МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМ. М. В. ЛОМОНОСОВА

ФИЗИЧЕСКИЙ ФАКУЛЬТЕТ

КАФЕДРА МАТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ И ИНФОРМАТИКИ

ГЕНЕРАЦИЯ ИЗОБРАЖЕНИЙ ДЛЯ АНАЛИЗА УСТОЙЧИВОСТИ НЕЙРОННЫХ СЕТЕЙ В ЗАДАЧЕ КЛАССИФИКАЦИИ

Курсовая работа

студента 435 группы

Клиентова Григория Алексеевича

Научные руководители:

д. ф.-м. н., профессор П. В. Голубцов

д. ф.-м. н., профессор РАН Дьяконов Александр Геннадьевич

Москва

2024

Оглавление

1. Введение

В последние годы машинное обучение стало неотъемлемой частью многих областей, включая компьютерное зрение, обработку естественного языка и медицинскую диагностику. Одной из ключевых задач в этих областях является классификация изображений, которая позволяет компьютерам автоматически определять содержание изображений. Однако, несмотря на значительные успехи в области классификации изображений, машинное обучение остается уязвимым для атак, известных как атаки с использованием adversarial examples.

Adversarial examples - это специально созданные входные данные, которые предназначены для обмана моделей машинного обучения. Такие примеры могут привести к тому, что модель выдаст неверный результат, даже если исходное изображение легко распознается человеком. Эта проблема стала предметом активных исследований, поскольку она ставит под сомнение надежность и безопасность систем машинного обучения.

В данной работе исследуется механизм adversarial attack, когда на вход сети подается некоторое универсальное изображение с примененным маскированием, которое модель будет относить к любому заданному классу, причем ответ модели будет зависеть от типа и расположения маски.

2. Цель работы

- Реализовать алгоритм генерации универсальных изображений для adversarial attack для различных типов маскирований
- Исследование результатов работы модели с применением таких изображений.

3. Используемая модель для атак

В данной работе используется модель ResNet-50 с предобученными весами IMAGENET1K_V2 из библиотеки PyTorch. Она имеет показатель точности top-1 порядка 80.858% на оригинальном датасете Imagenet и имеет 25.6 миллионов параметров. Модель принимает на вход изображение, а на выходе показывает вероятности принадлежности изображения ко всем классам.

4. Описание алгоритма для генерации изображений

В работе исследовалось изображение размером 64x64 с тремя каналами RGB. Для поиска был применен метод градиентного спуска. В качестве параметров для оптимизации использованы значения пикселей изображения в каждом из каналов. Псевдокод на Python для цикла обучения представлен ниже:

```
image = random_normal(3, 64, 64)
for epochs in range(EPOCHS):
    batch = []
    labels = []
    for class_ind in num_classes:
        batch.append(mask(image, class_ind))
        labels.append(class_ind)
    predictions = model(batch)
    loss = loss_function(predictions, labels)
    loss.backward()
    image.clip(0, 1)
```

Во всех экспериментах использовались следующие гиперпараметры:

- Количество эпох – не более 5000
- Функция ошибки – CrossEntropyLoss
- Оптимизатор весов – SGD с learning rate = 0.1
- Количество классов – от 2 до 62 с шагом 4.

5. Используемые маскирования

В работе были исследованы 3 типа маскирования изображения.

- Тип 1 - маскирование, когда все изображение делится на **k** равных частей и в зависимости от номера класса закрывается соответствующая ему часть изображения, то есть заполняется черными пикселями.
- Тип 2 - маскирование, когда все изображение делится на **k** равных горизонтальных полос и в зависимости от номера класса закрывается соответствующая ему полоса.
- Тип 3 - маскирование, когда для каждого класса генерируется своя маска, состоящая из черных пикселей, стоящих на случайных местах.

6. Эксперименты

В каждом из экспериментов исследовалась зависимости времени генерации изображения и получаемая точность в зависимости от количества требуемых классов, необходимых для нахождения на изображении. Точность рассчитывается как доля верно предсказанных классов.

