Задача по анализу изображений на позицию «Аналитик данных»

Евгений Мунин

31.05.18 Тулуза

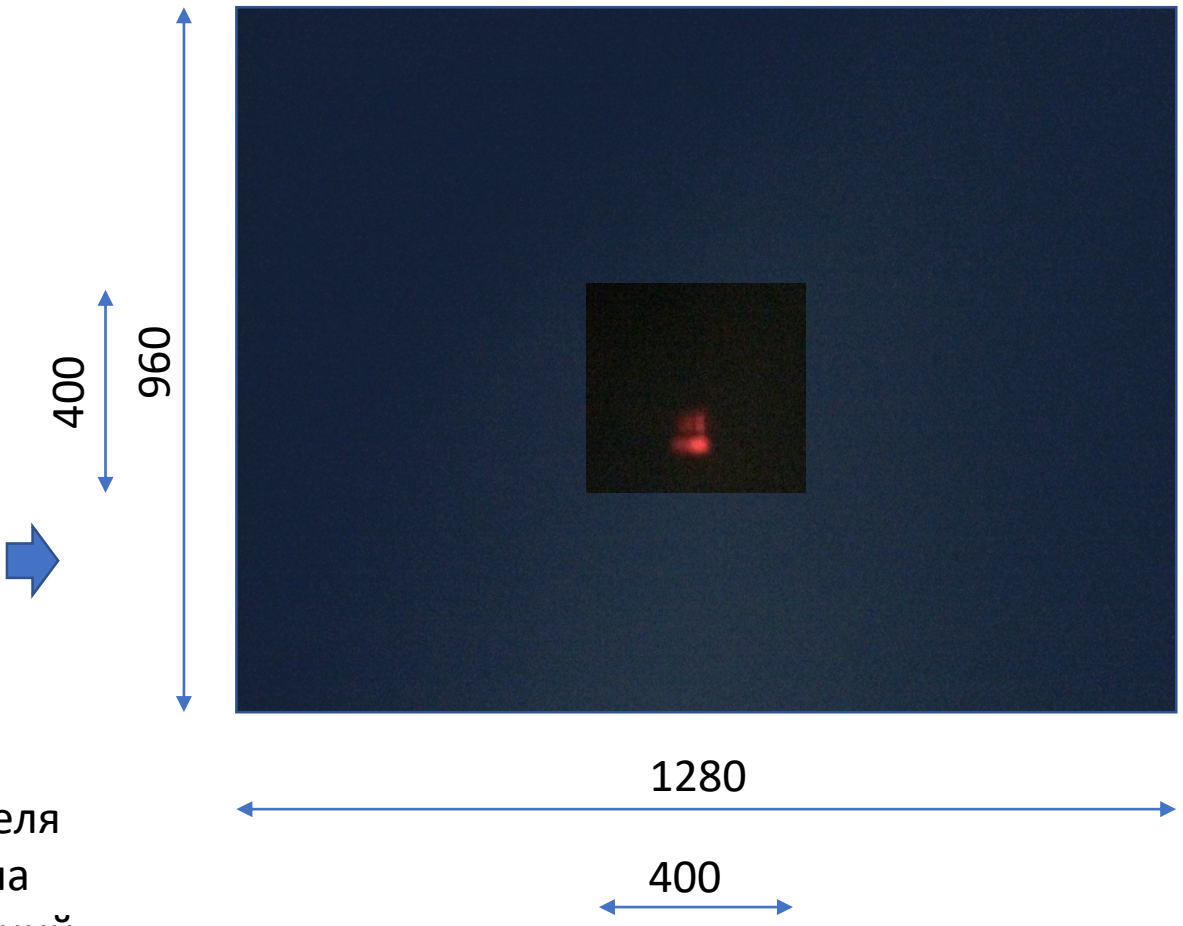
Предобработка исходных данных

Для повышения точности модели были применены следующие операции предобработки:

- Нормализация числовых признаков: приведение к значениям с мат. Ожиданием 0 и дисперсией 1.

- Для целевой переменной (категориальный признак) была применена операция **one-hot encoding**.

- Для сокращения времени на тренировку модели изображения были обрезаны, т.к. все изображения более или менее центрированы, и их полезная часть заключается в количестве и цвете центральных точек.

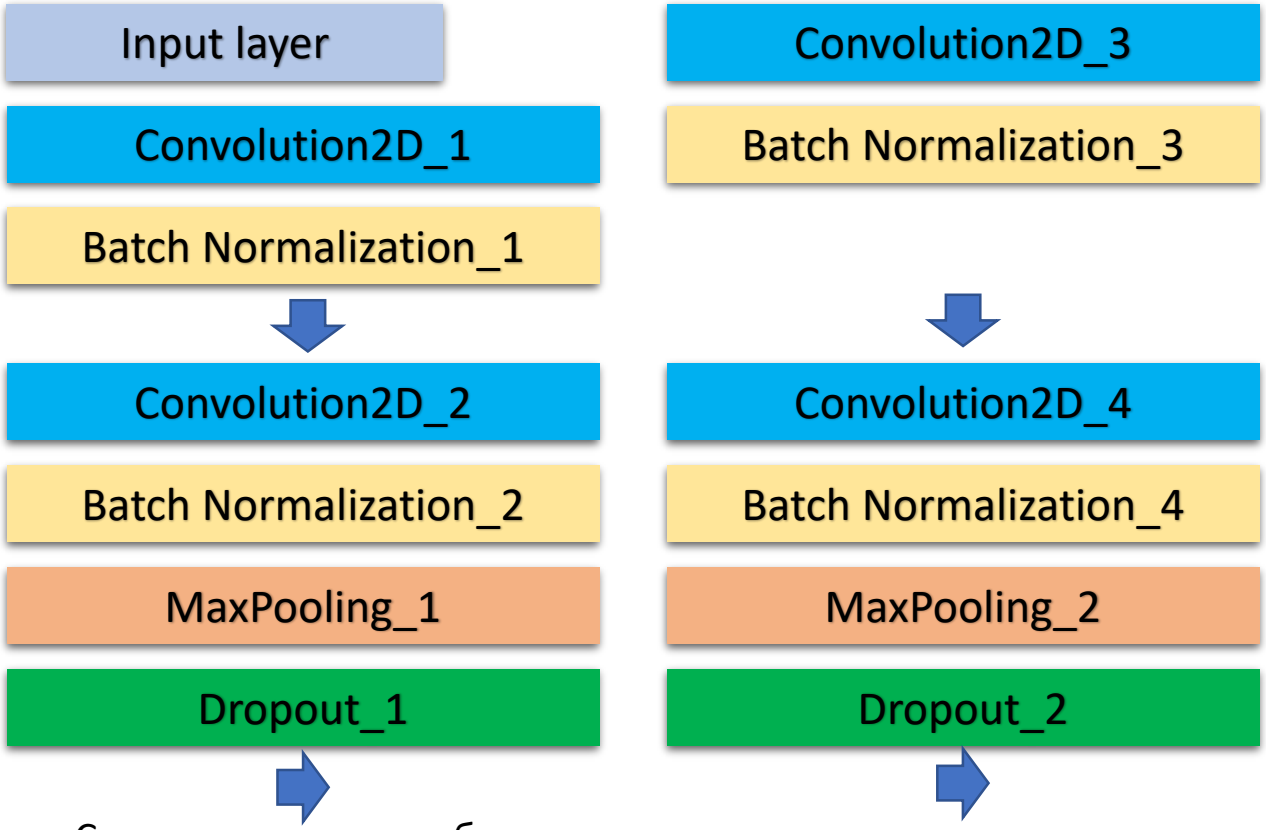


Возможные модели для анализа изображений:

- Многослойный перцептрон (MLP) – каждый канал пикселя считывается независимым входным параметром. MLP на прямую не преднозначен для распознавания изображений
- **Глубокая сверточная сеть (CNN). Используем KERAS на базе TensorFlow**

Структура сверточной нейронной сети (CNN)

- С целью ускорения обучения после каждого сверточного выполняется нормализация промежуточных данных **Batch Normalization**
- После каждого слоя подвыборки и между полносвязными слоями добавлены **Dropout** с параметрами вероятности $p = 0.25$ и $p = 0.5$ соотвественно



Комментарии к модели:

- На вход сети подается массив размерности **(71,400,400,3)**
- 4 двумерных сверточных слоя с возрастающим количеством фильтров **(16, 32, 32, 64)**
- 2 слоя подвыборки для субдескретизации **MaxPooling**
- В качестве функции активации на всех промежуточных слоях использована функция **ReLu**
- С целью превращения вектора действительных чисел в вектор вероятностей в завершающем полносвязном слое в качестве функции активации использована функция **softmax**
- Чтобы перейти от двумерной матрицы к одномерному вектору, перед 1-м полносвязным слоем установлен **Flatten**

Результаты моделирования

Модель	Улучшения настройки	Размер выборки (тренировка/валидация)	Количество итераций	Точность на тренировке	Точность на валидации
Basic model_3 classes	-	21 / 6 (9 изоб. на класс)	5	0.57	0.67
Adv model_3 classes	BatchNorm + Early Stopping	21 / 6	12	0.95	1
Basic model_7 classes	-	56 / 15	25	0.11	0.05
Adv model_7 classes	BatchNorm + Early Stopping	56 / 15	30	0.16	0.06

Выбор метрики: поскольку классы сбалансированы (количество изображений в каждом классе одинаковое) подходящей метрикой будет *точность (**accuracy**)*.

Размер ***Batch Size = 10***
(из-за ограничений CPU)

Вывод:

- При решении задачи классификации из 3-х классов модель показывает высокую точность на валидации.
- При увеличении количества классов точность резко падает даже несмотря на настройку и использование инструментов регуляризации.

Рекомендации:

- Представленных данных не достаточно для построения и тренировки модели. (Пример: для получения точности >0,76 для базовой модели CNN на данных CIFAR-10 требуется порядка 60000 изображений).
- Возможно порекомендовать увеличить набор данных до 2000…5000 изображений.
- Возможно использовать другие методы регуляризации, например **L2-регуляризацию**.

Перспективы поддержки 5000 сортов вина

Рекомендации:

- Для работы с большими объемами данных рекомендуется использовать средства online-learning такие как библиотека **Vowpal Wabbits** или **Apache Spark**.

- Для повышения точности также имеет смысл использовать **ансамбли** из нескольких сетей, каждая из которых вычисляет средний вектор предсказаний, который далее усредняется.

Вариант организации команды:

Должность	Количество	Обязанности
SQL Developer	1	Реализация функционала БД и ответственность аза хранение данных в требуемых форматах. Создание и оптимизация запросов SQL.
Data Engineer	2	Предобработка набора данных (обрезание изображений, нормализация, PCA, дескриптивный статистический анализ)
Data Scientist	2	Построение, оптимизация и тестирование предиктивной модели
Front-end Developer	1	Создание клиентской части Web приложения: верстка шаблона и создание пользовательского интерфейса
Back-end Developer	1	Проектирование архитектуры сервера и поддержка его архитектуры