Московский
государственный
университет
имени М.В. Ломоносова

ГЕНЕРАЦИЯ ИЗОБРАЖЕНИЙ ДЛЯ АНАЛИЗА УСТОЙЧИВОСТИ НЕЙРОННЫХ СЕТЕЙ В ЗАДАЧЕ КЛАССИФИКАЦИИ

Студент 435 группы | Клиентов Г.А.

Научные руководители:
д. ф.-м. н., профессор | Голубцов П.В.
д. м.-м. н., профессор РАН | Дьяконов А.Г.

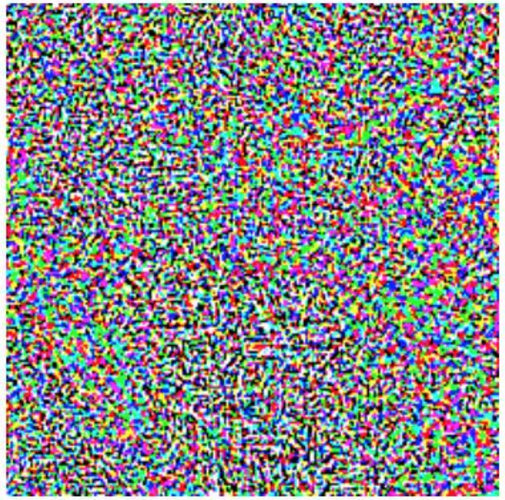
28.05.2025

Adversarial attack на примере FGSM

$\boldsymbol{x}$

“panda”
57.7% confidence

$+ .007 \times$



$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

“nematode”
8.2% confidence

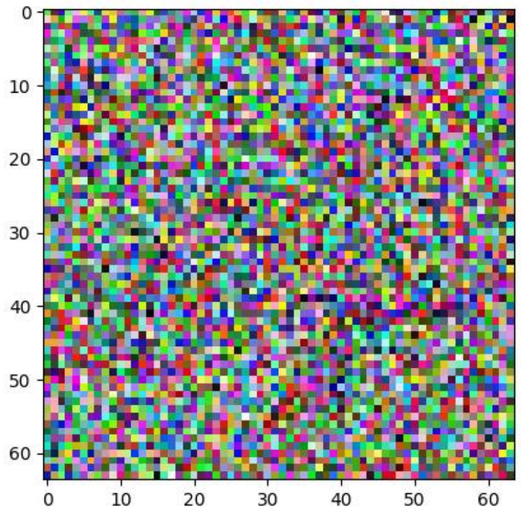
$=$

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

“gibbon”
99.3 % confidence

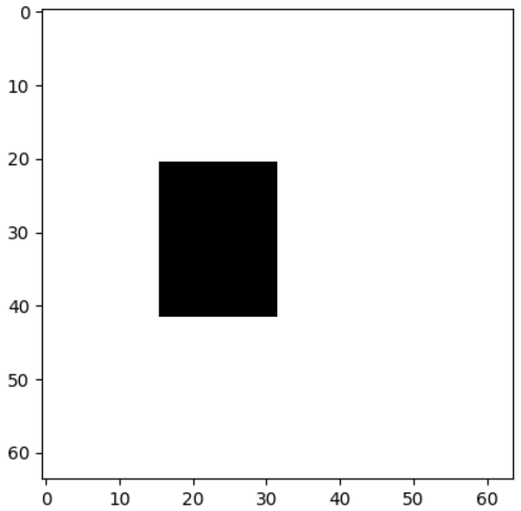
Adversarial Example Generation — PyTorch Tutorials https://pytorch.org/tutorials/beginner/fgsm_tutorial.html2.3.0+cu121 documentation

Можем ли мы подобрать изображение, которое бы относилось к любому классу?



