Научно-исследовательская работа

Предмет

**Тема работы**

***Выполнил****:*

*Сапаров Алмас*

учащийся 9 класса

Назарбаевской Интеллектуальной Школы г. Уральска

***Руководитель****:*

*ФИО научного руководителя*

с указанием должности и места работы,

возможной научной степени и звания

# **Аннотация**

Вспышка корона вируса COVID-19, а именно SARS-CoV-2, создала катастрофическую ситуацию по всему миру. Совокупная заболеваемость COVID-19 стремительно растет ото дня в день. Машинное обучение и модель корона вируса могут быть очень эффективно развернуты для прогнозирования роста эпидемии и для зарабатывания стратегии, политику по управлению ее распространением. Этот проект включает в себя анализирование текущею ситуацию COVID-19 в Казахстане, а также сравнение ситуаций корона вируса среди 3 стран, таких как Испания, Италия, Южная Корея. Далее в проекте исследуется глобальные данные корона вируса, а после используется и тренируется модель машинного обучения для прогнозирования чисел заболевших, выздоровевших, смертей глобально и локально. Для более глубокого понимания вируса, в исследования моделируется модель распространения корона вируса. Среди с распространением коронавируса, все чаще начали появляться случаи пневмоний. Из-за нехватки медицинских ресурсов и кадров число заболиваемых пневмонией может существенно возрасти. Для решений данной проблемы моделируется CNN которая сумеет точно и быстро сделать диагноз. Это может гарантировать своевременный доступ к лечению и сэкономить необходимое время и деньги для тех, кто уже находится в бедности.

**Введения**

В декабре 2019 года в Ухане, провинция Хубэй, Китай, была обнаружена новая цепь корона вируса (SARS-CoV-2), вызывающая тяжелый и потенциально смертельный респираторный синдром, то есть [COVID-19](https://www.who.int/news/item/27-04-2020-who-timeline---covid-19). С тех пор он стал пандемией, объявленной Всемирной организацией здравоохранения (ВОЗ) 11 марта, которая распространилась по всему миру. ВОЗ опубликовала на своем веб-сайте предварительное руководство по общественному здравоохранению для стран, которые должны бороться с пандемией. С тех пор инфекционное заболевание стало угрозой общественному здоровью. Италия и США серьезно пострадали из-за COVID-19. Национальные правительства вынуждают миллионы людей оставаться в самоизоляции и в тяжелых условиях. Заболевание быстро распространяется во многих странах мира. Из-за отсутствия подходящего лекарства или вакцины, в настоящее время социальное дистанцирование, самоизоляция и ношение лицевой маски стали наиболее широко используемой стратегией смягчения последствий пандемии и борьбы с ней.

В этом контексте необходима модель для оценки передачи заболеваний, выздоровления, смертности и других важных параметров. Также на ранних стадиях новой вспышки инфекционного заболевания крайне важно понимать динамику передачи инфекции. Оценка изменений в передаче с течением времени может дать представление об эпидемиологической ситуации и определить, оказывают ли меры борьбы со вспышкой измеримый эффект. Такой анализ может дать информацию для прогнозов о потенциальном будущем росте, помочь оценить риск для других стран.

Риск пневмонии огромен для многих, особенно в развивающихся странах, где миллиарды людей сталкиваются с энергетической бедностью и полагаются на загрязняющие виды энергии. По оценкам ВОЗ, ежегодно более 4 миллионов случаев преждевременной смерти происходят от болезней, связанных с загрязнением воздуха в домашних условиях, включая пневмонию. Ежегодно более 150 миллионов человек заражаются пневмонией, особенно дети в возрасте до 5 лет. В таких регионах ситуация может усугубиться в бедных странах из-за нехватки медицинских ресурсов и кадров. Например, в 57 странах Африки нехватка врачей и медсестер составляет 2,3 миллиона человек. Решением данной проблемы может послужит нейронная сеть для автоматического определения наличия у пациента пневмонии по рентгеновским снимкам грудной клетки.

**Основная часть**

Таймлайн коронавируса в Казахстане

Первые случаи корона вирусной инфекции COVID-19 были зарегистрированы на территории Казахстана 13 марта 2020 года. По официальной статистике, на 30 июля в Казахстане зарегистрированы 100 000 случаев заражения, 73 702 пациент выздоровел, 1269 пациентов не удалось спасти.

16 марта до 11 мая 2020 года в Казахстане Для предотвращения распространения заболевания был введён режим чрезвычайного положения

5 июля 2020 года в Казахстане начал действовать режим локдауна. Все объекты закрылись, кроме супермаркетов, аптек, кафе (с сохранением социального дистанцирования), аэропортов.

