

# Генерация обучающей выборки с помощью ограниченного набора данных

Крейнин Матвей

Московский физико-технический институт  
Кафедра интеллектуальных систем

*Научный руководитель:* к.ф.-м.н. А.В. Грабовой  
2024

# Проблематика работы

## Проблема

Недостаточное количество медицинских данных для обучения моделей сегментации, а также дороговизна разметки и неконсистентность разметки.

## Цель

Разработать процесс по генерации разметки и отбору, как и с априорным знанием о задаче, так и без них, качественных сэмплов для дальнейшего их использования .

## Решение

Обучение моделей сегментации на доступном наборе данных, разметка и отбор с их помощью, обучении модели на полученных данных и сравнение метрики с изначальным набором данных.

# Постановка и мотивация задачи

**Мотивация** Есть 3 гипотезы об улучшении качества моделей:

- 1 Увеличение размера модели.
- 2 Увеличение количества данных, на которых модель обучается.
- 3 Увеличение количества компьютера.

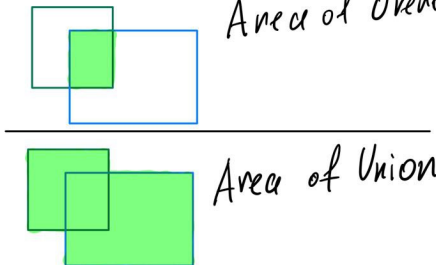
Хотим использовать вторую гипотезу и увеличить размер данных для обучения «бесплатно».

**Постановка задачи** Есть набор размеченных экспертами данных для задачи сегментации.

$X \in \mathbb{R}^{H \times W \times D}$ ,  $Y \in \mathbb{R}^{C \times H \times W \times D}$  – набор пар картинок, разметка

Задача: построить процесс разметки изображений, у которых нет разметки, и отбора, подходящих сэмплов, без участия эксперта, основанный на имеющемся ограниченном наборе данных. Отбор сложных кейсов для разметки экспертами.

# Напоминание IoU

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$


The diagram illustrates the formula for Intersection over Union (IoU). It is divided into two parts by a horizontal line. The top part shows two overlapping rectangles: a green one on the left and a blue one on the right. The intersection of these two rectangles is shaded in a darker green. The bottom part shows the same two rectangles, but both are filled with green, representing their union.

Рис.: IoU

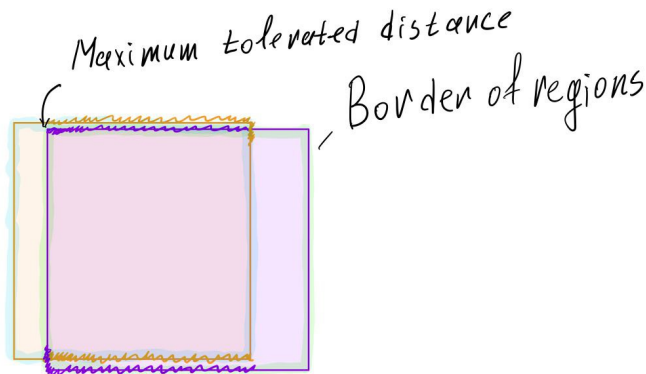
# Напоминание Hausdorff distance

$(M, d)$  - метрическое пространство. Для непустого множества  $X, Y \subset M$ , расстояние Хаусдорфа определяется как:

$$d_H(X, Y) = \max \left\{ \sup_{x \in X} d(x, Y), \sup_{y \in Y} d(X, y) \right\}$$

image3.jpg

# Напоминание Surface Distance



$$\text{Surface dice} = \frac{\begin{array}{c} \text{---} \\ | \\ \text{---} \end{array} + \begin{array}{c} \text{---} \\ | \\ \text{---} \end{array}}{\begin{array}{c} \square \end{array} + \begin{array}{c} \square \end{array}}$$

# Предложенная схема

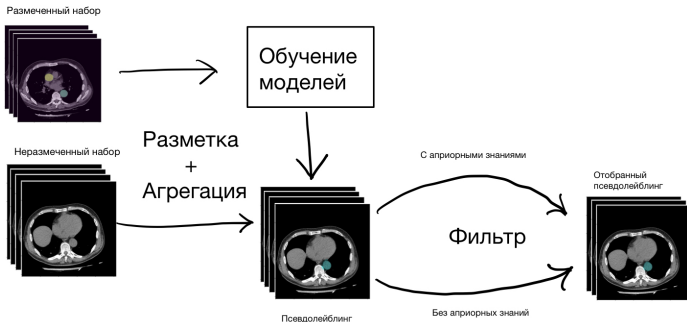


Рис.: Предложенная схема

# Критерии отбора без априорного набора знаний

Предлагаемые методы отбора:

- 1 Высокая волатильность между агрегированным ответом и отдельными моделями.
- 2 Высокая попарная волатильность между моделями.
- 3 Низкая согласованность между агрегированным ответом и отдельными моделями.
- 4 Низкая попарная согласованность между моделями.

Метрики, на основании, которых можно судить о высокой волатильности и низкой согласованности:

- 1 IoU - может быть неинформативен.
- 2 Hausdorff Distance - информативен только относительно других объектов выборки.
- 3 Surface Dice - информативен при минимальных знаниях о решаемой задаче.



# Предложенные методы отбора на практике

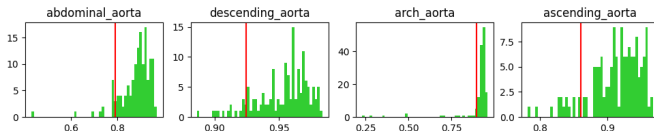


Рис.: Отбор "хороших" и "плохих" размеченных данных на основании среднего значения IoU между агрегированным ответом и ответом каждой из модели

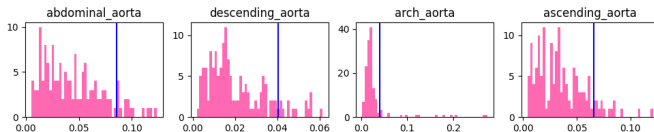


Рис.: Отбор «хороших» и «плохих» размеченных данных на основании волатильности разности между IoU между агрегированным ответом и ответом каждой из модели

# Предложенные методы отбора на практике

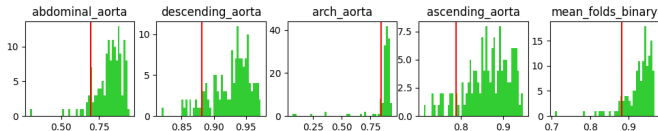


Рис.: Отбор "хороших" и "плохих" размеченных данных на основании среднего попарного значения IoU между моделями

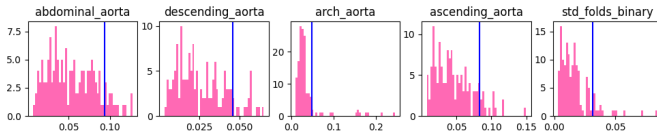


Рис.: Отбор "хороших" и "плохих" размеченных данных на основании волатильности попарного значения IoU между моделями

# Критерии отбора с априорным набором знаний

Примеры априорных знаний о решаемой задаче:

- 1 Человек (как правило) непрерывный, поэтому при отсутствии этого свойства возникают вопросы.
- 2 Эвристические правила на основании количества компонент предсказанной маски сегментации органа/патологии.
- 3 Физические соображения о возможном расположении компонент относительно друг друга и относительно других патологий/органов.

- 1 Данные для обучения на целевой задачи
- 2 Для претрейна модели и дотюнивание на данных от экспертов
- 3 Итерационный процесс улучшения модели и получения более качественных данных
- 4 Отбор сложных и уникальных кейсов

# Постановка эксперимента

- 1 SegResNet (UNet like архитектура) в качестве модели со стохастической глубиной (по некоторым результатам это SOTA в задаче 3D сегментации на данный момент).
- 2 Скользящее окно с пересечениями по всей картинке (изображение может быть какого угодно размера).
- 3 Выбираем медианный спейсинг между вокселями.
- 4 Выбираем максимальный медианный спейсинг, который доступен на наших вычислительных ресурсах.
- 5 Приоритизируем размер патча к размеру батча.
- 6 Клип картинки выбираем по 1 и 99 перцентилю интенсивности маски.
- 7 На инференсе используем пересечение в 0.5 между скользящим окном, транспонируем картинки по всем трем осям и используем гауссиану для сглаживания.

# Результаты эксперименты на псевдолейблнге

Набор данных	IoU
Отобранный псевдолейблинг	85.9 $\pm$ 0.57
Чистые данные	84.1 $\pm$ 0.39
Дообученный псевдолейблинг	86.7 $\pm$ 0.62

**Таблица:** Сравнение результатов обучения на целевой задаче локализации органа

Набор данных	IoU
Отобранный псевдолейблинг	88.9 $\pm$ 0.25
Чистые данные	87.4 $\pm$ 0.14
Дообученный псевдолейблинг	89.3 $\pm$ 0.33

**Таблица:** Сравнение результатов обучения на целевой задаче сегментации органа после локализации

- 1 Поставлена формулировка процесса.
- 2 Поставлены эксперименты и обучены модели для псевдолейблинга.
- 3 Реализован код по отбору «хороших» и «плохих» данных.
- 4 Поставлены эксперименты, которые не отвергают гипотезу о том, что увеличение количества данных в обучении с псевдолейблингом, приводит к итогову улучшению качества.
- 5 Выполнены эксперименты на закрытом датасете, которые не отвергают изначальную гипотезу.
- 6 Для каждой модели предложен алгоритм классификации новых траекторий
- 7 Поставлены первые вычислительные эксперименты по восстановлению параметров динам. систем и классификации

Будет сформулирована и описана математическая теория, дописана статья и поставлены эксперименты на открытом датасете.