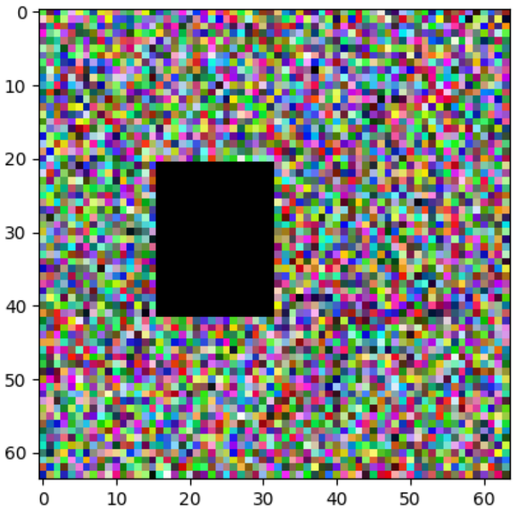
Универсальное изображение

×



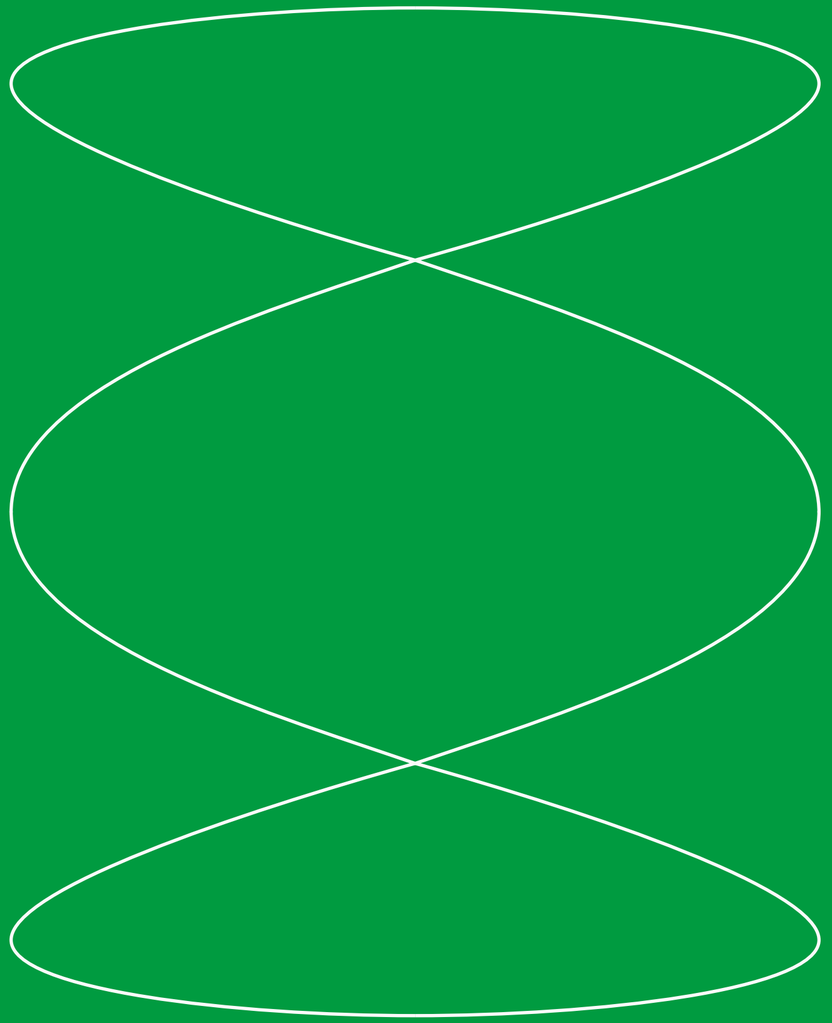
Маска для класса “gibbon”

=



“gibbon”

Способ генерации изображения



- Используем метод градиентного спуска
- Фиксируем веса модели
- В качестве параметров для оптимизации используются значения пикселей самого изображения
- Батч собираем из всего набора масок (размер батча равен количеству классов)

Используемая модель

ResNet-50 с 25.6 млн параметров (80% acc@1)

Предобученная на ImageNet-1K

Размер искомого изображения – 3x64x64 (RGB)

Количество классов – от 2 до 62 с шагом 4

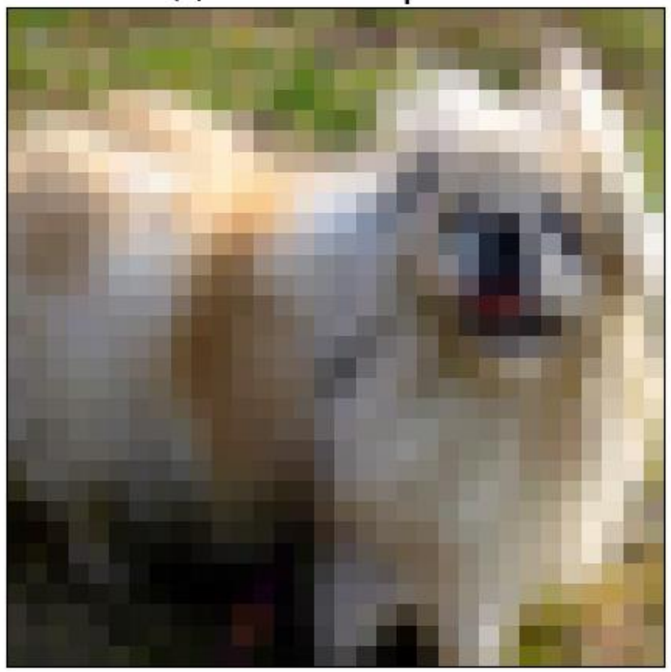
Количество эпох – до 5000

Loss-функция – CrossEntropyLoss

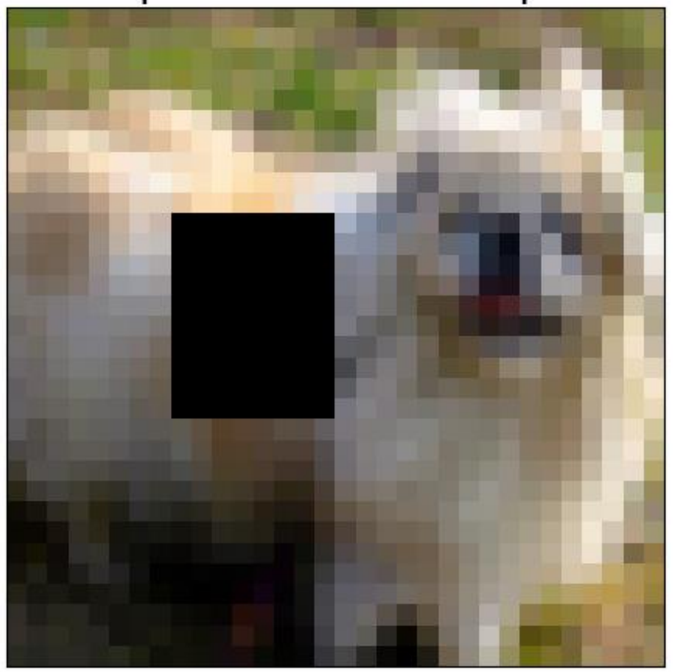
Оптимизатор – SGD с learning rate=0.1

Эксперимент 1
(маска, разбитая на k равных частей)