8 июля 2020 года президент Казахстана Касым-Жомарт Токаев объявил 13 июля днём национального траура по умершим из-за вируса COVID-19. *«Я выражаю соболезнования всем тем, кто потерял близких, поэтому 13 июля объявляю днем национального траура»*, — сказал президент Токаев во время телеобращения.

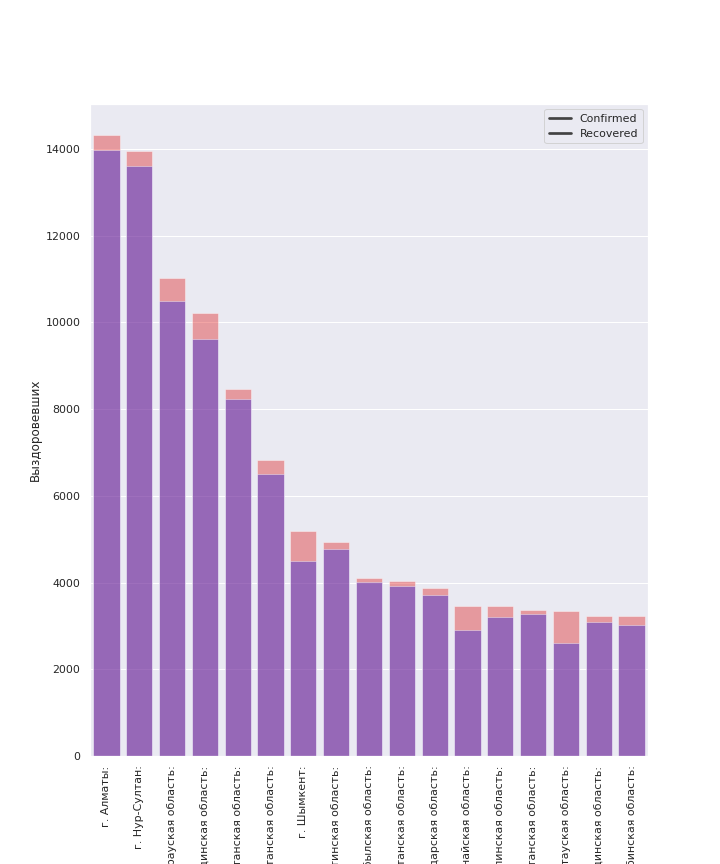
Анализирование текущего состояния Казахстана

Проанализируем текущего состояния Казахстана по числу заражённых,

умерших, активных и выздоровевших корона вирусом для дальнейшего исследование. Для анализа будут использоваться данные из открытых и официальных ресурсов РК.

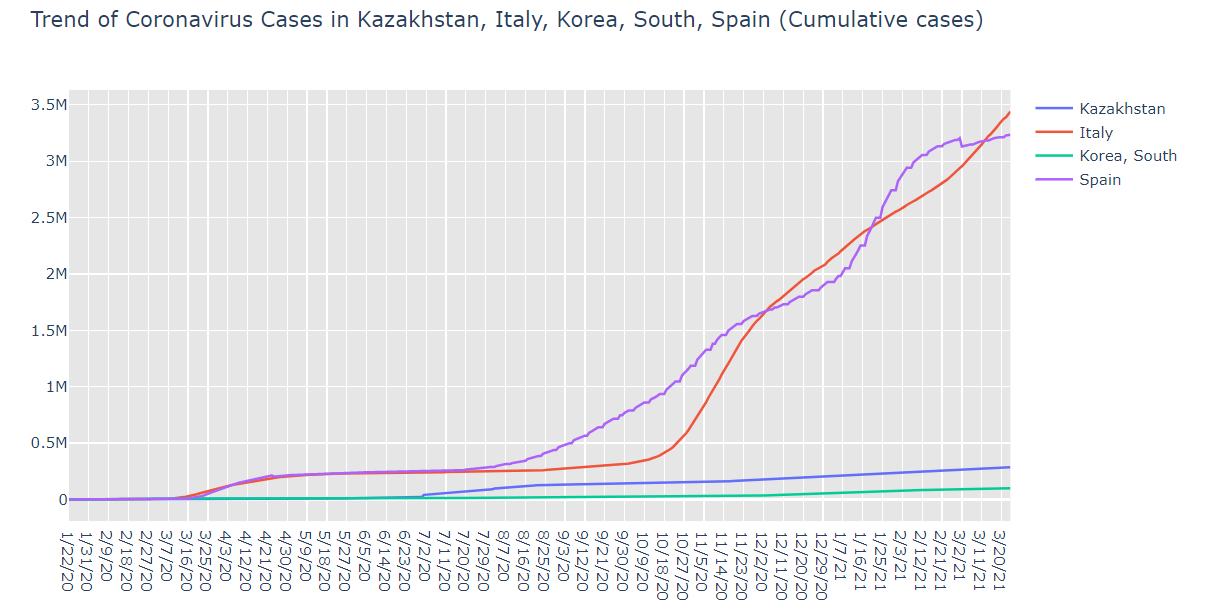
В визуализации можно наглядно распознать что огромное скопления зараженных сконцентрировано в крупных мегаполисов Казахстана, таких как г.Алматы и г.Нурсултан. Однако большую долю активно заражённых составляют Мангыстаукая область, что может спровоцировать увеличение заражённых.