6.1 Маскирование с разбиением изображения на k равных частей

В данном эксперименте для каждого класса закрывается своя часть изображения. Соответственно при применении данной маски от модели требуется отнести такое изображение к этому наперед заданному классу. Таким образом рецептивное поле (та часть изображения, на основе которое модель делает предположение о принадлежности к классу) будем состоять из всего изображения, за исключением маскированной прямоугольной части, расположенной согласно классу, к которому мы бы хотели, чтобы изображение было отнесено.

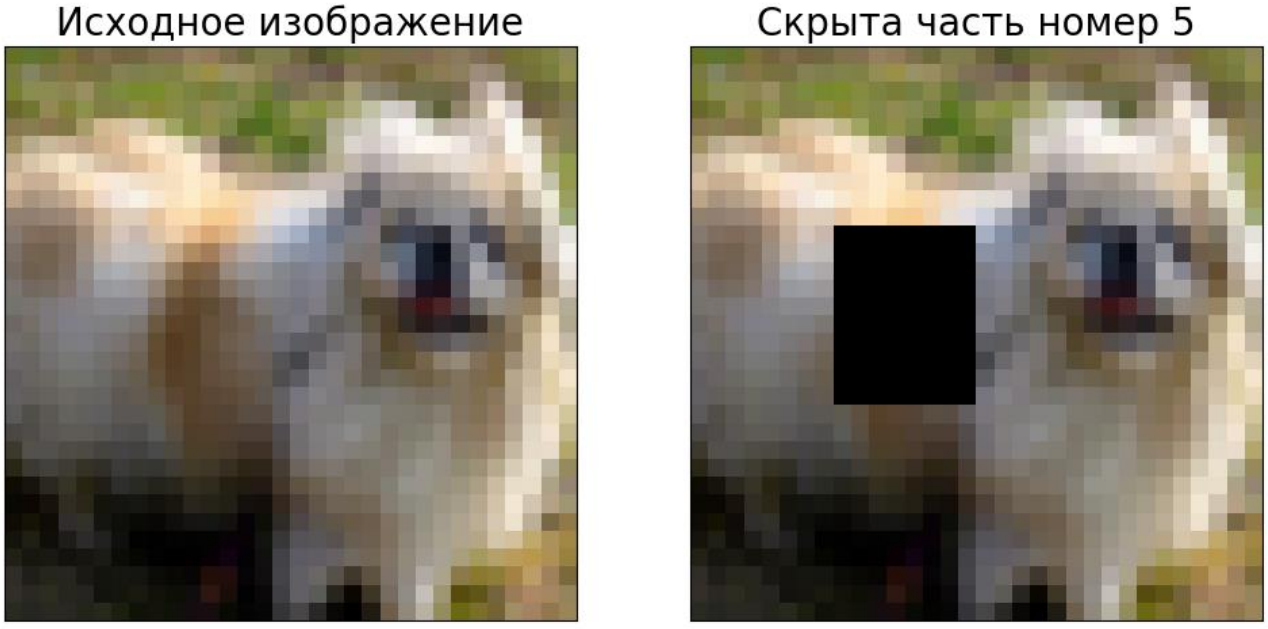

Рисунок 1. Пример маски типа 1 с разбиением на 10 прямоугольных частей. Картинка взята из датасета CIFAR-10

В эксперименте были взяты количество классов от 2 до 64 с шагом 4. Размер батча равен количеству классов. Таким образом ошибка усредняется по всем классам и градиентный спуск становится менее стохастическим.

В результате эксперимента видно, что точность нашей атаки равна почти единице при любом количестве классов, на которое разбивается маска. Это говорит о том, что искомое изображение, которое можно отнести к любому классу действительно находится. Однако количество эпох, затрачиваемое на поиск такого изображения, как правило, растет с увеличением количества классов. Флуктуации связаны со случайной инициализацией изображения.