Исходное изображение

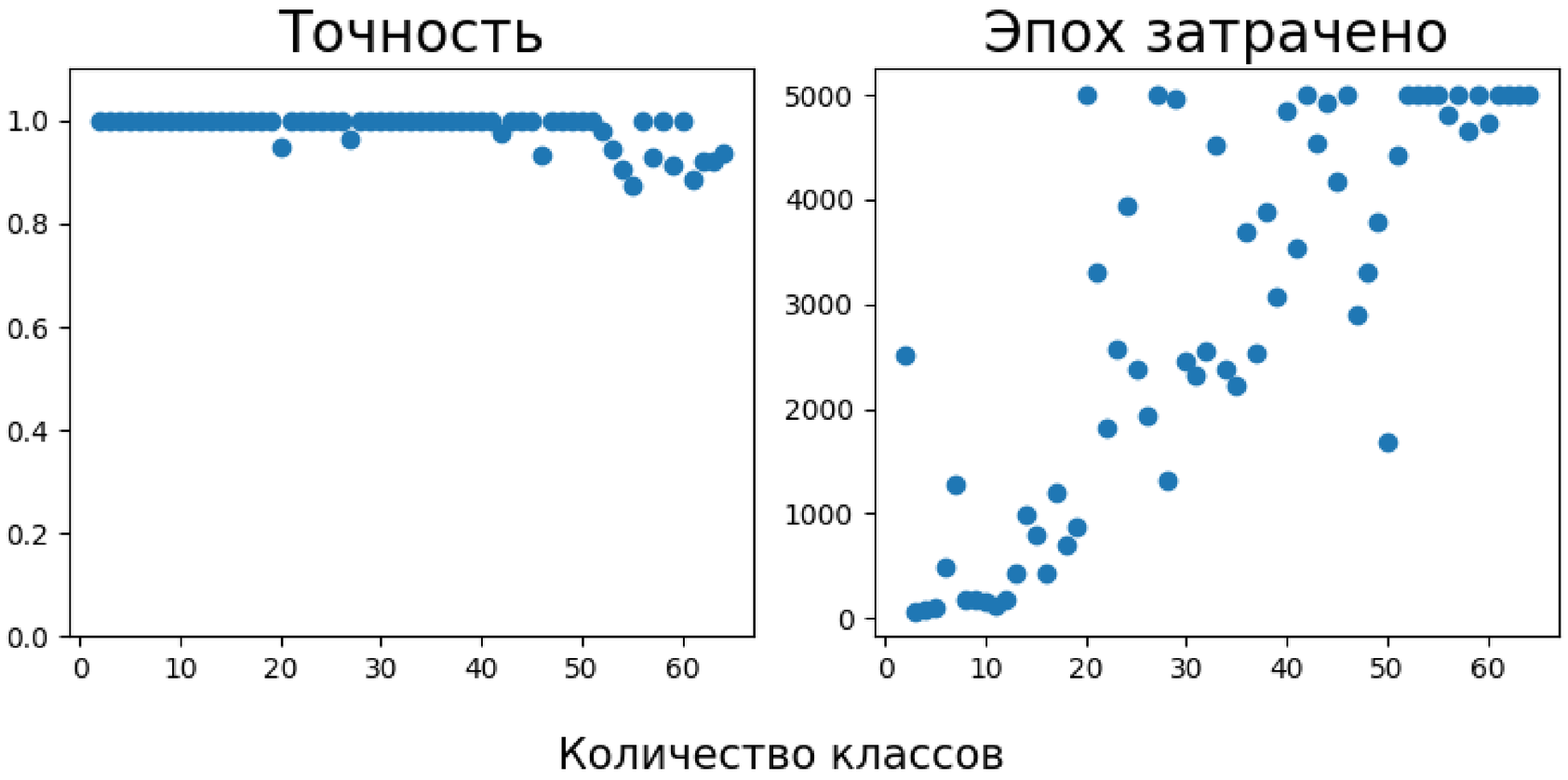


Скрыта часть номер 5

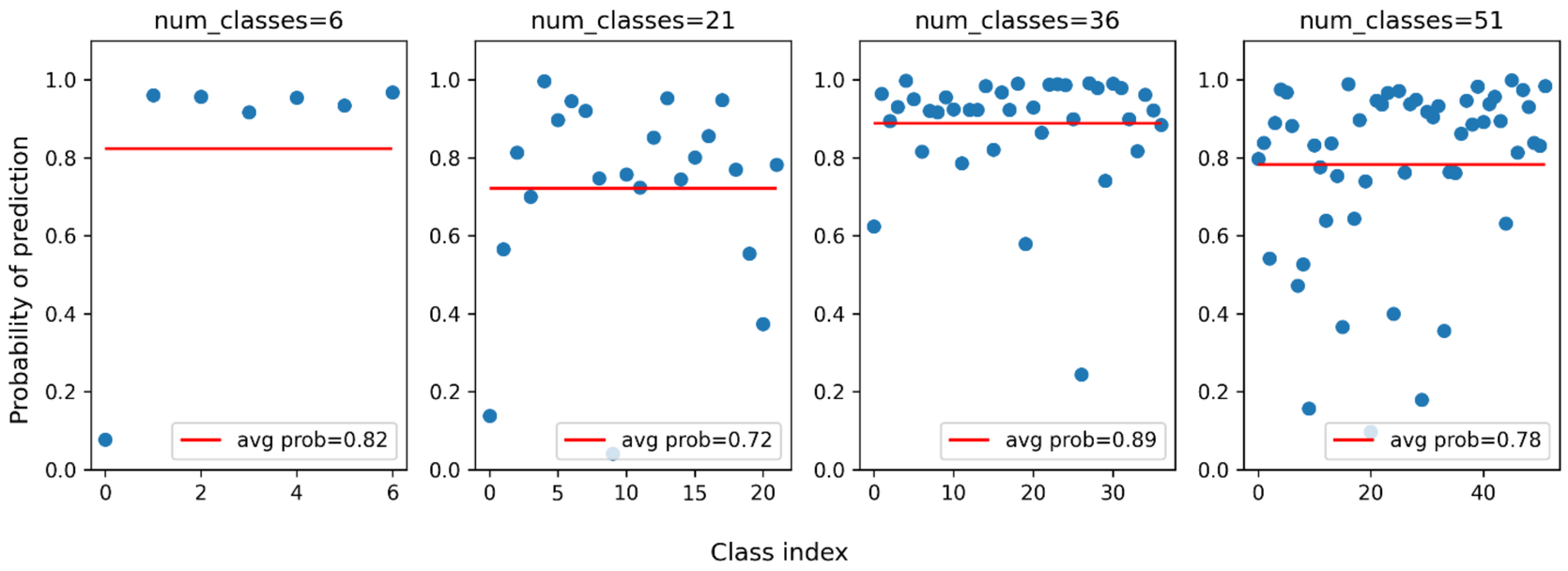


Уменьшаем рецептивное поле модели, из-за чего она классифицирует изображение только по немаскированной части

