

Командный практический тур НТО по профилю «Инфохимия»

Оглавление

Введение	2
Требования к команде	2
Оборудование и программное обеспечение	2
Описание задач	3
1 Задача (10 баллов)	3
2 Задача (10 баллов)	4
3 Задача (10 баллов)	5
4 Задача (15 баллов)	5
5 Задача (25 баллов)	6
6 Задача (25 баллов)	7
7 Задача (5 баллов)	9
Оценивание и подведение итогов	9

Введение

Ультразвуковое облучение жидкостей, например водно-спиртовых растворов, приводит к кавитации или образованию мелких пузырьков. Кавитационные пузырьки генерируются в реальных растворах без использования оптических ловушек, что делает систему максимально приближенной к реальным условиям. Под действием ультразвука пузырьки могут расти, колебаться и в итоге схлопываться или распадаться.

Оказывается, пузырьки в растворах разной концентрации (с разной поверхностной энергией и плотностью) проходят разные пути эволюции. Хотя трудно получить детерминированное описание того, как концентрация раствора влияет на динамику пузырьков, можно разделить изображения с различной концентрацией, применяя алгоритм искусственной нейронной сети. Это указывает на то, что методы искусственного интеллекта могут дополнять детерминированный анализ в неравновесных, почти нестабильных ситуациях.

Основной задачей финального этапа является обучить искусственную нейронную сеть определять октановое число бензина в водном растворе на основе изображений пузырьков.

Требования к команде

Задание разбито на отдельные задачи, которые необходимо распределить между участниками команды, в соответствии с их ролями.

Роли

1. Химик. Основные задачи - построение химической модели кавитационных пузырей в зависимости от содержания смеси.
2. Программист по машинному обучению. Основные задачи- анализ и обработка изображений, построение моделей и классификатора.
3. Программист-математик. Основные задачи – решение задач по математической обработке кавитационных пузырей и построение базы данных.

Теоретические задачи и задачи на программирование выполняются участниками стандартным образом, ответы предоставляются в виде текстовых файлов и кодов программ, формат которых задан в требованиях к выполнению заданий.

Оборудование и программное обеспечение

Требования к рабочему месту участника

Требования к компьютеру: не ниже Intel Core i5 (или аналогичный), минимум 8Gb ОЗУ, ОС Windows 7/10 x64, желательно наличие видеокарты 10 и выше серии для ускорения расчетов.

Скорость интернета: не менее 100 мбит/с.

Программное обеспечение, которое необходимо установить участнику до начала финала:

Для создания модели использовать язык программирования Python 3.8 и выше, менеджер пакетов Conda, фреймворк Pytorch и командную оболочку Jupyter Notebook.

Python 3.7 и выше (<https://www.python.org/>)

Conda (<https://www.anaconda.com/>)

Pytorch (<https://pytorch.org/>)

Jupyter Notebook (<https://jupyter.org/>)

Можно использовать дополнительные модули языка программирования Python. Numpy для матриц и общей математики, Matplotlib и Seaborn — для визуализации, Pandas — для работы с базами данных и создания фреймов данных, а openCV-Python — для предварительной обработки изображений.

Все компоненты используемых программы можно установить через командную строку или терминал Linux.

Оборудование

Оборудование находится на площадке проведения испытаний.

Перечень оборудования:

Персональный компьютер/ноутбук с установленной: ОС Linux, Windows, и Python.

Описание задач

Основная задача: создать модель машинного обучения, которая позволяет классифицировать фотографии кавитационных пузырей, полученные с помощью ультразвуковой обработки, по определению октанового числа в бензине (92,95,98).

1 Задача (10 баллов)

Условие

Ознакомьтесь со статьей: <https://pubs.acs.org/doi/full/10.1021/acs.jpcb.2c00948>

Ответить на вопросы:

- 1) С помощью каких параметров можно оценить отклик кавитационного пузырька на акустические колебания?
- 2) От изменения каких условий может зависеть отклик кавитационного пузырька на акустические колебания? И как изменения условий могут влиять на параметры кавитационного пузырька?
- 3) Какое применение ультразвука для качественного и количественного определения состава водных растворов представлено в статье? И как оно реализовано?

4) Какая архитектура нейронной сети наиболее релевантная для решения задачи определения концентрации спирта в водно-спиртовой смеси с помощью обработки изображений кавитационных пузырей? Предоставить развернутый ответ, с пояснениями почему именно данная нейронная сеть.