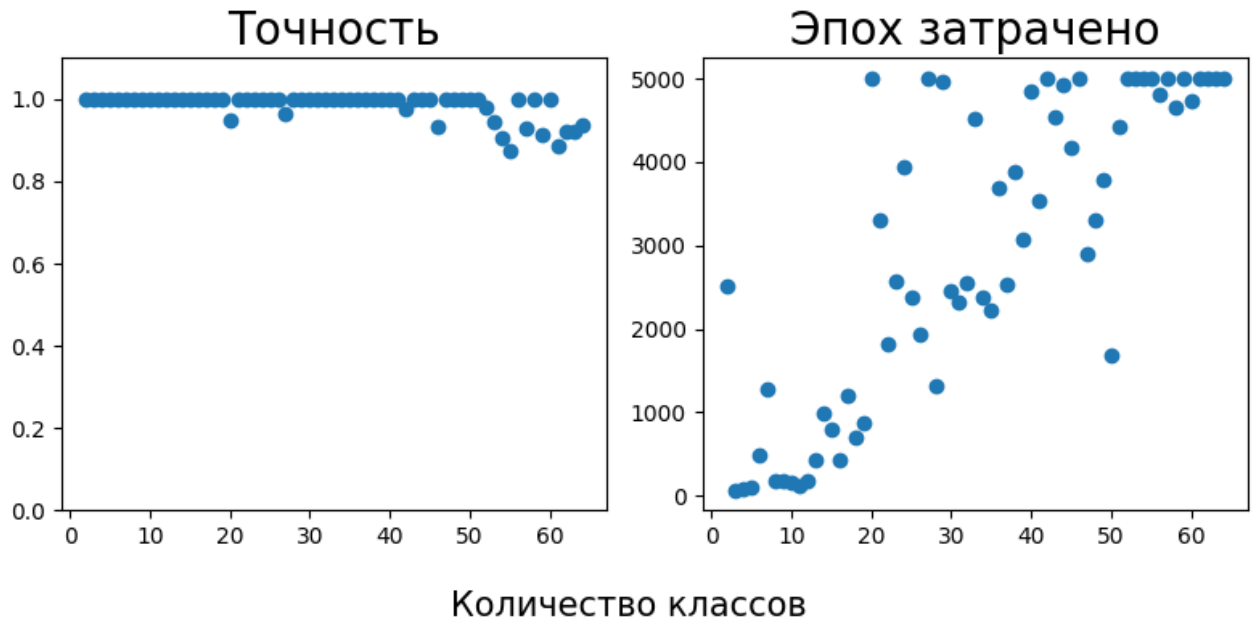

Рисунок 2. Зависимость точности и затраченных эпох при генерации изображения с маскированием типа 1

Однако стоит взглянуть на вероятности принадлежности к классу, которые выдает модель. На графике видно, что при любом количестве классов модель предсказывает заданный нами класс с хорошей вероятностью (более 70% в среднем). Это говорит об уверенности модели в своем предсказании.

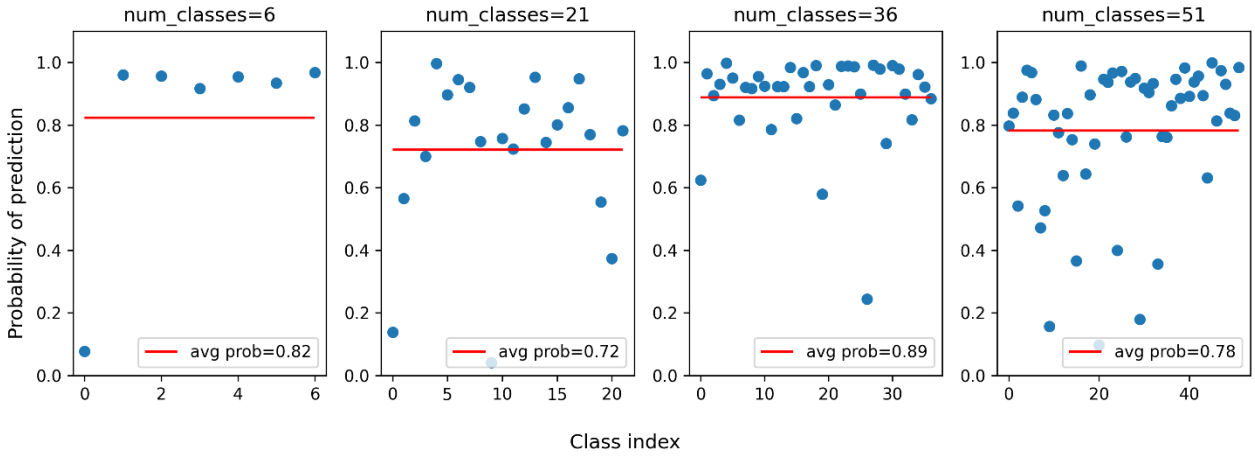

Рисунок 3. Зависимость предсказанной вероятности принадлежности изображения к заданному классу от номера класса при различном их количестве при применении маскирования типа 1.