Точность модели оказалась около единицы



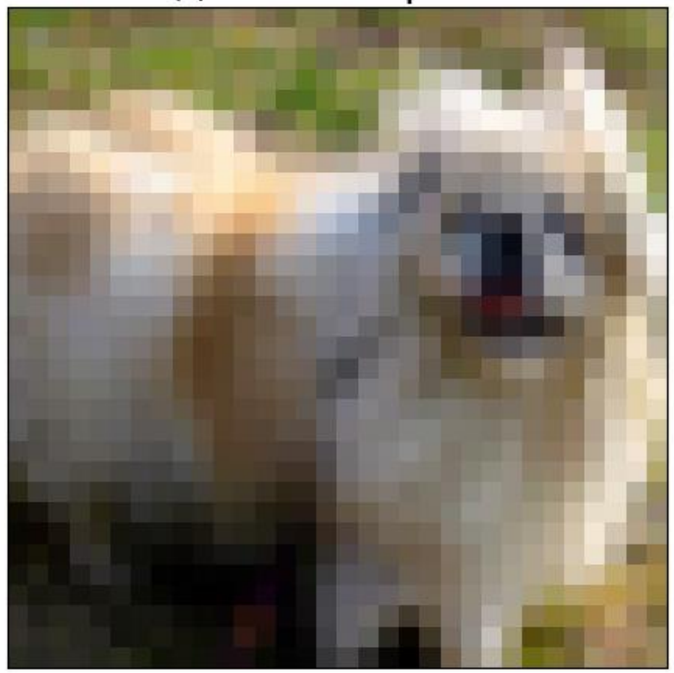
Вероятности, выдаваемые моделью



Высокие вероятности принадлежности к классу, которые выдает модель говорит о её уверенности в ответах

