy)** | **Классификация типа (accuracy)** | **Детектирование (IoU)** | **m** |
| [VILT-B32-FINETUNED-VQA](https://huggingface.co/dandelin/vilt-b32-finetuned-vqa) | 117.58M | 0.00 | 0.00 | - | - |
| [BLIP-VQA-CAPFILT-LARGE](https://huggingface.co/Salesforce/blip-vqa-capfilt-large) | 384.67M | 0.00 | 0.00 | - | - |
| [FLORENCE-2-LARGE](https://huggingface.co/microsoft/Florence-2-large) | 822.69M | - | - | **0.77** | - |
| [PALIGEMMA](https://huggingface.co/docs/transformers/main/en/model_doc/paligemma) | 2.92B | 0.53 | 0.47 | 0.25 | - |
| [INTERNVL-2](https://huggingface.co/OpenGVLab/InternVL2-8B) | 8.07B | **0.82** | 0.49 | 0.3 | - |
| [MINICPM-LLAMA3-V-2\_5](https://huggingface.co/openbmb/MiniCPM-Llama3-V-2_5) | 8.54B | 0.54 | **0.74** | 0.19 | + |

\*m – колонка, отражающая способность модели отвечать на несколько вопросов за одно обращение

**Вывод**

В результате проведённого исследования подтверждение нашла всего одна гипотеза – VLM способна решить несколько простых задач CV за одно обращение. Однако, хорошие результаты работы некоторых моделей (выделены жирным шрифтом) показывают их потенциал применения в качестве помощников в узком классе задач, например, как помощника в разметке для задачи детектирования. Дальнейшее развитие и исследование данной тематики можно считать актуальным.

Также, стоит отметить, что в текущих реалиях рассматривать какую-либо модель данного класса в качестве решения для real-time потоковой обработки данных, которая необходима в данном проекте – невозможно. Инференс моделей занимает много времени и требователен к ресурсам.