6.2 Маскирование с разбиением маски на k полос

В данном эксперименте для каждого класса закрывается своя вертикальная полоска на изображении. Таким образом рецептивным полем для модели будут являться верхний и нижний куски изображения.

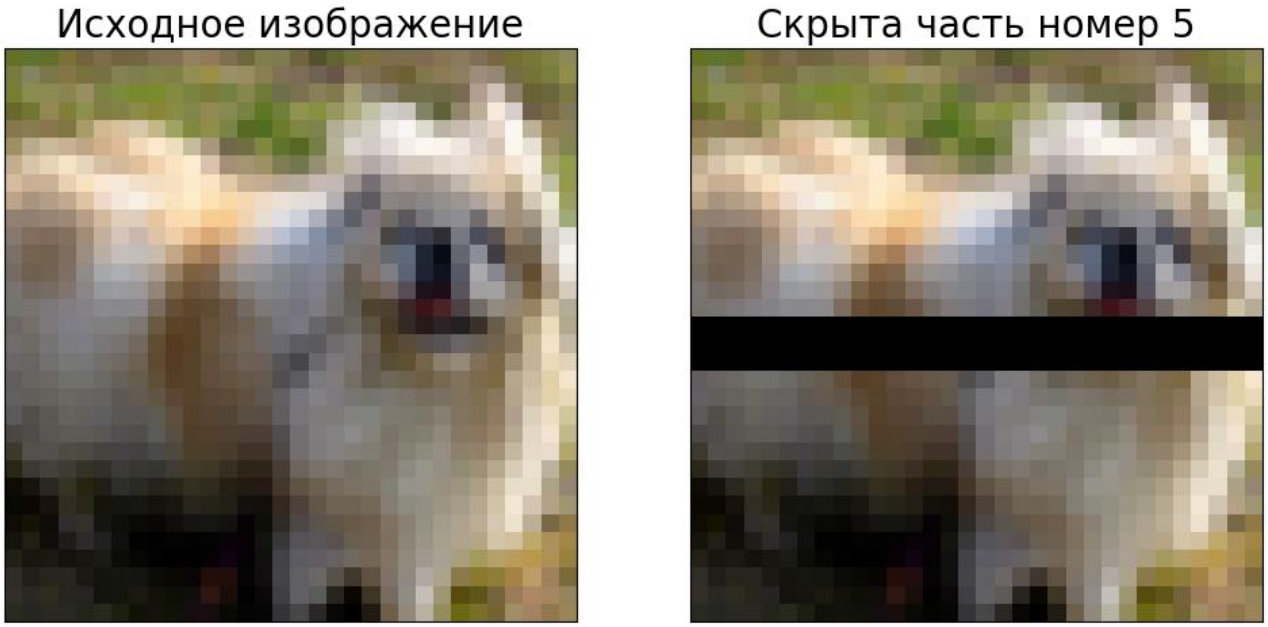

Рисунок 4. Пример маски типа 2 из горизонтальной линии с количеством классов равным 10. Картинка взята из датасета CIFAR-10

В результате эксперимента были получены схожие результаты, как и при использованы маски типа 1. Из-за того, что точность предсказаний равны 1, можно сделать вывод о том, что искомое изображение действительно находится,

причем с отличной точностью. Количество затрачиваемых эпох на поиск такого изображения возрастает с количеством классов, на которое мы разбиваем маску.