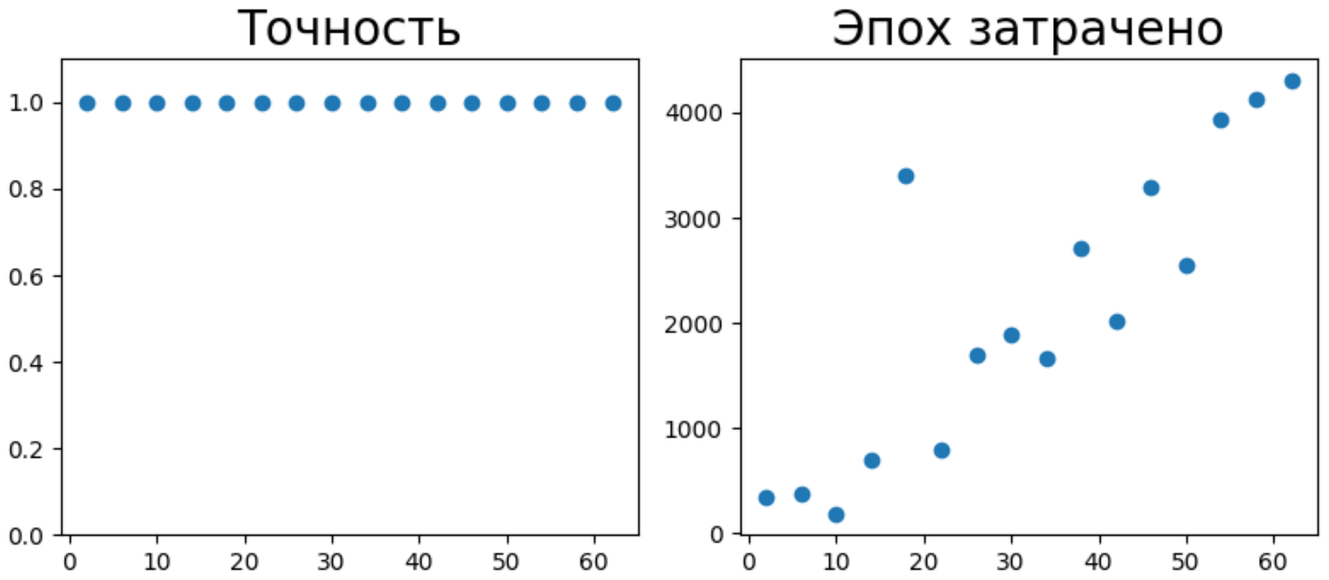
Эксперимент 2
(маска, разбитая на k равных полос)

