УДК 004.9

РЕШЕНИЕ ЗАДАЧИ РАСПОЗНАВАНИЯ И ИНТЕРПРЕТАЦИИ ИЗОБРАЖЕНИЙ ОБЪЕКТОВ НА ОСНОВЕ ГАРНИТУРЫ СМЕШАННОЙ РЕАЛЬНОСТИ

Андрианова Е.Г., Демидов Н.А.

Федеральное государственное бюджетное образовательное учреждение высшего образования «МИРЭА — Российский технологический университет», 119454, Российская Федерация, г. Москва, пр-т Вернадского, 78, e-mail: andrianova@mirea.ru, nick.a.demidov@rambler.ru

В статье рассмотрена задача распознавания и интерпретации изображений объектов на основе гарнитуры смешанной реальности Microsoft Hololens 2 в контексте работы по оптимизации процесса идентификации комплектующих персонального компьютера (КПК). Для решения поставленной задачи разработаны программные средства (ПС), имеющие клиент-серверную архитектуру. Клиентская часть ПС расположена на гарнитуре Microsoft Hololens 2 и отвечает за графический интерфейс, формирование снимков КПК, а также – отправку запросов на сервер. Серверная часть ПС содержит модуль аннотирования изображений, модуль перевода текстовых описаний, а также – модуль базы данных с информацией о названиях КПК, их текстовых описаниях и верифицирующих изображениях. Модули аннотирования и перевода текстов основаны на применении моделей нейронных сетей глубокого обучения – BLIP и T5 соответственно, представляющих собой модели-трансформеры. При этом для модели BLIP на наборе данных, содержащем примеры из предметной области в виде пар «изображение – аннотация», выполнено дообучение, позволившее в процессе распознавания изображений формировать точные аннотации КПК. Разработанные ПС могут быть использованы при выполнении инвентаризации КПК с использованием гарнитуры Microsoft Hololens 2 для оптимизации процесса их идентификации, а также при обучении персонала, работающего с КПК.

Ключевые слова: программные средства, Microsoft Hololens 2, нейронная сеть, трансформер, BLIP, T5, набор данных, предобучение, дообучение, аннотирование изображения, перевод текстового описания, комплектующая персонального компьютера.

SOLUTION TO THE TASK ON RECOGNITION AND INTERPRETATION OF OBJECT IMAGES BASED ON THE MIXED REALITY HEADSET

Andrianova E.G., Demidov N.A.

Federal State Budgetary Educational Institution of Higher Education "Moscow State University of Information Technologies, Radio Engineering and Electronics" (MIREA), 119454, Russia, Moscow, Vernadsky Avenue, 78, e-mail: andrianova@mirea.ru, nick.a.demidov@rambler.ru

The paper considers the task of recognition and interpretation of object images based on the mixed reality headset Microsoft Hololens 2 in the context of optimization of the process of identification of personal computer components (PCCs). To solve the task, the software tools (ST) with client-server architecture have been developed. The client part of the ST is located on the Microsoft Hololens 2 and is responsible for the graphical interface, generation of the PCC images, as well as sending requests to the server. The server part of the ST contains image annotation module, text description translation module, and also the database module with information about the PCC names, their text descriptions and verification images. The annotation and text translation modules are based on the application of deep learning neural network models such as BLIP and T5, respectively, which are transformer models. The fine-tuning of the BLIP model is performed on the dataset containing examples from the subject area in the form of pairs "image – annotation": it allowed to form accurate annotations of the PCCs in the process of image recognition. The developed STs can be used to inventorize the PCCs using Microsoft Hololens 2 to optimize the process of their identification, as well as in the training of personnel working with the PCCs.

Keywords: software tools, Microsoft Hololens 2, neural network, transformer, BLIP, T5, dataset, pre-training, fine-tuning, image annotation, text description translation, personal computer component.

Введение

В современном информационном обществе задачи анализа изображений регулярно решаются в таких сферах деятельности человека, как медицина, промышленность, бизнес, образование, наука и др. При этом в процессе анализа изображений может возникать потребность в их аннотировании, то есть в описании содержания изображений, с последующим переводом аннотаций и, возможно, имеющихся подробных текстовых описаний, соответствующих этим аннотациям, на другие языки при необходимости. Одна такая задача заключается в анализе и аннотировании изображений КПК. Эта задача является важной и актуальной для индустрии электроники и производства электронных устройств. В настоящее время эффективное решение этой задачи возможно с привлечением технологий искусственных нейронных сетей, которые представляют собой мощный инструмент глубокого обучения, способный адаптироваться к различным типам данных и задачам, включая задачи анализа и обработки как изображений, так и естественного языка. Благодаря использованию технологий искусственных нейронных сетей можно автоматизировать как процесс анализа и аннотирования изображений, так и процесс перевода аннотаций и соответствующих им текстовых описаний на удобный для восприятия информации язык, обеспечив при этом высокую точность результатов.

Очевидно, что реализация вычислительного процесса, включающего в себя анализ изображений, их аннотирование, а также – перевод аннотаций и соответствующих им текстовых описаний на удобный для восприятия информации язык, на сервере с привлечением технологий облачных вычислений должна позволить снизить нагрузку на локальные устройства пользователя и обеспечить доступ к высокопроизводительным вычислительным ресурсам.

Отображение результатов анализа и интерпретации визуальной информации, полученных с привлечением технологий облачных вычислений, с использованием гарнитуры смешанной реальности – очков Microsoft Hololens 2 (рисунок 1) – позволяет интегрировать виртуальные объекты с реальным миром, обеспечивая пользователю наглядное представление о результатах анализа и возможность взаимодействия с ними в реальном времени [1]. В итоге реализуется инновационный подход к анализу визуальной информации, объединяющий в себе передовые инструменты машинного обучения, облачных вычислений и смешанной реальности.

Рисунок 1 – Очки Microsoft Hololens 2

Целью работы является разработка программных средств для решения задачи распознавания и интерпретации изображений объектов при работе с гарнитурой смешанной реальности Microsoft Hololens 2 посредством разработки ПС реализации облачных вычислений.

Разработанные ПС предполагается использовать при проведении инвентаризации КПК с использованием гарнитуры смешанной реальности Microsoft Hololens 2 для оптимизации процесса идентификации КПК, а также при обучении персонала, работающего с КПК.

В связи с этим необходимо:

– разработать интерфейс для работы по идентификации КПК с применением гарнитуры смешанной реальности;

– сформировать информацию о КПК, в частности, о процессорах, материнских платах, видеокартах и других КПК, в виде множественных пар «изображение – аннотация», описывающих одни и те же экземпляры КПК, запечатленные с разных ракурсов, при разном освещении, разном фоне, разном масштабе и т.п.;

– разработать ПС реализации облачных вычислений, обеспечивающие аннотирование изображений КПК для точного определения типа и модели КПК, а также – перевод текстовых описаний КПК на язык, удобный для восприятия конечным потребителем, с использованием моделей глубокого обучения.

1. Выбор моделей глубокого обучения

Для распознавания изображений объектов с последующим их аннотированием и переводом тестовых описаний, соответствующих аннотациям, на язык, удобный для восприятия пользователем, целесообразно применять модели глубокого обучения. Обучение нейронных сетей – долгий и ресурсозатратный процесс. При

разработке программного обеспечения для решения практически важных задач обучение соответствующих моделей нейронных сетей может занимать несколько недель или месяцев даже на очень мощном оборудовании.