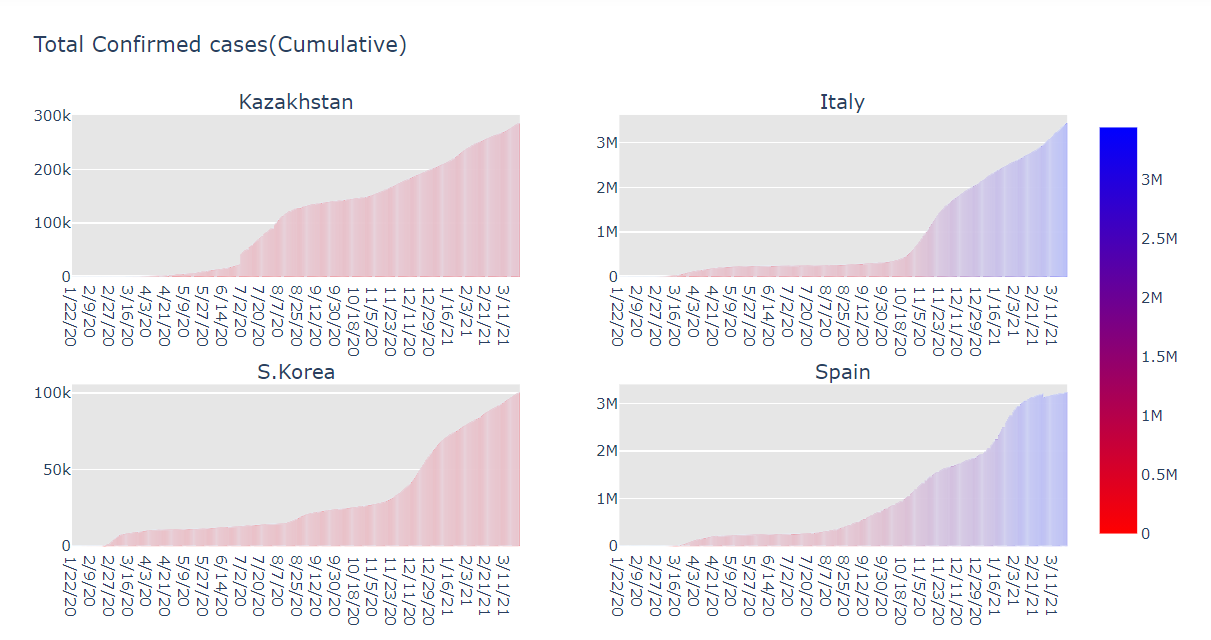
Для более детального воспринятия было вуалировано географическое представление заражённых на карте Казахстана.

Как можно заметить по первому графику, большое количество людей все же выздоравливает от коронавируса. Чтобы оценить и сравнить действительность гипотезу, был создан график.

Похожа ли тенденция распространения на Италию / Южную Корею / Испанию?

Чтобы получить данные, мы будем использовать репозиторий данных COVID-19 Центра системных наук и инженерии (CSSE) в Университете Джона Хопкинса, который доступен [здесь](https://github.com/CSSEGISandData/COVID-19). Из этого набора данных мы возьмем количество инфицированных для каждой страны благодаря библиотеке pandas, чтобы сравнить их между собой



Как мы можем наблюдать распространение сильно зависит от устройства здравоохранения и общего население в каждой конкретной стране. Цифровые значения забеливаемых Южной Корей значительно меньше, чем остальные страны. К 10 марту по официальному ресурсам было сделано 210 144 теста, это более 4000 тестов на миллион жителей. Также Южная Корею отличилось цифровой политикой, которое отслеживает всех контактировавших с больными. Такая политика совершенна невозможна в странах ЕС, с его законами о защите личных данных, среди с чем показатель заболиваемых значительно больше, чем Южная Корея

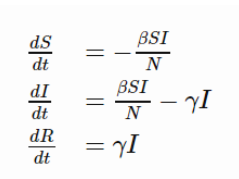
Также можно заметить вспышку число заболиваемых в Италий. 13 ноября 2020 года было зарегистрировано 40902 новых случая заболевания, что представляет собой самый большой рост ежедневных случаев заболевания в Италии с начала пандемии.