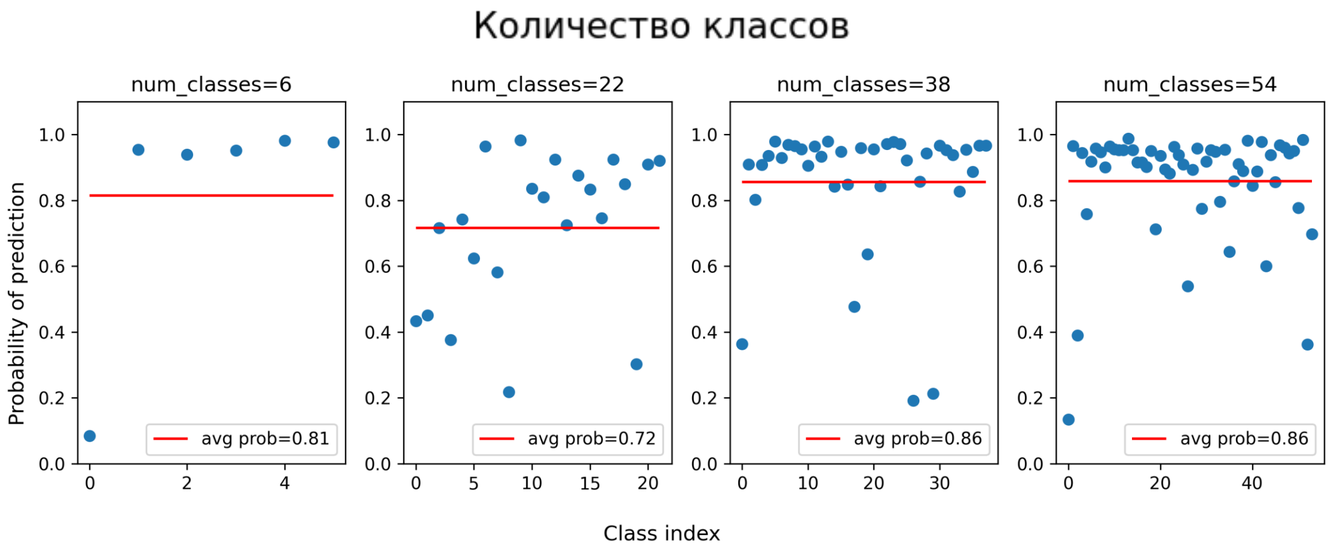
Исходное изображение



Скрыта часть номер 5

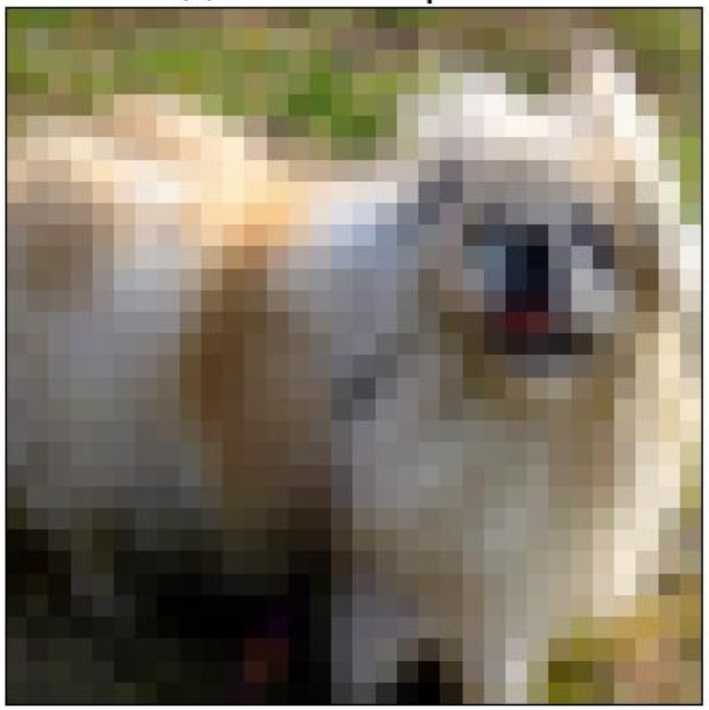
Результаты сходны с экспериментом №1



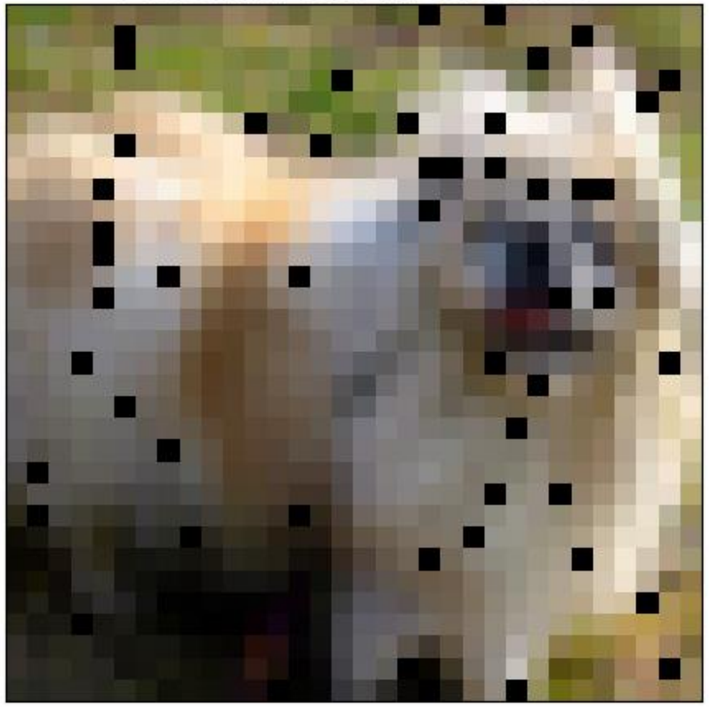


Эксперимент 3
(маска из случайно расположенных пикселей)