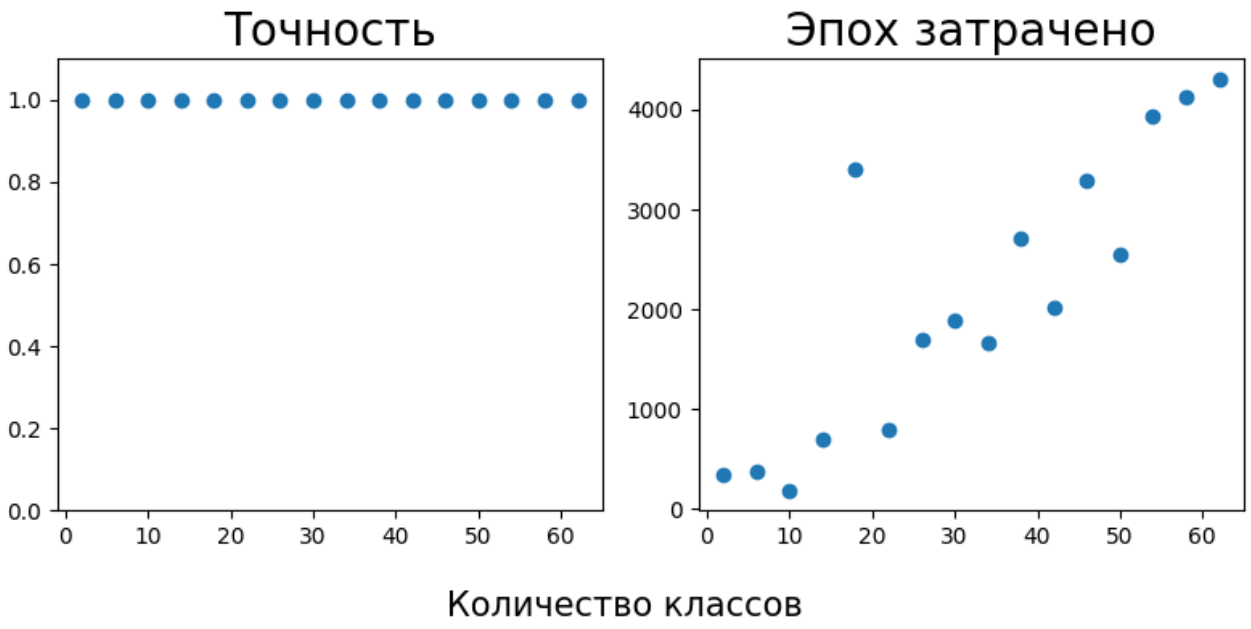

Рисунок 5. Зависимость точности и затраченных эпох при генерации изображения с маскированием типа 2.

Так же, как и в первом эксперименте, взглянем на вероятности, которые предсказывает модель, о принадлежности изображения к нужным нам классам.

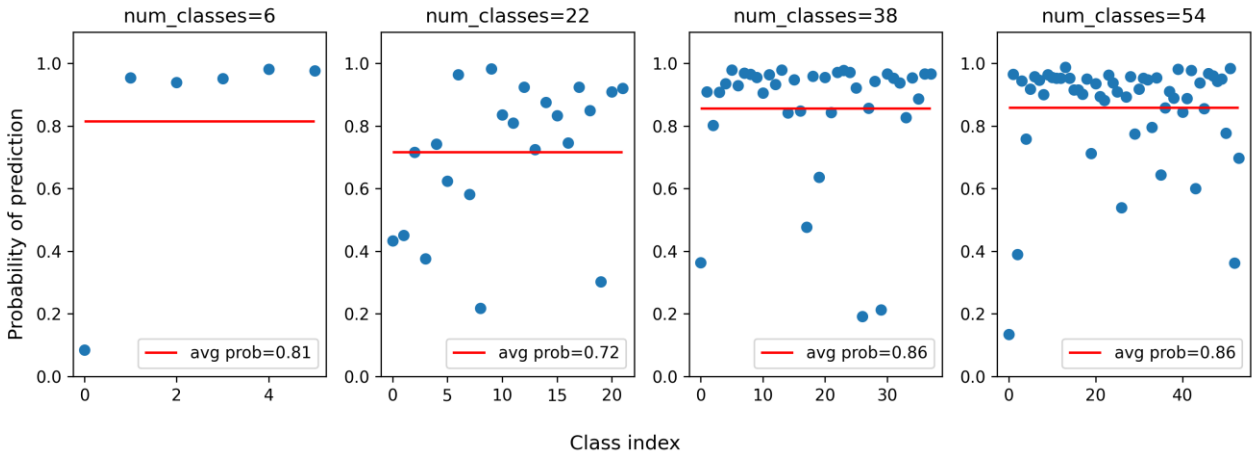

Рисунок 6. Зависимость предсказанной вероятности принадлежности изображения к заданному классу от номера класса при различном их количестве при применении маскирования типа 2.