Моделирования пандемий COVID с помощью математической модели SIR для оценки значения Rt

Одной из основных характеристик эпидемии является эффективное число воспроизводимых (Rt), которое указывает количество людей, которые каждый инфицированный человек будет инфицировать в любой момент времени. Возможность оценить Rt - важная задача, потому что это число определяет, будет ли эпидемия расти (Rt> 1) или начнет снижаться (Rt <1)

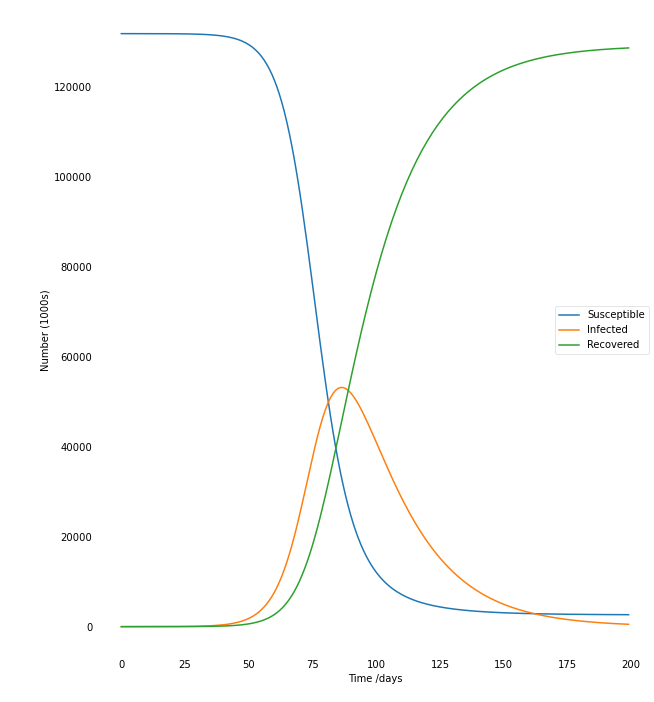
Для моделирования пандемий COVID я использовал математическую модель SIR. Эта модель состоит из трёх переменных:

* S (t) - Восприимчивый - количество людей, которые потенциально могут заразиться
* I (t) - Infected - количество инфицированных
* R (t) - Вылечены - количество людей, не восприимчивых к инфекции

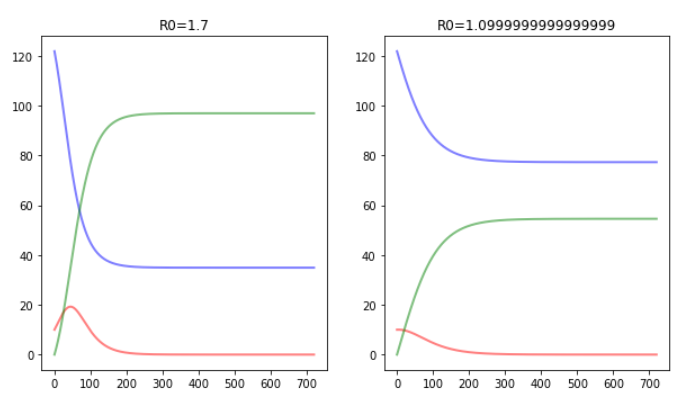
 Каждая переменная является функций времени, благодаря чему мы можем сформулировать следующие дифференциальные уравнения, описывающие поведение модели. Эта модель зависит от двух параметров:

* β - частота контактов, и мы предполагаем, что за единицу времени каждый инфицированный человек войдет в контакт с βN людьми. Среди этих людей доля восприимчивых людей составляет S / N, поэтому скорость, с которой возникают новые инфекции, определяется как -βSI / N.
* γ - скорость выздоровления, а число 1 / γ определяет количество дней, в течение которых человек остается инфицированным.
* Таким образом, термин γI определяет скорость, с которой инфицированные люди переходят от инфицирования к выздоровлению. (Hill.)