Исходное изображение

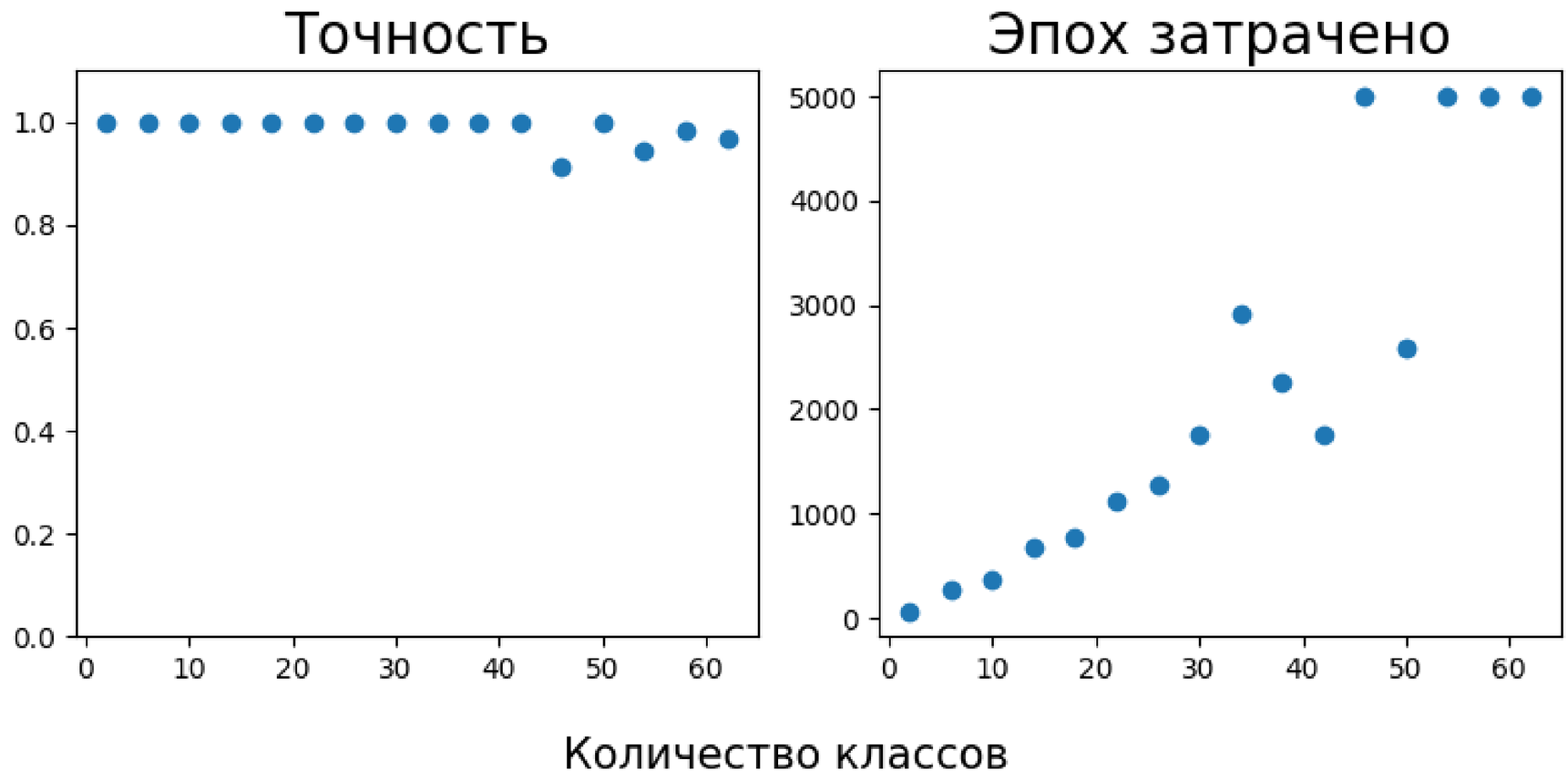


Маска для класса 5



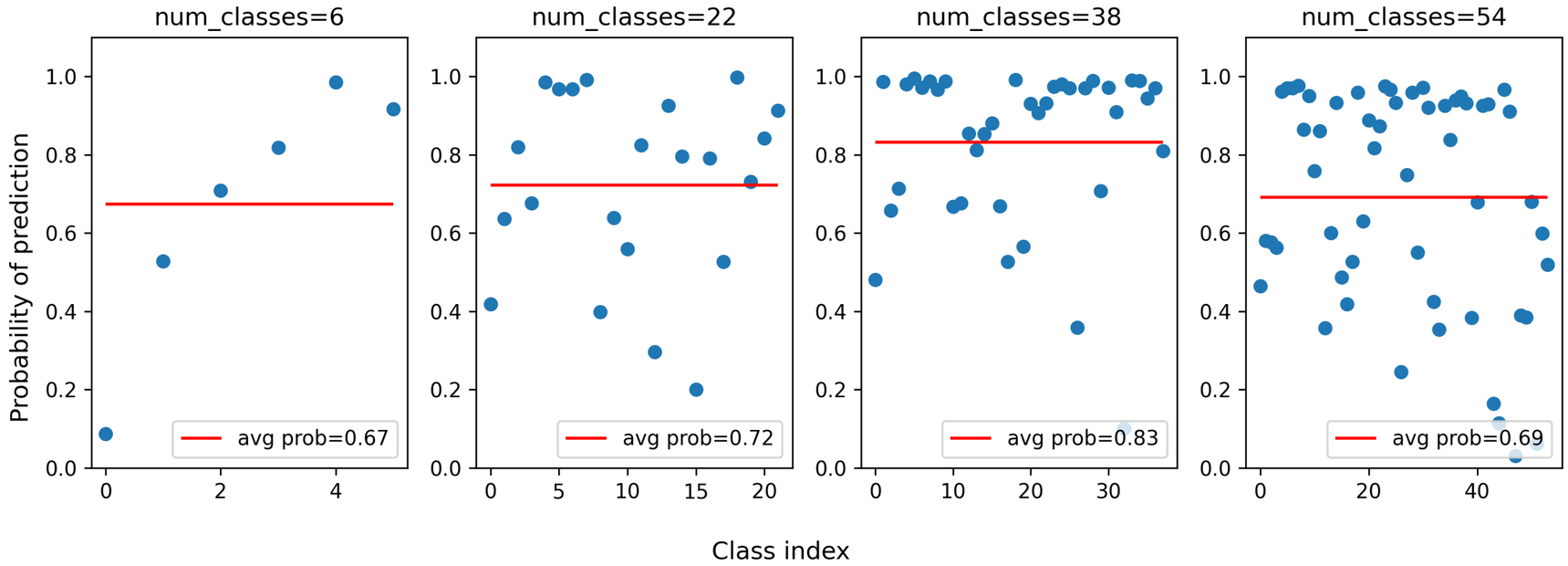
В данном случае рецептивное поле размазано по всему изображению, кроме отдельно стоящих пикселей.

Точность модели также около единицы



Вероятности, выдаваемые в эксперименте №3

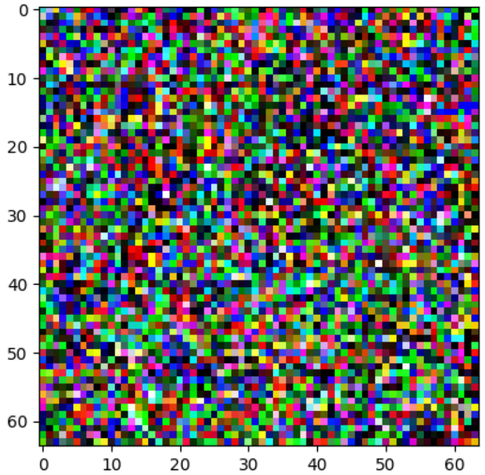
В данном случае, несмотря на высокую точность предсказаний, модель недостаточно уверена в своих ответах по поводу принадлежности изображения к своему классу.