Из графика видно, что модель предсказывает нужный нам класс при различном количестве разбиений (классов) изображения с довольно большой точностью, также как и в эксперименте 1, вероятность более 70% в среднем. Это говорит и достаточной уверенности модели в ответе.

6.3 Маскирование пикселями на случайных местах

В данном эксперименте для каждого класса закрываются некоторые случайные пиксели. Причем маска для каждого класса константная, и не зависит от итерации обучающего цикла. В данном случае рецептивное поле модели будет размазано по всей картинке, кроме некоторых случайных пикселей.

Рисунок 7. Пример маски типа 3 из случайных пикселей с количеством классов равным 10. Картинка взята из датасета CIFAR-10

В результате эксперимента были получены примерно схожие результаты, как и при применении масок типа 1 и 2.

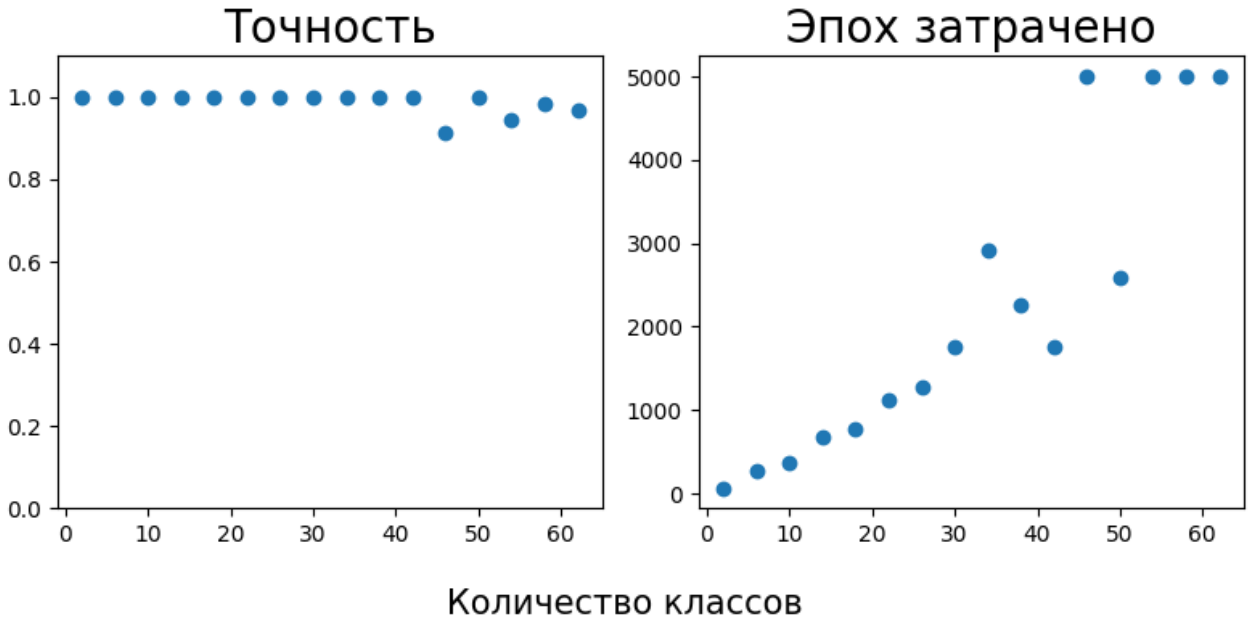

Рисунок 8. Зависимость точности и затраченных эпох при генерации изображения с маскированием типа 3.

Видно, что точность модели близка к единице при любом количестве различных масок, а также количество эпох, затрачиваемое на поиск такого изображения растет с увеличением количества классов.