В настоящее время при разработке моделей глубокого обучения активно используется трансферное обучение (transfer learning), применяемое в машинном обучении и представляющее собой повторное использование предварительно обученной модели для решения новой задачи. При трансферном обучении новая модель использует знания, полученные ранее в результате выполнения некоторой задачи, для улучшения обобщения в новой задаче.

1.1. Выбор модели глубокого обучения для аннотирования изображений

При разработке ПС целесообразно использовать какую-либо уже предобученную модель распознавания изображений объектов, реализующую концепцию «Image-to-Text» («изображение в текст») и обеспечивающую решение задачи аннотирования изображений. Такую модель можно дообучить на новых данных из целевой предметной области, связанной описанием КПК. Так как предобученная модель уже «выучила» много сложных зависимостей, затраты времени и ресурсов на её дообучение будут существенно меньшими, чем в случае, если бы обучение модели выполнялось «с нуля».

В 2022 году практически одновременно появились две конкурирующие модели, такие как GIT (GenerativeImage2Text) [2] и BLIP (Bootstraping Language-Image Pre-training) [3], основанные на трансформерах, представляющих собой архитектуры глубокого обучения, разработанные компанией Google в 2017 году [4].

Трансформеры используют в своей работе механизм многоголового внимания (Multi-Head Attention, MHA) и предполагают работу с функцией softmax, реализующей обобщение логистической функции для многомерного случая. Функция softmax преобразует некоторый вектор x размерности n в вектор z размерности n, у которого каждая координата z_i вычисляется как $z_i = \frac{e^{x_i}}{\sum_{k=1}^{n} e^{x_k}}$ $(i = \overline{1,n})$. При этом каждая координата z_i является вещественным числом из интервала [0,1], а сумма всех координат равна 1. В трансформерах текстовые данные преобразуется в числовые представления, называемые токенами, при этом каждый токен преобразуется в вектор посредством поиска в таблице встраивания слов. Каждый токен контекстуализируется в пределах контекстного окна с другими (немаскированными) токенами с помощью механизма MHA, позволяющего усиливать роль ключевых токенов и уменьшать роль менее важных токенов. Преимущество трансформеров состоит в том, что они не имеют рекуррентных ячеек (units) и, следовательно, требуют меньше времени на обучение, чем рекуррентные нейронные архитектуры, такие как, например, длинная краткосрочная память (Long Short-Term Memory, LSTM).

Модель GIT была предложена в [2]. GIT – трансформер-декодер, работающий как с токенами изображения CLIP (Contrastive Language-Image Pretraining), так и с текстовыми токенами. Модель обучается с использованием «принуждения учителя» на множестве пар (изображение, текст). Цель модели – спрогнозировать следующий текстовый токен, имея токены изображения и предыдущие текстовые токены. Модель имеет полный доступ к токенам патча изображения (т. е. используется двунаправленная маска внимания), но имеет доступ только к предыдущим текстовым токенам (т.е. используется причинно-следственная маска внимания) при предсказании следующего текстового токена. Авторы модели GIT утверждают, что генеративные модели обеспечивают согласованную сетевую архитектуру между предварительным обучением и точной настройкой, в то время как существующие модели обычно имеют сложную архитектуру: они содержат различные структуры (одно-/многомодальный кодер/декодер) и зависят от внешних модулей, таких как детекторы объектов/теггеры и инструменты оптического распознавания символов (OCR, Optical Character Recognition). В модели GIT архитектура упрощена: они содержат один кодер изображений и один декодер текста в рамках одной задачи моделирования языка.

Модель GIT может быть использована для решения таких задач, как:

– визуальное вопросно-ответное (VQA, Visual Question Answering) моделирование для изображений и видео;

– создание аннотаций к изображениям и видео;

– классификация изображений (посредством согласования модели с изображением и формированием запроса на генерирование класса изображения в виде текста).

Модель BLIP была предложена в [3]. В модели BLIP используется новая структура VLP (Vision-Language Pre-training), которая гибко переносится как на визуально-языковое понимание, так и на задачи генерации. Модель BLIP использует аннотатор для создания синтетических подписей к веб-изображениям и фильтр – для удаления зашумленных пар «изображение-текст». Аннотатор и фильтр инициализируются на основе одной и той же предварительно обученной модели и настраиваются индивидуально на наборе данных, аннотированных человеком. Авторы модели BLIP утверждают, что большинство существующих предварительно обученных

моделей хорошо работают только на задачах, основанных на понимании или задачах, основанных на генерации. При этом повышение производительности в этих моделях во многом достигается посредством увеличения набора данных за счет зашумленных пар «изображение-текст», собранных из Интернета, что является неоптимальным источником контроля.

Модель BLIP может быть использована для решения таких задач, как:

– визуальное вопросно-ответное (VQA, Visual Question Answering) моделирование для изображений;
– создание аннотаций к изображениям;
– сопоставление изображения и текста.

Программные реализации моделей GIT [5] и BLIP [6] доступны на сайте предлагаемой на сайте моделей глубокого обучения HuggingFace. Они способны успешно решать задачу аннотирования изображений. Однако выбор будет сделан в пользу модели BLIP, поскольку при разработке ПС не планируется работать с видеоданными, и, следовательно, имеет смысл отказаться от модели GIT, предлагающей дополнительный функционал.

На рисунке 2 представлена структура фреймворка BLIP [3].

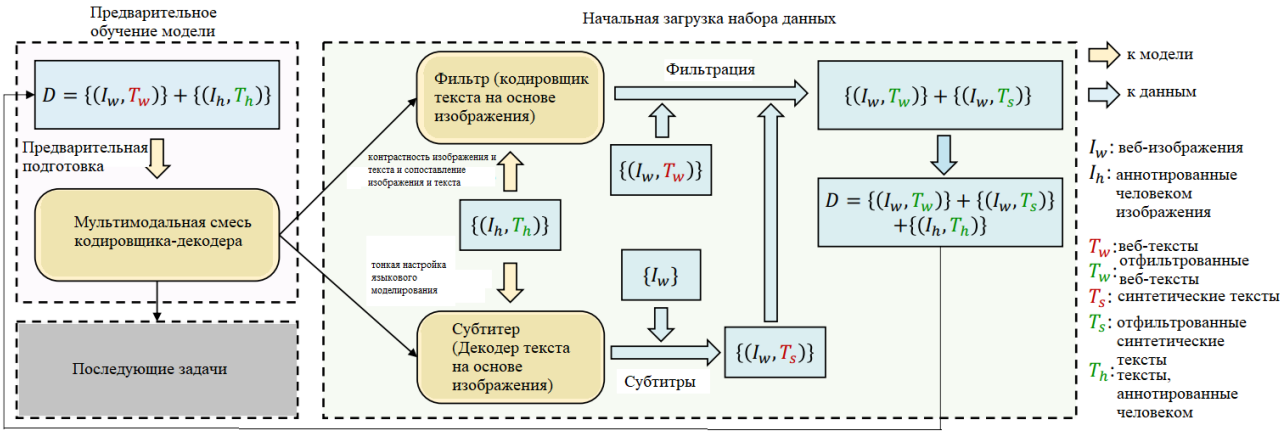


Рисунок 2 – Структура фреймворка BLIP [3]

При разработке ПС модель BLIP планируется дообучить на наборе данных, содержащих пары «изображение – текст», описывающие КПК.

1.2. Выбор модели глубокого обучения для перевода текстововых описаний

При разработке ПС целесообразно использовать уже предобученную модель перевода текстов. В связи с этим было принято решение об использовании модели T5 (Text-To-Text Transfer Transformer) [7], программные реализации которой доступны на сайте моделей глубокого обучения Hugging Face. В частности, модель T5 [8] представляет собой трансформер, работающий в многозадачном режиме перевода на требуемый язык. Она настроена на машинный перевод по парам между такими языками, как русский, английский и китайский. Модель обеспечивает приемлемое качество перевода при высокой скорости его выполнения. Модель T5 [8] будет использоваться в представленном виде (без каких-либо изменений с целью дообучения).

2. Разработка программных средств

При разработке ПС была определена их архитектура, разработан набор данных, описывающий целевую предметную область, выполнено дообучение модели BLIP, разработана база данных, содержащая дополнительную информацию о КПК, разработаны программные компоненты ПС и выполнена их интеграция.

2.1. Разработка архитектуры программных средств

При разработке ПС было принято решение об использовании клиент-серверной архитектуры. Клиентская часть ПС расположена на гарнитуре Microsoft Hololens 2 и отвечает за графический интерфейс, формирование снимков КПК, а также – отправку запросов на сервер.

Серверная часть ПС содержит модуль аннотирования изображений, модуль перевода текстовых описаний, а также – модуль базы данных с информацией о названиях КПК, их текстовых описаниях и верифицирующих изображениях.

Модули аннотирования и перевода текстов основаны на применении моделей нейронных сетей глубокого обучения – BLIP и T5 соответственно, представляющих собой модели-трансформеры.

На рисунке 3 представлена архитектура разработанных ПС.

Клиентская часть ПС написана на основе среды разработки Unity с использованием MRTK на языке C#. Среда разработки Unity предоставляет удобную среду для разработки приложений, а MRTK содержит инструменты, необходимые для разработки приложений со смешанной реальностью.

С помощью MRTK был разработан графический интерфейс пользователя с использованием технологий смешанной реальности и голограмм. Клиентская часть ПС позволяет пользователю удобно отправлять запросы к серверно части ПС с использованием графического интерфейса пользователя.

Серверная часть ПС написана на языке Python с использованием фреймворков Flask, Flask-RESTful, Flask-SQLAlchemy и библиотеки transformers. Фреймворки Flask и Flask-RESTful позволяют легко создать сервер с API. Фреймворк Flask-SQLAlchemy позволяет легко взаимодействовать с базой данных. Библиотека transformers необходима для работы с моделями глубокого обучения с архитектурой «трансформер».

Серверная часть ПС имеет несколько конечных точек (эндпоинтов), которые способны принимать запросы от клиентской части ПС и обрабатывать эти запросы, производя распознавание и аннотирование изображений КПК, поиск информации о КПК в базе данных и перевод текстовых описаний КПК.

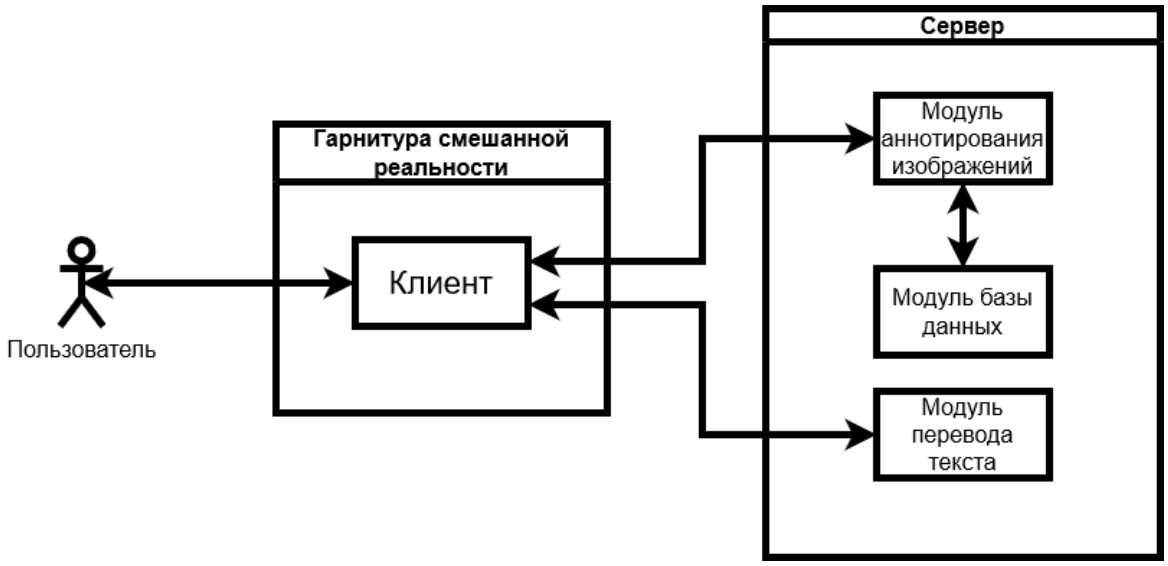


Рисунок 3 – Архитектура программных средств

Разработанные клиентская и серверная части ПС были интегрированы с помощью API (Application Programming Interface), который позволяет клиентской части ПС отправлять запросы к серверной части и получать необходимую информацию.

2.2. Разработка набора данных для дообучения модели BLIP

Для адекватного функционирования ПС в контексте решения задачи распознавания и интерпретации изображений КПК было принято решение о необходимости дообучения модели глубокого обучения BLIP, применяемой для аннотирования изображений.

Для дообучения модели BLIP в контексте задачи распознавания и интерпретации изображений КПК был создан набор данных в формате JSON (JavaScript Object Notation) [9], который является стандартным текстовым форматом для хранения и передачи структурированных данных. Этот формат основан на синтаксисе объекта в JavaScript, однако не привязан к нему. Работу с этим форматом поддерживают многие современные языки программирования, в том числе и Python.

В соответствии с руководством [10] был разработан набор данных, включающий в себя два столбца с названиями file_name и text. В столбце file_name, отвечающем за изображения КПК, указаны названия файлов в формате JPEG с изображениями ППК. В столбце text, отвечающем за названия КПК, указаны официально установленные названия КПК. При этом одной и той же КПК сопоставлено несколько её изображений – от 2 до 6 (в разных ракурсах, при разном освещении и т.д.). Например, для комплектующей ASUS PRIME A520M-E Motherboard в наборе данных предусмотрены изображения, представленные файлами 001.jpeg, 002.jpeg, 003.jpeg, 004.jpeg и 005.jpeg. Разработанный набор данных содержит 125 строк, описывающих 27 КПК.

На рисунке 4 представлен пример, демонстрирующий организацию структуры набора данных.

```
{"file_name": "001.jpg", "text": "ASUS PRIME A520M-E Motherboard"}
{"file_name": "002.jpg", "text": "ASUS PRIME A520M-E Motherboard"}
{"file_name": "003.jpg", "text": "ASUS PRIME A520M-E Motherboard"}
{"file_name": "004.jpg", "text": "ASUS PRIME A520M-E Motherboard"}
{"file_name": "005.jpg", "text": "ASUS PRIME A520M-E Motherboard"}

{"file_name": "006.jpg", "text": "ASUS TUF GAMING B550M-PLUS WI-FI II Motherboard"}
{"file_name": "007.jpg", "text": "ASUS TUF GAMING B550M-PLUS WI-FI II Motherboard"}
{"file_name": "008.jpg", "text": "ASUS TUF GAMING B550M-PLUS WI-FI II Motherboard"}
{"file_name": "009.jpg", "text": "ASUS TUF GAMING B550M-PLUS WI-FI II Motherboard"}
{"file_name": "010.jpg", "text": "ASUS TUF GAMING B550M-PLUS WI-FI II Motherboard"}
{"file_name": "011.jpg", "text": "ASUS TUF GAMING B550M-PLUS WI-FI II Motherboard"}
```


Рисунок 4 – Организация структуры набора данных в соответствии с руководством [10]

2.3. Дообучение модели BLIP

Дообучение модели BLIP проводилось в среде Google Collab с использованием GPU, который позволяет ускорить выполнение тензорных вычислений, реализуемых в нейросетевых моделях.

Дообучение модели BLIP на примерах из целевой предметной области необходимо в связи с тем, что, хотя модель и обучена на очень большом числе примеров, а именно – на 129 млн. изображений, собранных из наборов COCO [11], VG [12], CC [13], SBU [14] и LAION [15] (с низким разрешением 384×384, что считается достоинством модели, поскольку обеспечивает более быстрый вывод результатов, чем у других моделей, обученных на изображениях с более высоким разрешением), она «не знает» («не выучила») специфику целевой предметной области. Следовательно, знания модели BLIP, не дообученной примерах из целевой предметной области, скорее всего, будут поверхностными, хотя она и сможет генерировать аннотации, близкие по сути, но не вполне точные.

Особое внимание при дообучении модели BLIP было уделено настройке значений гиперпараметров модели BLIP, информация о которых приведена в таблице 1. Настройка значений гиперпараметров позволила дообучить модель BLIP так, чтобы она с высокой точностью распознавала новые объекты, но при этом не была переобучена [при выборе значений этих гиперпараметров выполнялся анализ значений LM-функции потерь языкового моделирования (Language Modeling, LM) на обучающей и валидационной выборках]. Выбор значений гиперпараметров был выполнен с использованием Optuna [16] – фреймворка, позволяющего осуществить поиск оптимальных значений гиперпараметров моделей машинного и глубокого обучения.

Таблица 1 – Гиперпараметры модели BLIP

Название гиперпараметра	**Описание гиперпараметра**	**Установленное значение**
epochs	число эпох	100
learning_rate	начальная скорость обучения	10^{-4}
train_batch_size	размер батча в обучающей выборке	4
valid_batch_size	размер батча в валидационной выборке	6
scheduler	закон изменения скорости обучения	CosineAnneaLinqLR (косинусный отжиг)
min_lr	нижняя граница скорости обучения	10^{-6}
T_max	максимальное число итераций	500
weight_decay	коэффициент L2 регуляризации	10^{-6}
n_accumulate	коэффициент аккумуляции	1

Результаты дообучения модели BLIP (в том числе – сама дообученная модель BLIP и различные графические зависимости, характеризующие процесс дообучения, были сохранены в каталоге на платформе Weights&Biases [17], предоставляющей разработчикам инструментов искусственного интеллекта возможности для хранения полученных результатов и оптимизации рабочего процесса.

Изображения в наборе данных были предварительно подготовлены с помощью препроцессора BLIP с целью правильного дообучения модели BLIP. Предварительная обработка (препроцессинг) включает в себя настройку яркости, контрастности и ориентации входного изображения с одновременным удалением шума.

При расчете LM-функции потерь языкового моделирования активируется декодер текста на основе изображения, целью которого является создание текстовых описаний для изображения. Декодер текста оптимизирует функцию потерь на основе перекрестной энтропии CrossEntropyLoss (из библиотеки torch, предоставляющей инструменты для работы с тензорами и динамическии нейронными сетями на Python с сильным ускорением графического процессора) [18]. При этом модель BLIP обучается максимизировать

вероятность текста авторегрессионным способом. При расчете LM-функции потерь выполняется сглаживание меток с коэффициентом 0,1. LM-функция потерь позволяет модели BLIP преобразовывать (с возможностью обобщения) визуальную информацию в связные аннотации (подписи).

Функция потерь на основе перекрестной энтропии CrossEntropyLoss – метрика, позволяющая оценить, насколько хорошо функционирует модель глубокого обучения. Чем ближе значение функции потерь к нулю, тем лучше: значение функции потерь должно быть минимизировано. Функция потерь позволяет модели определить, насколько она «неправильная» и на основании этой «неправильности» – улучшить себя.

Пусть $y_1, y_2, \dots, y_n$ – токены в предложении. Тогда в случае языкового моделирования вероятность $p(y_1, y_2, \dots, y_n)$ появления последовательности токенов в указанном порядке вычисляется как:

$$p(y_1, y_2, \dots, y_n) = p(y_1) \cdot p(y_2|y_1) \cdot p(y_3|y_1, y_2) \cdot \dots \cdot p(y_n|y_1, \dots, y_{n-1}) = \prod_{t=1}^{n} p(y_t|y_{<t}) . \quad (1)$$

В формуле (1) вероятность $p(y_1, y_2, \dots, y_n)$ появления последовательности токенов в указанном порядке представлена как декомпозиция на условные вероятности каждого токена с учётом предыдущего контекста.

Нейронные языковые модели можно рассматривать как классификаторы, которые классифицируют префикс текста на $|V|$ классов, где классы являются словарными токенами.

Для нейронных языковых моделей различных архитектур процесс вывода выглядит следующим образом:

– встроить слова для предыдущих (контекстных) слов в нейронную сеть;

– получить векторное представление контекста из нейронной сети;

– спрогнозировать распределение вероятностей для следующего токена на основе векторного представления.

Пусть h_t – векторное представление контекста $y_1, y_2, \dots, y_{t-1}$, а e_{y_t} и e_ω выходные векторные вложения.

Тогда

$$p(y_t|y_{<t}) = \frac{exp(h_t^T \cdot e_{y_t})}{\sum_{\omega \in V} exp(h_t^T \cdot e_\omega)}. \quad (2)$$

Для всей последовательности токенов LM-функция потерь может быть вычислена как

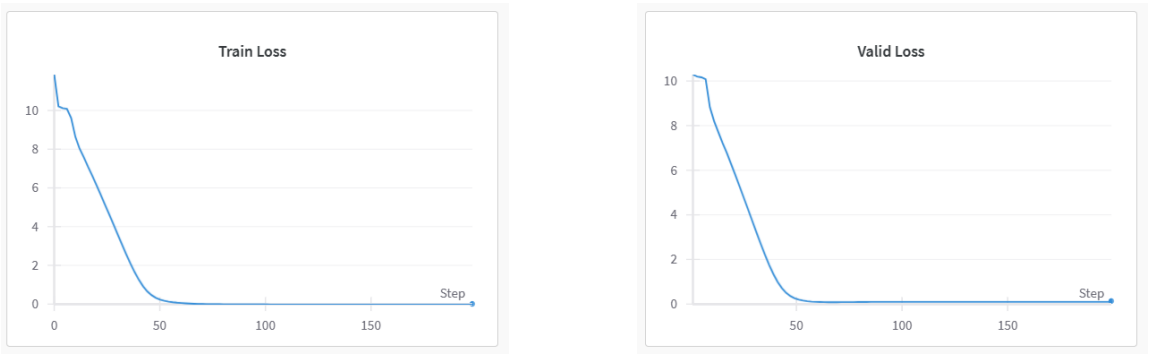
$$Loss = -\sum_{t=1}^{n} p(y_t|y_{<t}). \quad (3)$$

В процессе дообучения модели BLIP на каждой эпохе выполнялось оценивание значений текущей скорости обучения LR и функции потерь $Loss$ (3) на обучающей и валидационной выборках.

На 100-й эпохе дообучения оказалось, что LR = 6.58e–5, Train_Loss = 0.00026, Valid_Loss = 0.00380. Лучшее значение функции потерь $Loss$ (3) на валидационной выборке составило 0.00358.

При этом время дообучения модели BLIP оказалось равно 1 часов 11 минут 05 секунд.

На рисунках 5,а и 5,б приведены графики для функции потерь $Loss$ (3) на обучающей и валидационной выборках, построенные средствами платформе Weights&Biases по результатам дообучения.



а б

Рисунок 5 – Графики для функции потерь:
а – на обучающей выборке; б – на валидационной выборке

На рисунке 6,а приведено изображение КПК – материнской платы – с официально установленным названием ASUS TUF GAMING B550M-PLUS WI-FI II Motherboard, на рисунке 6,б приведена аннотация, сопоставленная этому изображению недообученной моделью BLIP, а на рисунке 6,в приведена аннотация, сопоставленная этому изображению дообученной моделью BLIP.

а

msx b450a - gaming motherboard

б

ASUS TUF GAMING B550M-PLUS WI-FI II Motherboard

в

Рисунок 6 – Аннотирование изображения моделью BLIP:
а – изображение; б – аннотация на основе недообученной модели; в – аннотация на основе дообученной модели

Из представленного примера видно, что недообученная модель смогла угадать «суть» изображения, но оказалась не настолько точна, как дообученная модель, которая сгенерировала абсолютно правильную аннотацию.

Таким образом, можно сделать вывод о целесообразности выполненного дообучения модели BLIP.

2.4. Разработка базы данных комплектующих персонального компьютера

Разработка базы данных, содержащая информацию о КПК, их подробном описании и визуальном представлении, была выполнена с целью предоставления более подробной информации о КПК и визуальной верификации КПК на основе их аннотаций.

База данных состоит из одной таблицы. Таблица содержит три столбца, определяющих название КПК, описание КПК и изображение КПК. В эту таблицу вводится информация о названии и подробном описании каждой КПК. Эта таблица позволяет пользователям быстро находить нужную информацию о каждой КПК и узнавать её основные характеристики и особенности, а также получать визуальное представление о каждой КПК. Изображения КПК, извлеченные из базы данных, отображаются в клиентской части ПС – в голографическом окне гарнитуры Hololens 2 – для удобства пользователей, в том числе – для верификации результатов распознавания КПК.

База данных размещена на серверной части ПС, который обеспечивает постоянный доступ к данным и позволяет клиентской части ПС отправлять запросы к базе данных.

После распознавания изображения клиентская часть ПС отправляет запрос к базе данных, находящейся на сервере, и получает дополнительную информацию о распознанном объекте, если КПК, соответствующая объекту, будет обнаружена в базе данных. Эта дополнительная информация будет отображена в голографическом окне гарнитуры Hololens 2. Если КПК, соответствующая объекту, не будет обнаружена в базе данных, в голографическом окне гарнитуры Hololens 2 будет выведено соответствующее сообщение.

В таблице 2 приведен пример дополнительной информации, которую можно извлечь из базы данных для КПК Intel PWLA8391GT PRO1000 GT PCI Network Adapter.

Таблица 2 – Пример дополнительной информации, извлеченной из базы данных, для КПК Intel PWLA8391GT PRO1000 GT PCI Network Adapter

Название комплектующей	**Описание комплектующей**	**Верифицирующее изображение**
Intel PWLA8391GT PRO1000 GT PCI Network Adapter	ASUS A9200SE/T/P/128M/A 128MB AGP PC GPU – это видеокарта производства компании ASUS. Она имеет 128 МБ видеопамяти и работает через интерфейс AGP. Эта видеокарта была выпущена для установки в старые компьютеры с разъемом AGP, и является обычно используемой в компьютерах конца 1990-х – начала 2000-х годов.	

2.5. Реализация перевода текстовых описаний

Для реализации перевода текстового описания КПК в ПС используется модель глубокого обучения [8], представляющая собой трансформер T5.

После того, как выполнено распознавание объекта, получена его аннотация и из базы данных извлечено описание соответствующей КПК, у пользователя есть возможность осуществления перевода полученного описания на другой язык, удобный для восприятия информации [19].

Клиентская часть ПС может отправить на сервер запрос, содержащий текстовое описание КПК на исходном языке и желаемый язык перевода. Этот запрос обрабатывается сервере. В процессе обработки текстового описания КПК с использованием модели глубокого обучения осуществляется его перевод, отличающийся высокой точностью. Используемая модель глубокого обучения позволяет выполнить перевод описания КПК на английский и китайский языки. После обработки на сервере текст перевода описания КПК возвращается клиентской части ПС: результаты перевода отображаются в голографическом окне гарнитуры Microsoft Hololens 2.

3. Апробация программных средств

Для работы с ПС пользователю требуется подключиться к Wi-Fi или WLAN, после чего запустить клиентскую часть ПС, расположенную на гарнитуре Microsoft Hololens 2. В результате пользователь увидит перед собой главное голографическое окно ПС (рисунок 7). Чтобы выполнить распознавание и аннотирование изображений, пользователь должен посмотреть на объект, который он хочет распознать и проаннотировать, и нажать кнопку «Сделать снимок» в главном голографическом окне ПС (рисунок 7), получив снимок КПК (рисунок 8).

ПС произведут распознавание объекта (рисунок 9), затратив некоторое время на обращение к серверной части ПС, и перед пользователем появится голографическое окно с результатом распознавания изображения КПК, если оно прошло успешно, в виде аннотации, а также – с извлеченными из базы данных текстовым описанием КПК и верифицирующим изображением КПК (рисунок 10). Вывод изображения КПК в голографическом окне полезен с точки зрения верификации результатов распознавания. Окно с результатом распознавания изображения будет содержать кнопки «Английский» и «Китайский», предназначенные для реализации опций перевода на соответствующие языки. В любой момент времени пользователь может осуществить перевод текстового описания на выбранный им язык или же перейти к распознаванию нового изображения КПК. Такие действия возможны благодаря наличию в поле зрения пользователя нескольких голографических окон одновременно. При выборе пользователем опции перевода на тот или иной язык ПС, затратив некоторое время на обращение к серверной части ПС (рисунок 11), выполнят обновление информации в голографическом окне (рисунок 12). Если КПК не распознана, в голографическом окне будет выведено сообщение «Объект не распознан» (рисунок 13). В любом случае – успешно распознан объект или нет – пользователь имеет возможность выполнить распознавание еще одной КПК. Для завершения работы с ПС пользователя следует нажать кнопку «Закрыть» (рисунки 10 и 13).

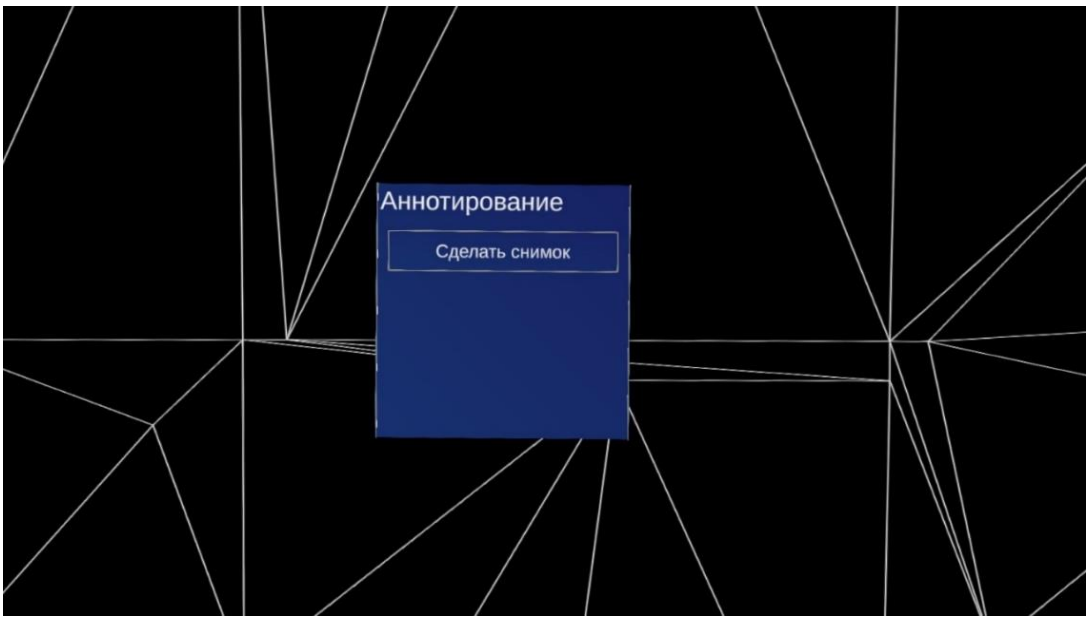


Рисунок 7 – Главное голографическое окно программных средств с приглашением сделать снимок КПК

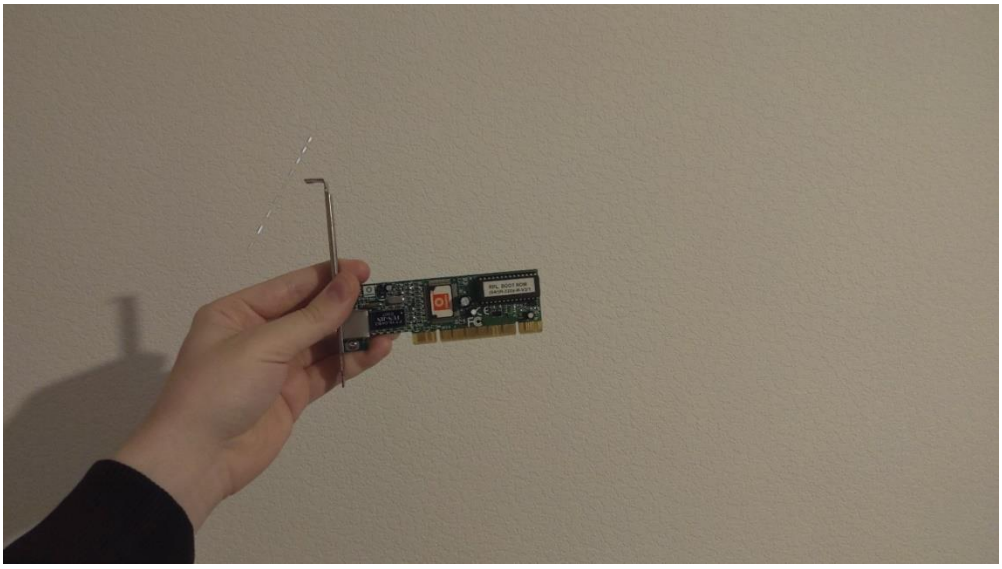


Рисунок 8 – Снимок КПК
Intel PWLA8391GT PRO1000 GT PCI Network Adapter

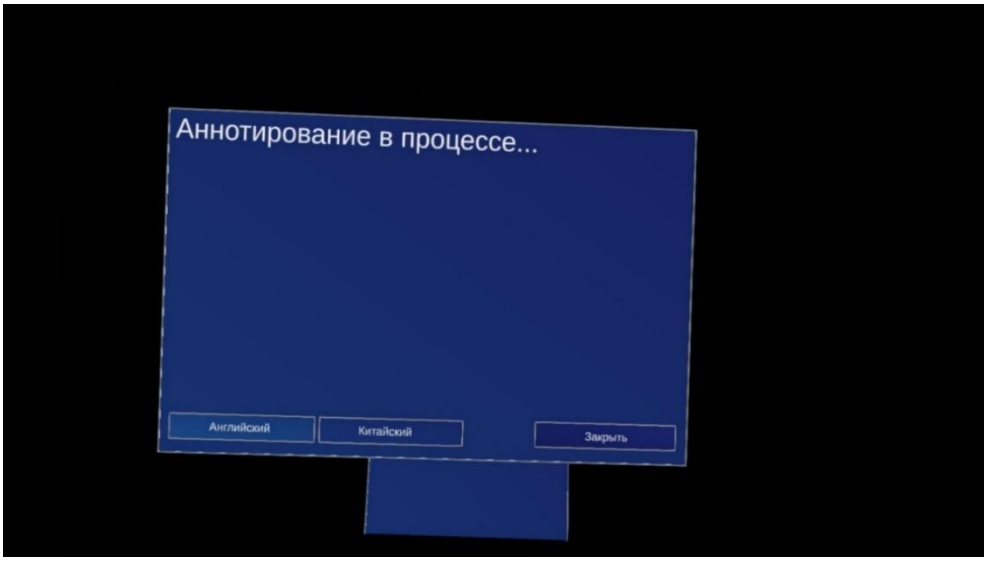


Рисунок 9 – Голографическое в состоянии ожидания ответа от сервера

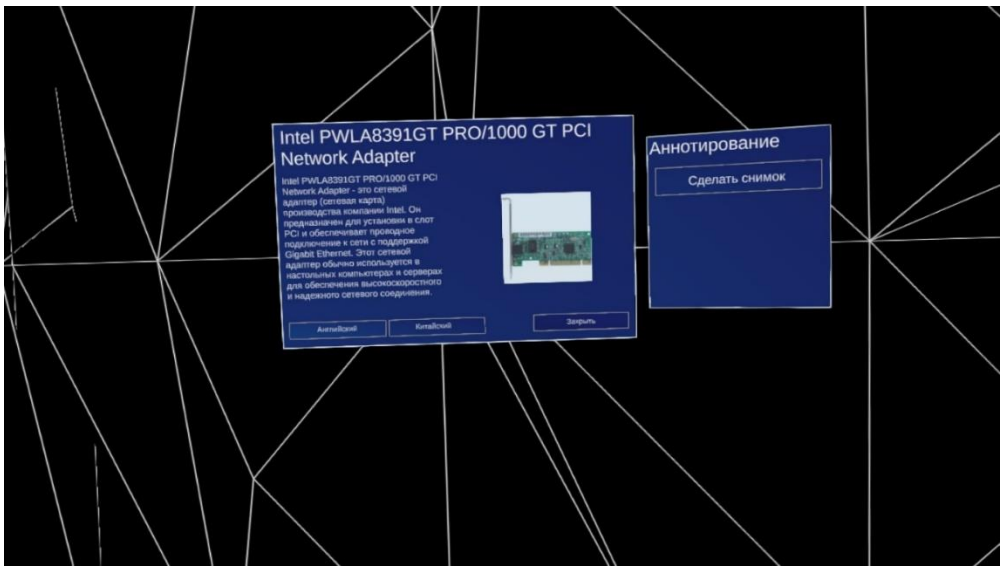


Рисунок 10 – Голографическое окно с результатами успешного распознавания КПК Intel PWLA8391GT PRO1000 GT PCI Network Adapter

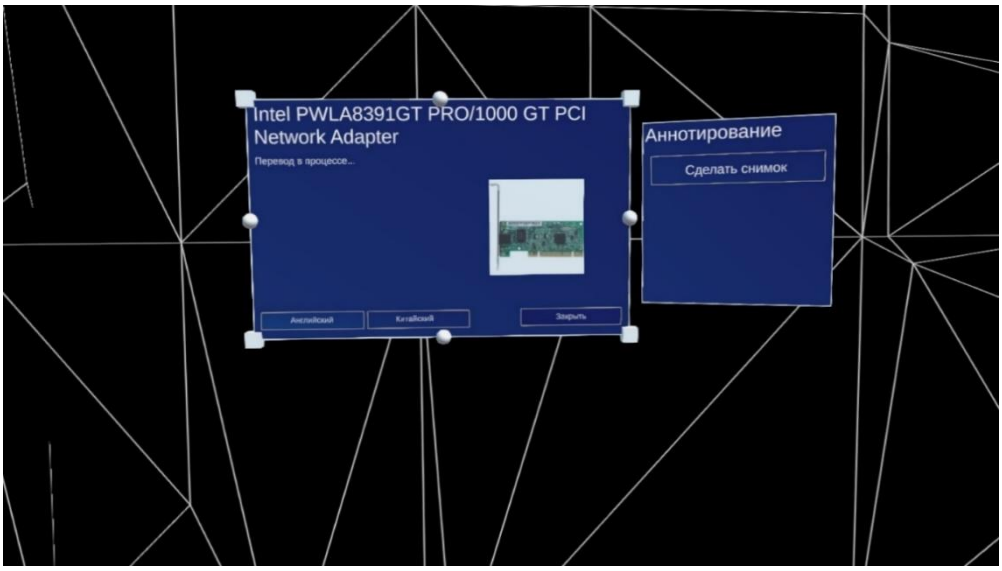


Рисунок 11 – Голографическое окно в состоянии ожидания результата перевода на английский язык описания КПК Intel PWLA8391GT PRO1000 GT PCI Network Adapter

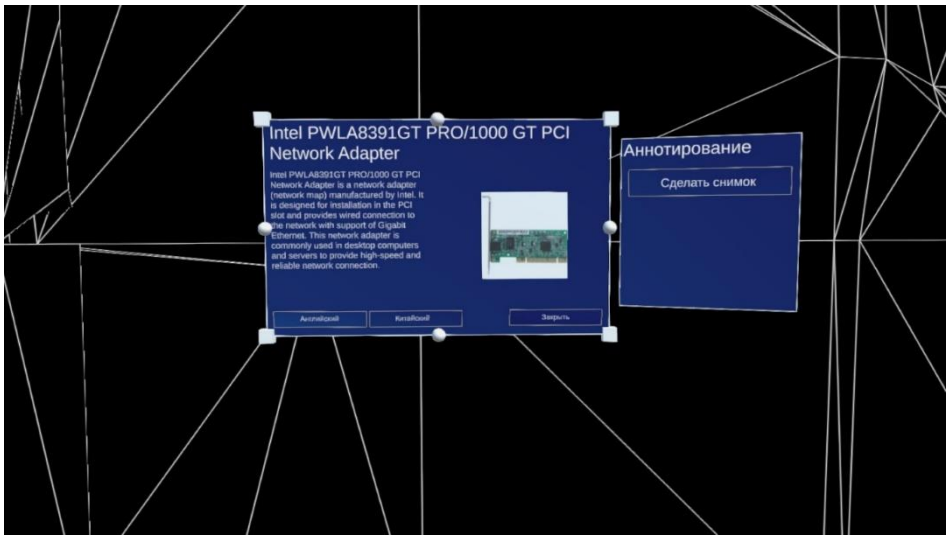


Рис. 12. Голографическое окно с результатами перевода на английский язык описания КПК Intel PWLA8391GT PRO1000 GT PCI Network Adapter

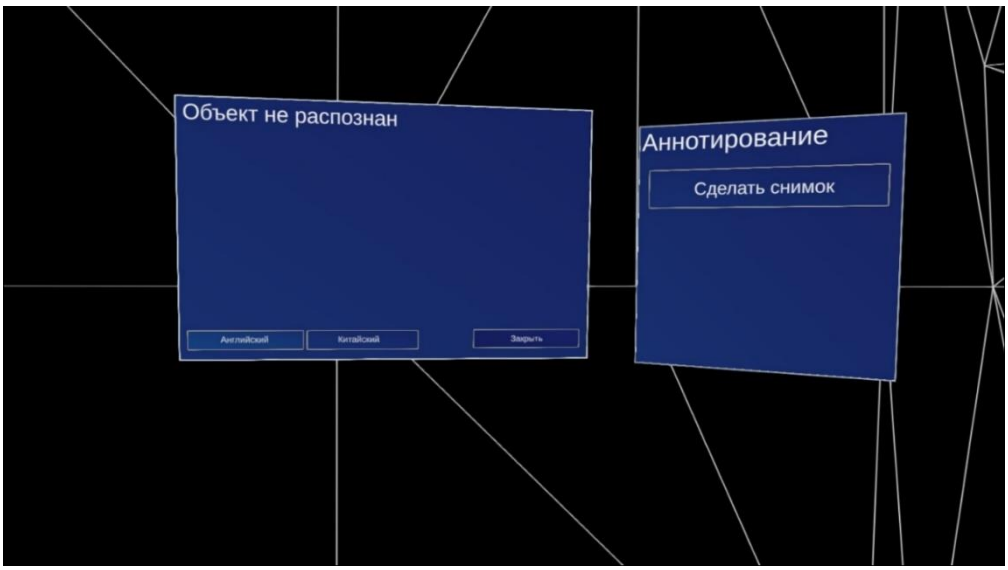


Рис. 13. Голографическое окно с результатами неуспешного распознавания КПК

Заключение

При работе с ПС пользователь, взаимодействующий с гарнитурой Microsoft Hololens 2, может сделать снимок КПК и отправить запрос на сервер для аннотирования изображения и перевода текстового описания комплектующей ПК на удобный для восприятия язык. Результаты работы ПС отображаются в соответствующих голографических окнах гарнитуры Microsoft Hololens 2.

В результате работы были созданы ПС реализации облачных вычислений, обеспечивающие решение задачи распознавания и интерпретации изображений с использованием гарнитуры смешанной реальности Microsoft Hololens 2 в контексте работы по оптимизации процесса идентификации комплектующих ПК.