5) Предложить возможность улучшения алгоритма определения концентрации спирта с помощью методов машинного обучения и кавитационных пузырей.

Формат ответа: текстовый файл с ответами на вопросы в формате .pdf, названный по форме (номер_задачи_название_команды.pdf).

Основная роль: химик. **Вспомогательные роли:** программист по машинному обучению, программист математик.

Количество попыток: 1

Критерии оценивания

- Максимальное количество баллов за 1 задачу 10 баллов. За полный ответ на вопросы № 1, 2, 3, 4, 5 ставится по 2 балла.

2 Задача (10 баллов)

Условие

Сделайте раскадровку видео файлов <https://disk.yandex.ru/d/YAGwL12IFAo2Mg>, для получения около 67000 изображений.

Справочные материалы: библиотека языка программирования Python OpenCV

Основные роли: химик, программист по машинному обучению, программист математик.

Формат ответа:

Скрипт, названный по форме (номер_задачи_название_команды.py) на языке программирования Python версии 3.7 или выше, выполняющий раскадровку видео формата mp4(или avi), которое лежит в папке, расположенной рядом со скриптом и папка должна быть названа “input_data_raw”, и сохраняющий полученные кадры в формате jpg в папку рядом со скриптом с названием “output_data_raw”.

Скрипт должен быть формата .py

Количество попыток: 1

Критерии оценивания

Скрипт на языке программирования Python версии 3.7 или выше, выполняющий задачу в соответствии с условием. При соблюдении формата ответа ставится 10 баллов. При невозможности запуска скрипта, несоблюдения форматов входных и выходных данных, ставится 0.

Содержание критерия	Баллы
Ответ полностью соответствует критериям	10
Скрипт не запускается, формат ответа не соответствует критериям оценивания	0
<i>Максимальный балл</i>	10

3 Задача (10 баллов)

Условие

Напишите скрипт, который при запуске через командную строку будет на вход получать размер изображения в формате двух чисел, написанных через пробел, и приводить все изображения к входному размеру.

Пример.

Размер изображения 224 на 224

Вход: 224 224

Справочные материалы: библиотека языка программирования Python OpenCV

В качестве ответа предоставьте скрипт, который приводит изображения к определенному размеру.

Основная роль: программист математик. **Вспомогательные роли:** химик, программист по машинному обучению.

Формат ответа: Скрипт, названный по форме (номер_задачи_название_команды.ру), на языке программирования Python версии 3.7 или выше, соответствующий условию задачи. На вход программе с помощью CLI должны подаваться два числа, соответствующие размерности выходного изображения. Входные изображения формата jpg, полученные в задании 2, должны находиться в папке, находящейся рядом со скриптом, “output_data_raw”, измененные изображения должны сохраняться в формате jpg в папку, находящуюся рядом со скриптом “output_data”.

Количество попыток: 1

Критерии оценивания

Скрипт на языке программирования Python версии 3.7 или выше, выполняющий задачу в соответствии с условием. При соблюдении формата ответа ставится 10 баллов. При невозможности запуска скрипта, несоблюдения форматов входных и выходных данных, ставится 0.

Содержание критерия	Баллы
Ответ полностью соответствует критериям	10
Скрипт не запускается, формат ответа не соответствует критериям оценивания	0
<i>Максимальный балл</i>	10

4 Задача (15 баллов)

Условие

Используя исходные изображения или изображения, полученные с помощью скрипта задачи 2, создайте алгоритм, позволяющий выявлять контуры кавитационных пузырей на изображении. Для этого, необходимо написать скрипт, позволяющий получать из каждого изображения площадь, радиус, и длину окружности кавитационного пузыря.

Рекомендовано использовать кадры или видео, в которых пузыри четко сегментированы или когда наблюдается один пузырь.

Справочные материалы: библиотеки языка программирования Python OpenCV, NumPy

В качестве ответа предоставьте скрипт, который изменяет входные изображения, рисуя на них контуры кавитационного пузыря, и формирующий таблицу параметров формата .csv, в которой будет содержаться информация (метрики): «номер изображения»-«площадь»-«радиус»-«длина окружности»-«октановое число бензина».

Номер/название изображения	Площадь (пиксель)	Радиус (пиксель)	Длина окружности (пиксель)	Октановое число
1	S1	R1	L1	N1
2	S2	R2	L2	N2
...				

Основная роль: программист математик. **Вспомогательные роли:** химик, программист по машинному обучению.

Формат ответа: Скрипт, названный по форме (номер_задачи_название_команды.ру), на

языке программирования Python версии 3.7, берущий на вход изображения формата jpg, полученные в задаче 2 или 3, находящиеся в папке рядом со скриптом с названием “input_data”. На выходе скрипта должен получаться файл формата csv соответствующий условию задачи с названием “output_data.csv”

Количество попыток: 1

Критерии оценивания

Скрипт на языке программирования Python версии 3.7 или выше, выполняющий задачу в соответствии с условием. При соблюдении формата ответа ставится 10 баллов. При невозможности запуска скрипта, несоблюдения форматов входных и выходных данных, ставится 0.

Содержание критерия	Баллы
Ответ полностью соответствует критериям	15
Полученная таблица содержит три из четырех требуемых параметров (площадь, радиус, длина окружности, октановое число)	10
Полученная таблица содержит два из четырех требуемых параметров (площадь, радиус, длина окружности, октановое число)	5
Полученная таблица содержит один из четырех требуемых параметров (площадь, радиус, длина окружности, октановое число)	2,5
Скрипт не запускается, формат ответа не соответствует критериям оценивания	0
<i>Максимальный балл</i>	15

5 Задача (25 баллов)

Условие

Используя таблицу, полученную в задаче 3, реализуйте модель классификации фотографий кавитационных пузырей по содержанию октанового числа (92, 95, 98) с помощью методов классического машинного обучения.

Справочные материалы: библиотека языка программирования Python Scikit-Learn.

В качестве ответа представить скрипт Python с решенной задачей, и таблицу в формате .csv, где будут отображены: название модели, precision, recall, f1-score (примеры данных параметров можно увидеть в функции библиотеки `scikit-learn.metrics.classification_report`).

Пример таблицы,

Название модели	Precision	Recall	f1-score
Случайный лес			
Метод опорных векторов			
...			

Основная роль: программист по машинному обучению. **Вспомогательные роли:** химик, программист математик.

Формат ответа: Скрипт, названный по форме (номер_задачи_название_команды.py), на языке программирования Python версии 3.7, получающий на вход данные в формате csv с названием “output_data.csv”, лежащие рядом со скриптом, полученные в задаче 4. На выходе должна получиться таблица формата csv с названием “output_metrics.csv” с наполнением в соответствии с условием, лежащая рядом со скриптом.

Количество попыток: 1.

Критерии оценивания

Скрипт на языке программирования Python версии 3.7 или выше, выполняющий задачу в соответствии с условием. При соблюдении формата ответа ставится 10 баллов и производится оценивание точности модели по f1-score. При невозможности запуска скрипта, несоблюдения форматов входных и выходных данных, ставится 0. За каждую обученную модель ставится 1 балл, максимальное количество моделей 10. В соответствии с метрикой f1-score ставится максимальный балл за задачу при построении рейтинга моделей всех участвующих команд. Рейтинг строится по максимальной точности из всех представленных моделей по метрике f1-score. Количество баллов за точность варьируются от 0 до 15, т.е команда набравшая наибольшую точность получает 15 баллов, остальные команды получают баллы, исходя из их позиции в рейтинге. В случае обучения модели на тренировочном, валидационном и тестовом датасете, с целью получения 100% точности, задания не засчитываются.

Содержание критерия	Баллы
Ответ полностью соответствует критериям	25*
Построено максимальное количество моделей, в таблице	10

заполнены все метрики	
Скрипт не запускается, формат ответа не соответствует критериям оценивания	0
<i>Максимальный балл</i>	25

* - 25 баллов получает команда набравшая наибольший показатель f1-score.

6 Задача (25 баллов)

Условие

Используя исходные изображения, или изображения, полученные в задаче 2, реализовать модель классификации фотографий кавитационных пузырей по определению октанового числа бензина (92,95,98) с помощью сверточной нейронной сети, обученной методом перенесенного обучения (предобученные нейронные сети, например, ResNet50), с точностью выше 70%.

Справочные материалы: библиотеки языка программирования Python версии 3.7 или выше TensorFlow или Pytorch

В качестве ответа предоставить скрипт Python с решенной задачей, и таблицу в формате csv., где будут отображены: название модели, precision, recall, f1-score (примеры данных параметров можно увидеть в функции библиотеки `scikit-learn.metrics.classification_report`).

Пример таблицы,

Название модели	Precision	Recall	f1-score
ResNet50			
VGG16			
...			

Основная роль: программист по машинному обучению. **Вспомогательные роли:** химик, программист математик.

Формат ответа: Скрипт, названный по форме (номер_задачи_название_команды.py), на языке программирования Python версии 3.7, получающий на вход данные в формате jpg из папки с названием "output_data" ("output_data_raw"), лежащие рядом со скриптом,

полученные в задаче 2 или 3. На выходе должна получиться таблица формата csv с названием “output_metrics_nn.csv” с наполнением в соответствии с условием, лежащая рядом со скриптом.

Количество попыток: 1.

Критерии оценивания

Скрипт на языке программирования Python версии 3.7 или выше, выполняющий задачу в соответствии с условием. При соблюдении формата ответа ставится 10 баллов и производится оценивание. При невозможности запуска скрипта, несоблюдения форматов входных и выходных данных, ставится 0. За каждую рассчитанную модель ставится 1 балл, максимальное количество моделей 10. В соответствии с метрикой f1-score ставится максимальный балл за задачу при построении рейтинга моделей всех участвующих команд. Рейтинг строится по максимальной точности из всех представленных нейронных сетей по метрике f1-score. Количество баллов за точность варьируются от 0 до 15, т.е команда набравшая наибольшую точность получает 15 баллов, остальные команды получают баллы, исходя из их позиции в рейтинге. В соответствии с метриками точности ставится максимальный балл за задачу. В случае обучения нейронной сети на тренировочном, валидационном и тестовом датасете, с целью получения 100% точности, задания не засчитываются.

Содержание критерия	Баллы
Ответ полностью соответствует критериям	25*
Построено максимальное количество нейронных сетей, в таблице заполнены все метрики	10
Скрипт не запускается, формат ответа не соответствует критериям оценивания	0
<i>Максимальный балл</i>	25

* - 25 баллов получает команда набравшая наибольший показатель f1-score.

7 Задача (5 баллов)

Условие

Сопоставить методы классического машинного обучения и методы перенесенного обучения для сверточных нейронных сетей и сделать выводы, на примере задачи классификации фотографий кавитационных пузырей по определению октанового числа бензина. В документе опишите преимущества и недостатки методов, ключевые данные проведенных анализов.

В качестве ответа предоставить файл в формате pdf. с развернутым ответом по условиям задачи 7.

Формат ответа: файл в формате pdf ., названный по форме (номер_задачи_название_команды.pdf).

Количество попыток: 1.

Основная роль: химик. **Вспомогательные роли:** программист по машинному обучению, программист математик.

Критерии оценивания

5 баллов ставится за полный аргументированный ответ, подкрепленный научными публикациями и собственными выкладками касательно данных и моделей.

Содержание критерия	Баллы
Ответ полностью соответствует критериям	5
Неполный ответ или отсутствие аргументации	0
<i>Максимальный балл</i>	5

Таким образом на оценку сдается решения по задачам 1-7, содержащее ответы по всем задачам, загруженное на облачное хранилище. Ссылку на облачное хранилище выдает администратор капитану команды.

Итоговое решение должно быть загружено в облачное хранилище до **16:00 (МСК) 11 апреля 2024** года.

Оценивание и подведение итогов

Максимальное количество баллов за командную часть - 100 баллов. Задачи оцениваются согласно критериям, указанным к задачам.

Максимальное количество баллов за каждую задачу, количество попыток на задачу и рекомендуемые роли указаны в таблице ниже.

Команда-победитель определяется по наибольшей сумме баллов за задачи командного этапа. При одинаковом количестве баллов выше по рейтингу ставится та команда, которая использовала меньшее количество попыток по всем задачам.

Задача №	Кол-во баллов	Кол-во попыток	Рекомендованные роли
1	10	1	Основная роль: химик. Вспомогательные роли: программист по машинному обучению, программист математик.

2	10	1	Основные роли: химик, программист по машинному обучению, программист математик.
3	10	1	Основная роль: программист математик. Вспомогательные роли: химик, программист по машинному обучению.
4	15	1	Основная роль: программист математик. Вспомогательные роли: химик, программист по машинному обучению.
5	25	1	Основная роль: программист по машинному обучению. Вспомогательные роли: химик, программист математик.
6	25	1	Основная роль: программист по машинному обучению. Вспомогательные роли: химик, программист математик.
7	5	1	Основная роль: химик. Вспомогательные роли: программист по машинному обучению, программист математик.

Максимальное количество баллов 100.