Однако из графика вероятностей отнесения к классу, которые выдает модель, можно видеть, что в зависимости от количества классов, вероятность довольно сильно скачет. Это говорит о недостаточной уверенности модели в ответе.

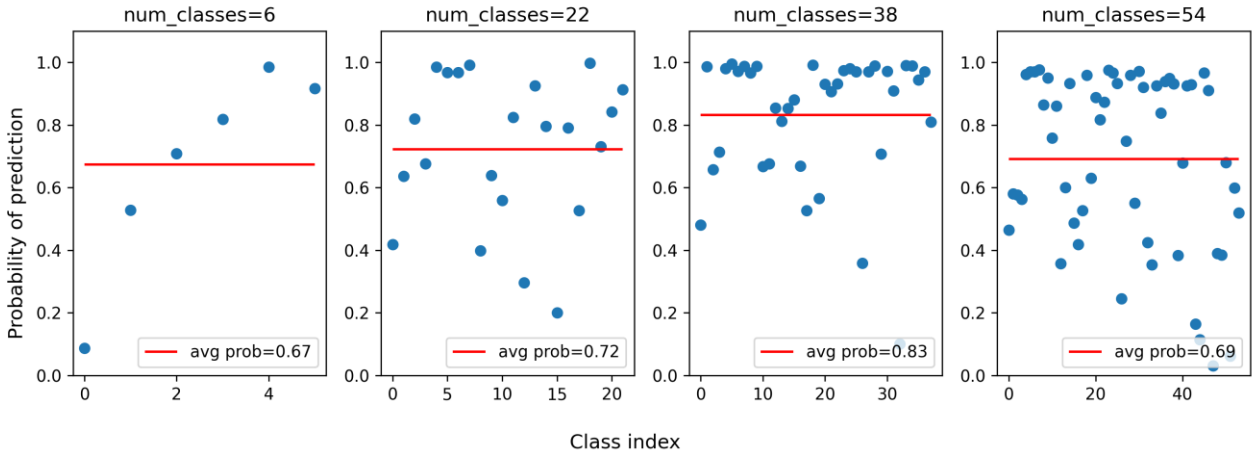

Рисунок 9. Зависимость предсказанной вероятности принадлежности изображения к заданному классу от номера класса при различном их количестве при применении маскирования типа 3.

7. Интерпретация результатов

В результате экспериментов были получены изображения, которые могут быть отнесены к любому классу из заданного набора. В первых двух экспериментах (маски типа 1 и 2) маскировались связные области на изображениях. Причем характерный размер таких областей был больше, чем размер свёрток, используемых в модели ResNet-50. Из-за этого часть сверточных слоев может ошибаться в ответах. Однако в третьем эксперименте маска представляла собой набор случайно расположенных черных пикселей, из-за чего область маскирования не была связной и не была локализована на изображении. В связи с чем сверточные слои, используемые в модели могли выдавать правильный ответ с большей вероятностью, ибо соседние от замаскированного случайного пикселя несли в себе информацию об изображении для того, чтобы его классифицировать. Это и объясняет хорошую точность и уверенность модели

в первых двух экспериментах и посредственную в третьем. Само изображения не несет какого-либо смысла для человеческого восприятия и имеет вид белого шума.