Список литературы

1. Microsoft HoloLens 2. For precise, efficient hands-free work [Электронный ресурс]. – Режим доступа: https://www.microsoft.com/en-us/hololens, свободный (дата обращения: 01.03.2024).
2. Wang J., Yang Z., Hu X., Li L., Lin K., Gan Z., Liu Z., Liu C., Wang L. GIT: A Generative Image-to-text Transformer for Vision and Language // arXiv:2205.14100v5. – 2022. https://doi.org/10.48550/arXiv.2205.14100.
3. Li J., Li D., Xiong C., Hoi S. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation // arXiv:2201.12086v2. – 2022. https://doi.org/10.48550/arXiv.2201.12086.
4. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. Attention Is All You Need // arXiv:1706.03762v7. – 2023. https://doi.org/10.48550/arXiv.1706.03762.
5. Hugging Face. microsoft/git-large [Электронный ресурс]. – Режим доступа: https://huggingface.co/microsoft/git-large, свободный (дата обращения: 01.03.2024).
6. Hugging Face. Salesforce/blip-image-captioning-large [Электронный ресурс]. – Режим доступа: https://huggingface.co/Salesforce/blip-image-captioning-large, свободный (дата обращения: 01.03.2024).
7. Raffel C., Shazeer N., Roberts A, Lee K., Narang Sh., Matena M., Zhou Y., Li W., Liu P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer // Journal of Machine Learning Research. – 2020. – Vol. 21. – pp. 1–67.
8. Hugging Face. utrobinmv/t5_translate_en_ru_zh_large_1024 [Электронный ресурс]. – Режим доступа: https://huggingface.co/utrobinmv/t5_translate_en_ru_zh_large_1024, свободный (дата обращения: 01.03.2024).
9. Введение в JSON [Электронный ресурс]. – Режим доступа: https://www.json.org/json-ru.html, свободный (дата обращения: 01.03.2024).
10. Hugging Face. Create an image dataset [Электронный ресурс]. – Режим доступа: https://huggingface.co/docs/datasets/image_dataset, свободный (дата обращения: 01.03.2024).
11. Lin T., Maire M., Belongie S.J., Hays J., Perona P., Ramanan D., Doll´ar P., Zitnick C.L. Microsoft COCO: common objects in context. In Fleet D.J., Pajdla T., Schiele B., Tuytelaars T. (eds.). The European Conference on Computer Vision (ECCV). – 2014. – Vol. 8693. – pp. 740–755.

12. Krishna R., Zhu Y., Groth O., Johnson J., Hata K., Kravitz J., Chen S., Kalantidis Y., Li L., Shamma D.A., Bernstein M.S., Fei-Fei L. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision (IJCV). – 2017. – Vol. 123(1). – pp. 32–73.
13. Changpinyo S., Sharma P., Ding N., Soricut R. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts // arXiv:2102.08981v2. – 2021. https://doi.org/10.48550/arXiv.2102.08981.
14. Ordonez V., Kulkarni G., Berg T.L. Im2text: Describing images using 1 million captioned photographs. In Shawe-Taylor J., Zemel R.S., Bartlett P.L., Pereira F.C.N., Weinberger K.Q. (eds.). The Neural Information Processing Systems. – 2011. – pp. – 1143–1151.
15. Schuhmann C., Vencu R., Beaumont R., Kaczmarczyk R., Mullis C., Katta A., Coombes T., Jitsev J., Komatsuzaki A. Laion-400m: Open dataset of clipfiltered 400 million image-text pairs. arXiv preprint arXiv:2111.02114. 2021.
16. Optuna – a hyperparameter optimization framework [Электронный ресурс]. – Режим доступа: https://optuna.org/, свободный (дата обращения: 01.03.2024).
17. Weights & Biases: The AI Developer Platform [Электронный ресурс]. – Режим доступа: https://wandb.ai/site, свободный (дата обращения: 01.03.2024).
18. CrossEntropyLoss [Электронный ресурс]. – Режим доступа: https://pytorch.org/docs/stable /generated/torch.nn.CrossEntropyLoss.html, свободный (дата обращения: 01.03.2024).
19. Демидов Н.А. Аспекты решения задачи распознавания и перевода текстов с использованием гарнитуры Hololens 2 // Решение: материалы XII Всероссийской научно-практической конференции (г. Березники, 21 октября 2023 г.). – Пермь: Изд-во Перм. нац. исслед. политехн. ун-та, 2023. – С. 108–110.

References

1. Microsoft HoloLens 2. For precise, efficient hands-free work [Electronic resource]. – Access mode: https://www.microsoft.com/en-us/hololens, free (access date: 01.03.2024).
2. Wang J., Yang Z., Hu X., Li L., Lin K., Gan Z., Liu Z., Liu C., Wang L. GIT: A Generative Image-to-text Transformer for Vision and Language // arXiv:2205.14100v5. – 2022. https://doi.org/10.48550/arXiv.2205.14100.
3. Li J., Li D., Xiong C., Hoi S. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation // arXiv:2201.12086v2. – 2022. https://doi.org/10.48550/arXiv.2201.12086.
4. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. Attention Is All You Need // arXiv:1706.03762v7. – 2023. https://doi.org/10.48550/arXiv.1706.03762.
5. Hugging Face. microsoft/git-large [Electronic resource]. – Access mode: https://huggingface.co/microsoft/git-large, free (access date: 01.03.2024).
6. Hugging Face. Salesforce/blip-image-captioning-large [Electronic resource]. – Access mode: https://huggingface.co/Salesforce/blip-image-captioning-large, free (access date: 01.03.2024).
7. Raffel C., Shazeer N., Roberts A, Lee K., Narang Sh., Matena M., Zhou Y., Li W., Liu P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer // Journal of Machine Learning Research. – 2020. – Vol. 21. – pp. 1–67.
8. Hugging Face. utrobinmv/t5_translate_en_ru_zh_large_1024 [Electronic resource]. – Access mode: https://huggingface.co/utrobinmv/t5_translate_en_ru_zh_large_1024, free (access date: 01.03.2024).
9. Introduction to JSON [Electronic resource]. – Access mode: https://www.json.org/json-ru.html, free (access date: 01.03.2024).
10. Hugging Face. Create an image dataset [Electronic resource]. – Access mode: https://huggingface.co/docs/datasets/image_dataset, free (access date: 01.03.2024).
11. Lin T., Maire M., Belongie S.J., Hays J., Perona P., Ramanan D., Doll'ar P., Zitnick C.L. Microsoft COCO: common objects in context. In Fleet D. J., Pajdla T., Schiele B., Tuytelaars T. (eds.). The European Conference on Computer Vision (ECCV). – 2014. – Vol. 8693. – pp. 740–755.
12. Krishna R., Zhu Y., Groth O., Johnson J., Hata K., Kravitz J., Chen S., Kalantidis Y., Li L., Shamma D.A., Bernstein M.S., Fei-Fei L. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision (IJCV). – 2017. – Vol. 123(1). – pp. 32–73.
13. Changpinyo S., Sharma P., Ding N., Soricut R. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts // arXiv:2102.08981v2. – 2021. https://doi.org/10.48550/arXiv.2102.08981.
14. Ordonez V., Kulkarni G., Berg T.L. Im2text: Describing images using 1 million captioned photographs. In Shawe-Taylor J., Zemel R.S., Bartlett P.L., Pereira F.C.N., Weinberger K.Q. (eds.). The Neural Information Processing Systems. – 2011. – pp. – 1143–1151.
15. Schuhmann C., Vencu R., Beaumont R., Kaczmarczyk R., Mullis C., Katta A., Coombes T., Jitsev J., Komatsuzaki A. Laion-400m: Open dataset of clipfiltered 400 million image-text pairs. arXiv preprint arXiv:2111.02114. 2021.

16. Optuna – a hyperparameter optimization framework [Electronic resource]. – Access mode: https://optuna.org/, free (access date: 01.03.2024).
17. Weights & Biases: The AI Developer Platform [Electronic resource]. – Access mode: https://wandb.ai/site, free (date of access: 01.03.2024).
18. CrossEntropyLoss [Electronic resource]. – Access mode: https://pytorch.org/docs/stable /generated/torch.nn.CrossEntropyLoss.html, free (access date: 01.03.2024).
19. Demidov N.A. Aspects of solving the problem of text recognition and translation using the headset Hololens 2 // Solution: materials of the XII All-Russian Scientific and Practical Conference (Berezniki, October 21, 2023). – Perm: Perm Publishing House. national research Polytechnic Univ., 2023. – pp. 108–110.