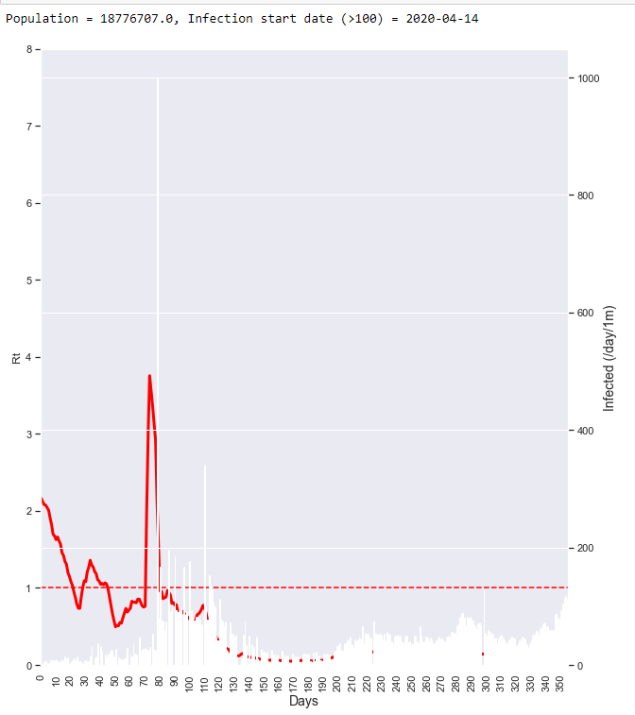
Смоделируем пандемию в городе с населением 131800 человек точно так же, как и в Уральске, установив значение частоты контактов - 0,2, а период восстановления - 20 дней.



При моделировании SIR обычно оценивают базовое репродуктивное число эпидемии R₀ = β / γ. В нашем примере R₀ = 4, что очень похоже на оценочные цифры для заражения COVID. Таким образом, эта модель говорит нам, что без каких-либо карантинных мер в таком городе, как Уральск, мы окажемся в итоге с примерно 55000 тысяч человек инфицированных людей за один раз примерно за 80 дней

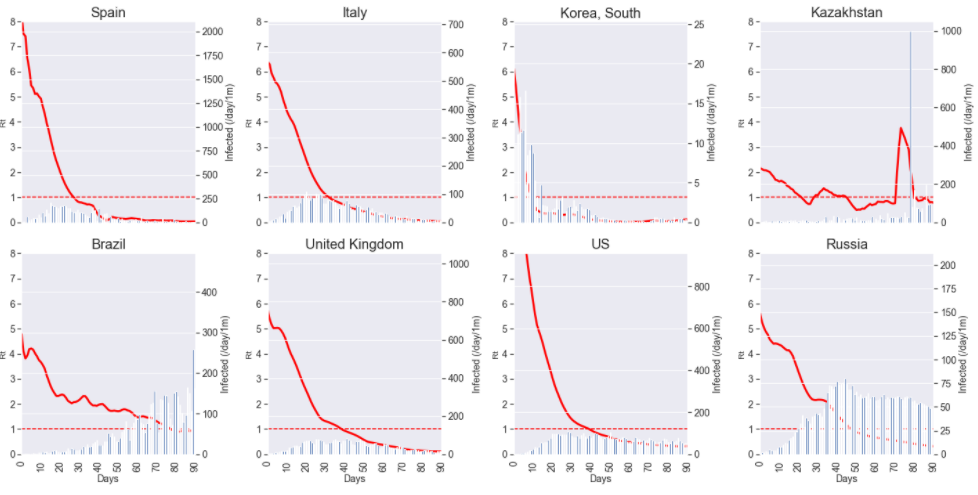
 Как можно заметить, чем ниже R, тем меньше количество людей, инфицированных одновременно, что помогает минимизировать перегрузку системы здравоохранения.

В нашем примере мы несколько свободно взяли значения β и γ. Однако, если у нас есть реальные эпидемиологические данные, мы можем попытаться определить эти значения, подогнав полученную модель SIR так, чтобы она наиболее точно соответствовала реальным данным. Используя уже вышеперечисленную [базу данных](https://github.com/CSSEGISandData/COVID-19), можно вычислить Rt для каждой страны. Для вычисления мы возьмем начальную точку, скажем, когда число инфицированных превысило 1000. Затем мы сосредоточимся на периоде длинной 8 дней, пытаясь оптимизировать значение β и γ благодаря методу минимизации из scikit learn. Функция ошибки будет возвращять нормальный вектор различия вектора ежедневных случаев между вектором I + R. Так как 8 точек данных не хватает для оптимизаций β и γ, я взял фиксированное значение γ, в соответствии с [медицинскими исследованиям](https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf)и, которые указывают на среднее время восстановления 30 дней.

После получения значение Rt Казахстана, я изобразил его в виде графика. Как можно заметить пандемия быстро замедляется, когда значения Rt опускается до 1.

База данных хранить в себе информацию о 188 стран, давая свободу для различных исследований и сравнений.

Я изобразил 8 стран: Испания, Италия, Южная Корея, Казахстан, Бразилия, Великобритания, США и Россия

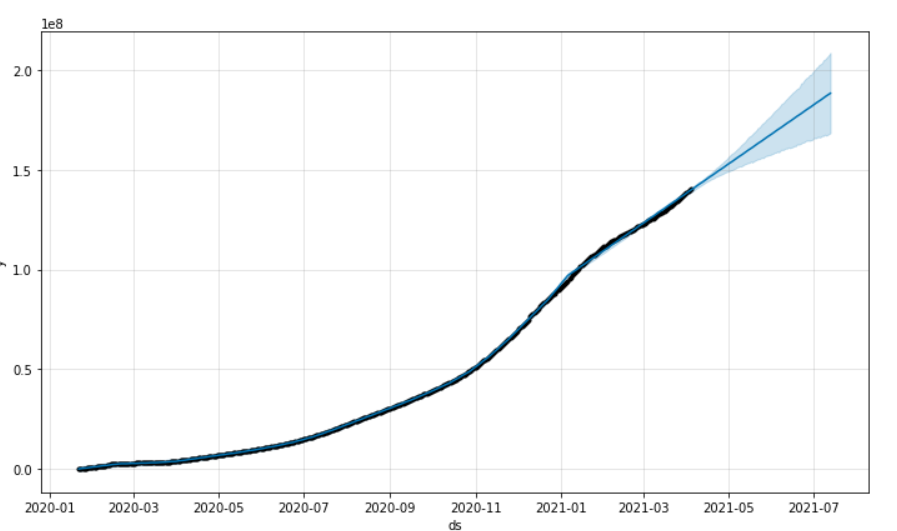


По этому графику можно сравнить:

* Насколько эффективны защитные меры в разных странах,
* Как быстро эпидемия берется под контроль (т.е. Когда график Rt становится ниже 1),
* Насколько серьезно это влияет на население в целом (по нормализованному количеству новых случаев)

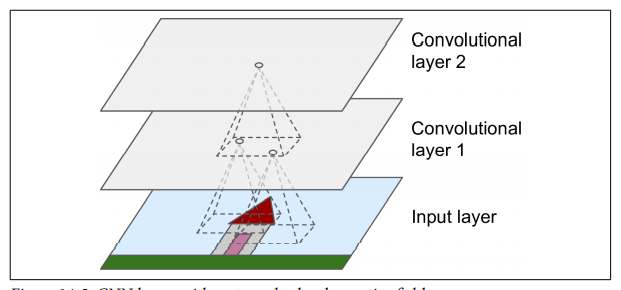
Прогнозирование тенденций заражение с помощью Prophet

Prophet — это программное обеспечение с открытым исходным кодом, выпущенное командой Facebook Core Data Science. Он лучше всего работает с временными данными, которые имеют сильные сезонные эффекты и несколько сезонов исторических данных. Prophet устойчив к отсутствию данных и сдвигам в тренде и, как правило, хорошо справляется с выбросами. В контексте прогноза пандемий это нейросеть окажется очень эффективной с вышеперечисленными причинами

Я создал прогноз подтвержденных случаев COVID-19 на 100 дней вперед с интервалом прогноза 95%.

Использование технологий CNN для классификаций пневмоний

Cверточная нейронная сеть (CNN) - это класс глубоких нейронных сетей, наиболее часто применяемых для анализа визуальных образов. В контексте распознавание пневмоний по рентгеновским снимкам, они окажутся очень эффективными. Сама сверточная нейронная сеть состоит из двух главных компонентах : сверточный и объединяющие слой. Отличием сверточных слоев в том, что он связан только с нейронами в его рецептивной области. Эта архитектура позволяет нейронной сети сконцентрироваться на небольших низкоуровневых свойствах в первых слоях, собирая их в более сложные объекты. (Géron, 2017) Эта иерархическая структура характерна для изображений реального мира, что является одной из причин, почему CNN так хорошо работают для распознавания изображений.

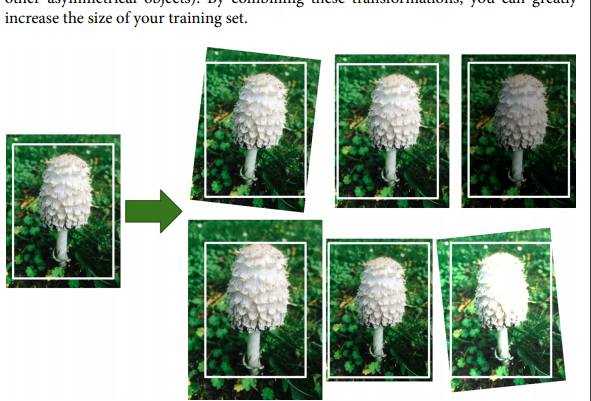


В качестве данных для тренировки я использовал открытые данные из сайта [Kaggle](https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia). В этом ресурсе насчитывается 5 863 рентгеновских снимка разделённые на 2 категорий. Рентгеновские снимки грудной клетки (передне-задние) были отобраны из ретроспективных когорт педиатрических пациентов в возрасте от одного до пяти лет из Медицинского центра женщин и детей Гуанчжоу, Гуанчжоу.



Нормальный рентген грудной клетки (левый снимок) показывает чистые легкие без каких-либо участков аномального помутнения на изображении. Бактериальная пневмония (центральный снимок) обычно проявляется фокальной долевой консолидацией, в данном случае в правой верхней доле (белые стрелки), тогда как вирусная пневмония (правый снимок) проявляется более диффузным «интерстициальным» паттерном в обоих легких.

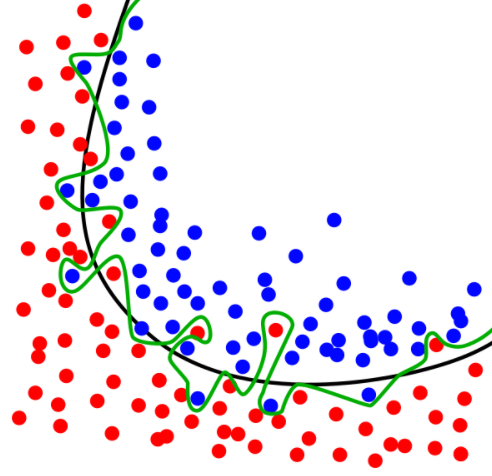
При всей своей эффективности нейронные сети зачастую требуют большое количество данных, в связи с чем я увеличил количество данных вида изменяя изначальный снимок. Практика увеличения данных - эффективный способ увеличить размер обучающей выборки. Дополнение обучающих примеров позволяет сети «видеть» более разнообразные, но все же репрезентативные точки данных во время обучения. (Géron, 2017)

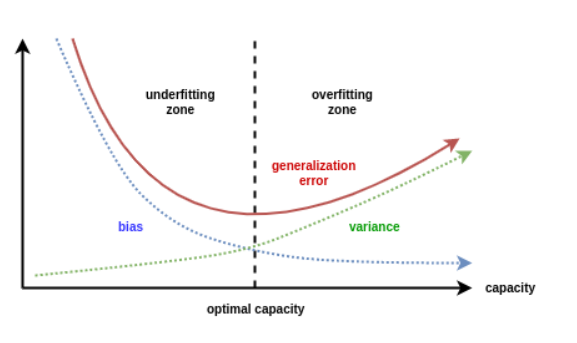


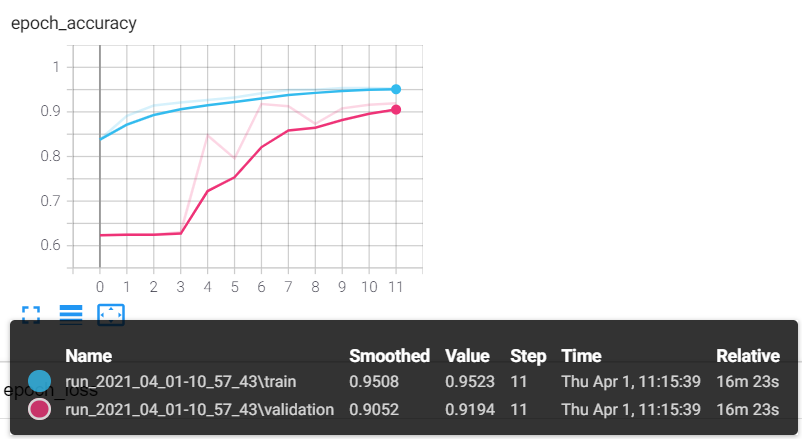
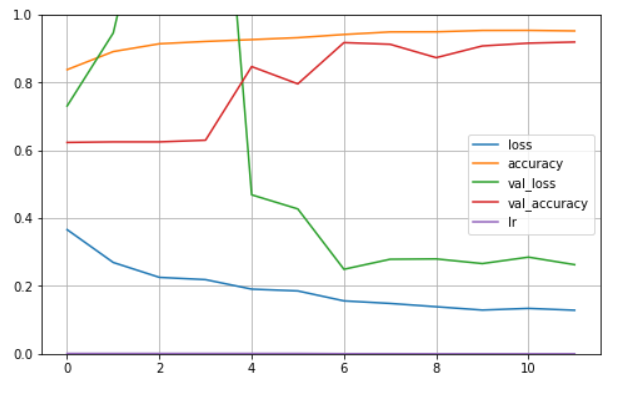
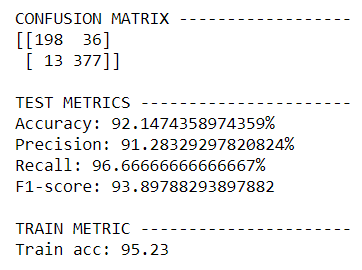
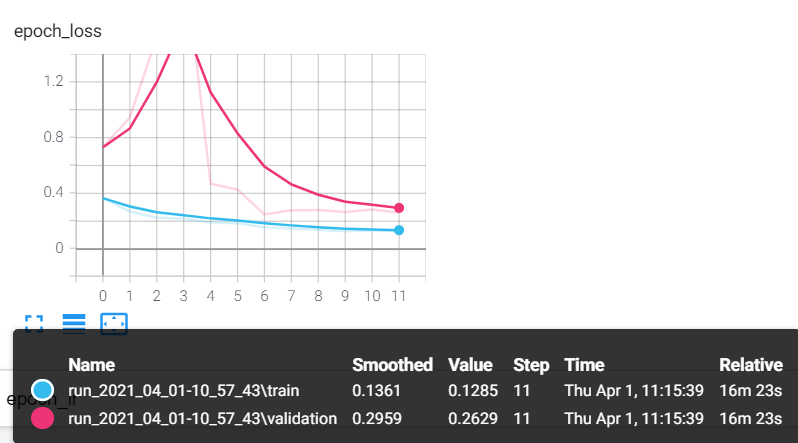
Далее я с моделировал архитектуру нейронной сети. Сама структура состоит из пяти сверточных блоков состоящих слоёв из сверточного, максимальной субдискретизации и пакетной нормализации. Далее идёт флаттен слои и четыре полно связные слои. Также в промежутках я использовал дропауты, чтобы уменьшить переобучение. В качестве функций активаций я использовал функцию relu, за исключеннием последнего слоя, где я использывал функцию активаций sigmoid. Я также использовал оптимизатор Адам и кросс-энтропию в качестве функций потерь

*Изображение выглядит как текст, парковка, счетчик, несколько

Автоматически созданное описание*

Вовремя обучения модели, нейронная сеть может часто переобучаться на данных, относительно плохо работая на примерах, не участвовавших в обучении. Это связано с тем, что при обучении модели, в обучающей выборке обнаруживаются некоторые случайные закономерности, которые отсутствуют в генеральной совокупности

Для избежание данной проблемы я использовал колбэк – EarlyStopping. Этот метод помогает модели в процессе обучений прекратить обучение, когда разрыв обобщение существенно ухудшается

Я обучил нейросеть на 10 эпох и сумел достигнуть точности 92.15% на проверочных данных. Также эта нейросеть сумела достичь 93.89% f1 очков. Конечно этот результат далёк от идеального, но стоит отметить успех глубокого обучения в разнообразных проблем реального мира.

Заключение

Благодаря этому проекту мы смогли оценить реакцию и меры здравоохранений разных стран. Вовремя сравнение мы сумели рассчитать Rt в зависимости от времени. В добавок, мы спрогнозировали тенденций заражение на месяц. Мы также смогли, достичь высоких успехов в распознавания пневмоний по рентгеновскому снимку. Конечно, результаты, которые мы достигли благодаря технологиям глубокого обучения далеки от идеала, но стоит отметить эффективность применение нейронной сети для решение глобальной проблемы.

Использованные литературы

* Géron, A. (2017). Hands-on machine learning with Scikit-Learn and TensorFlow concepts, tools, and techniques to build intelligentsystems O'Reilly Media Paperback .
* Christian Hill. Learning Scientific Programming with Python. — *Cambridge University Press, ISBN: 9781107428225*

*.*