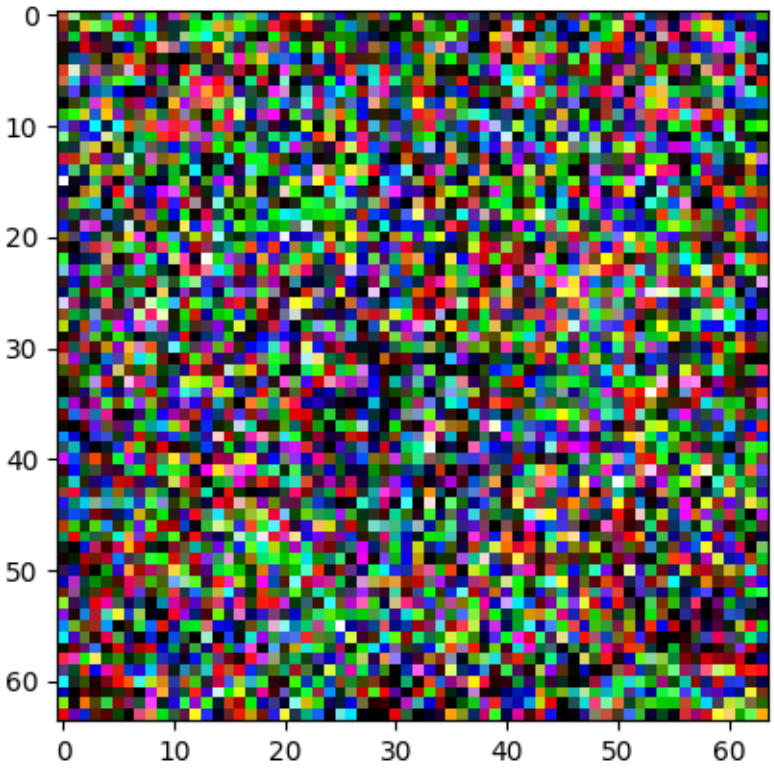

Рисунок 10. Пример сгенерированного изображения при применении маски типа 3 в случае 62 классов.

7.1 Изображение как вектор в линейном пространстве

Взглянем на задачу с теоретической точки зрения. Любое RGB-изображение можно представить как вектор в ограниченной области линейного пространства, если принять значение пикселей как действительное число от 0 до 1, причем размерность его будет, в нашем случае, равна $3 \times 64 \times 64 = 12\,288$. Когда мы применяем маскирование, мы берем вектор из этого пространства, который по норме немного отличается от исходного (т.е. универсального изображения). Например, в случае маски горизонтальными линиями при 64 классах маскированное изображение отличается по норме не более, чем на 1.6%. Получается, что объекты из довольно маленького ограниченного шара из множества изображений могут переводиться моделью во все возможные классы.

Причем вектором, которым мы отступаем от центра этого шара для получения изображения другого класса, является маска.

8. Основные итоги работы

В результате работы были исследованы различные типы маскирований для получения изображения, которое могло быть отнесено нейронной сетью к любому классу из набора заданных. Причем для каждого из рассмотренных типов масок такое изображение было найдено. Также, было показано, что ответы модели при применении маскирования типа 1 и 2 были достаточно уверенны. Что говорит о применимости adversarial attack для такого типа модели.

Также, был сделан вывод о том, что в довольно узкой окрестности такого сгенерированного изображения содержатся изображения, относящиеся нейронной сетью к любому классу из заданного набора. Такое заключение может помочь в теоретических исследованиях о том, как именно работают нейронные сети.

В дальнейшем, мы планируем провести эксперименты на большем числе масок и с использованием других моделей нейронных сетей. Однако такие эксперименты требуют большого количества времени ввиду высокой вычислительной сложности моделей из-за огромного числа параметров.

9. Литература

1. Sen, Jaydip. (2023). Adversarial Attacks on Image Classification Models. 10.13140/RG.2.2.19431.01449.
2. Khamaiseh, Samer & Bagagem, Derek & Al-Alaj, Abdullah & Mancino, Mathew & Alomari, Hakam. (2022). Adversarial Deep Learning: A Survey on Adversarial Attacks and Defense Mechanisms on Image Classification. IEEE Access. PP. 1-1. 10.1109/ACCESS.2022.3208131.
3. Xu, Han & Ma, Yao & Liu, Haochen & Deb, Debayan & Liu, Hui & Tang, Ji-Liang & Jain, Anil. (2020). Adversarial Attacks and Defenses in Images, Graphs and Text: A Review. International Journal of Automation and Computing. 17. 10.1007/s11633-019-1211-x.
4. Akhtar, Naveed & Mian, Ajmal & Kardan, Navid & Shah, Mubarak. (2021). Advances in Adversarial Attacks and Defenses in Computer Vision: A Survey. IEEE Access. PP. 1-1. 10.1109/ACCESS.2021.3127960.
5. Akhtar, Naveed & Mian, Ajmal. (2018). Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey.