**Разработка системы доразметки данных в задачах классификации изображений**

**Состав команды:**

* Гудкова Д.А. (Б20-504)
* Каретникова О.А. (Б20-514)
* Мальгина Э.А. (Б20-514)
* Фигурин И.И. (Б20-514)
* Фонов Н.Д. (Б20-514) (тимлид)

**Техническое задание**

1. **Системные требования**

* Система, на которой будет исполняться приложение на основе Electron (в случае полного автономного десктопа - работа Python + TensorFlow)

1. **Функционал приложения**
   1. Указание настроек алгоритма доразметки
   2. Загрузка датасета с частичной разметкой или без нее
   3. Загрузка собственного алгоритма предобработки изображений
   4. Загрузка конфигурации модели и/или весов модели
   5. Выгрузка датасета с частичной разметкой
   6. Выгрузка весов модели и конфигурации
2. **Порядок (Flow) работы с приложением**
   1. Пользователь задает настройки алгоритма в приложении
   2. Пользователь загружает частично размеченный или неразмеченный датасет изображений
   3. Пользователь указывает собственный алгоритм предобработки изображений или выбирает из пресетов.
   4. Пользователь указывает модель из существующих пресетов или предоставляет собственную. К этой модели также есть возможность загрузить собственные веса.
   5. В случае, если нет размеченных данных в датасете, пользователю предлагается провести начальную разметку до работы с доразметкой.
   6. После указания всех пунктов a-e становится доступной функция доразметки. Это итеративный алгоритм, в частности, повторяются следующие пункты:
      1. Происходит обучение модели классификатора на существующих данных.
      2. Происходит разбиение неразмеченных данных на временные подклассы.
      3. Из одного из временных подклассов пользователю предлагается выбрать только те элементы, которые принадлежат одному из существующих (исходных) классов.
      4. Алгоритм на основе полученных вводных совершает доразметку и оценивает точность и полноту полученной разметки.
      5. Пользователю предоставляется возможность выполнить еще одну итерацию или остановить работу алгоритма с выгрузкой результатов.
   7. После окончания работы алгоритма пользователь имеет опции выгрузить следующие данные:

* Датасет с доразмеченными данными;
* Полученную конфигурацию модели;
* Конечные веса модели.

**Пользовательские требования**

1. Программа должна иметь функцию импорта частично размеченного или неразмеченного датасета в виде архива изображений формата .png;
2. Программа должна иметь функцию экспорта доразмеченного датасета в виде архива изображений формата .png;
3. Программа должна иметь функцию загрузки алгоритма предобработки данных в формате .py;
4. Программа должна иметь функцию загрузки ML-модели в формате .py;
5. Программа должна иметь функцию выгрузки обученной ML-модели в формате .py;
6. Программа должна предоставлять возможность просматривать файлы в датасете:
   1. группируя по классам,
   2. случайные изображения из всего набора;
7. Программа должна предоставлять пользователю возможность выбора алгоритма предобработки данных из ранее загруженных пользователем или из предоставляемых системой алгоритмов (Batch norm, Scalling, Gaussian Blur, VGG16);
8. Программа должна предоставлять пользователю возможность выбора оптимизируемого ML-алгоритма из ранее загруженных пользователем или из предоставляемых системой ML-алгоритмов (Байесовская классификация, Логистическая регрессия, KNN, Модели ансамблей);
9. Программа должна предоставлять пользователю выбирать метод кросс-валидации из предоставляемых системой методов (K-Fold, Repeated Random Sub-sampling, Leave-one-out, Holdout, Resubstitution);
10. Программа должна отображать пользователю метрики классификации из предоставляемых системой метрик (error, accuracy, recall, specificity, precision, fall-out, F-score, Cohen’s Kappa, Микро и Макро усреднение по метрикам) по выбору пользователя;
11. Программа должна отображать пользователю графики обучения из предоставляемых системой графиков (AUC ROC, PR ROC) алгоритма по выбору пользователя;
12. Программа должна предоставлять пользователю возможность доразметки изображений\*;
13. Программа должна иметь функцию автосохранения (каждые 10 минут работы системы) доразмеченных данных и весов модели для предотвращения потери данных пользователя.

\* – алгоритм выбора данных для оптимальной доразметки предоставляет пользователю набор изображений, в котором пользователю необходимо отметить изображения указанного алгоритмом класса.