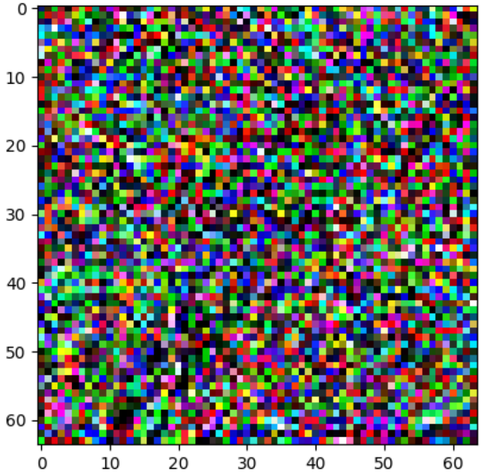
Примеры генерируемого изображения

Изображение не несет какого-либо смысла для человеческого восприятия и имеет вид белого шума.
Изображения получены в экспериментах с количеством классов равным 62.

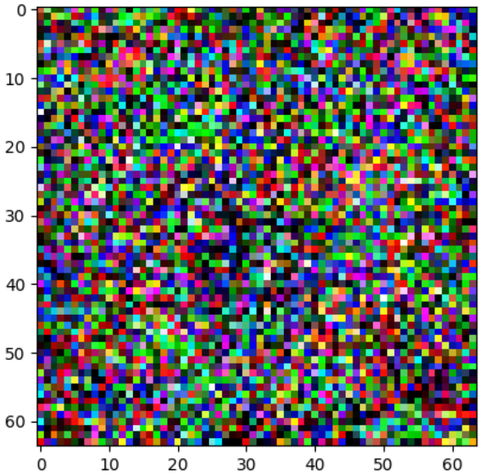
Эксперимент 1 (маска из равных квадратов)



Эксперимент 2 (маска из горизонтальных полос)



Эксперимент 3 (маска из случайных пикселей)



Интерпретация результатов

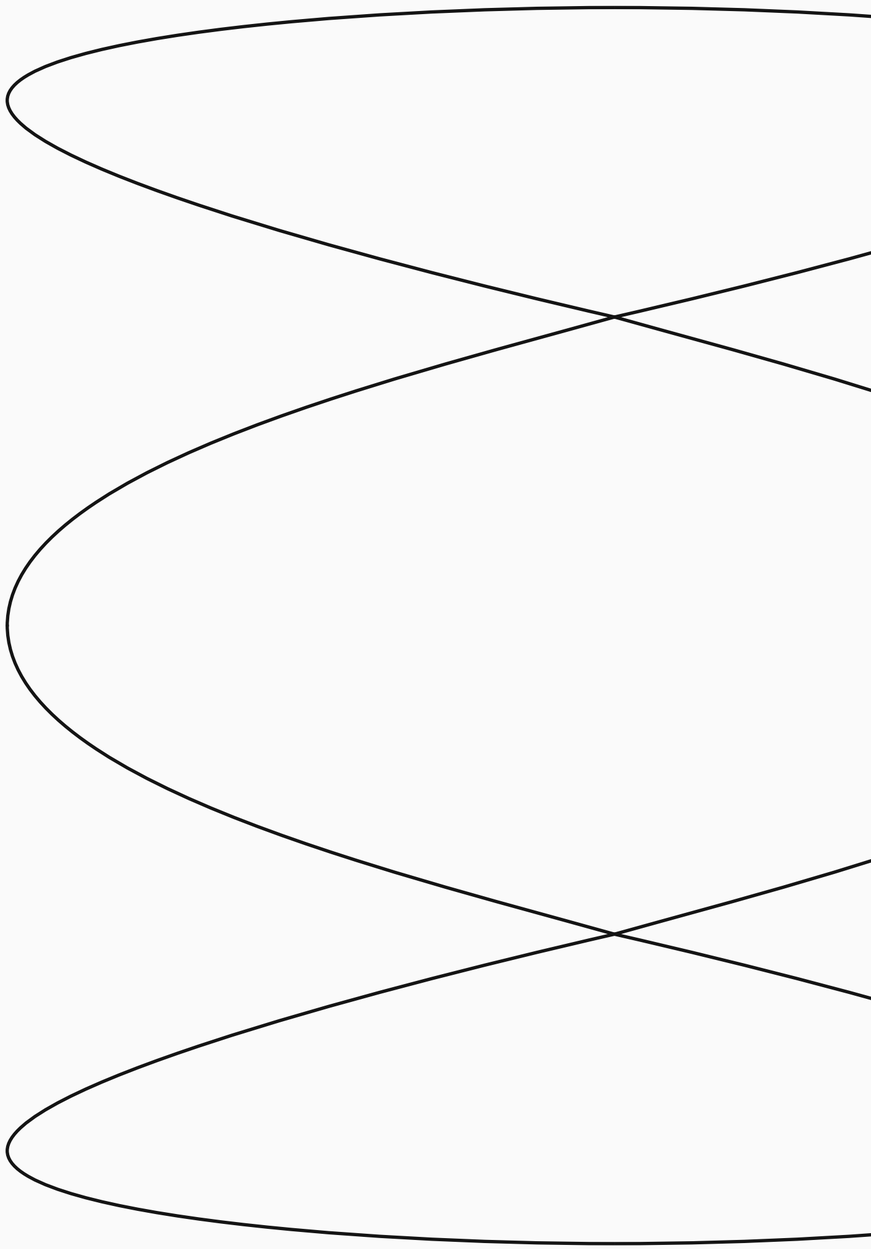
Видно, что во всех случаях универсальное изображение находится

Причем в первых двух экспериментах модель с достаточной уверенностью выдает ответы.

Размер свертки меньше характерного размера маски

Однако в третьем эксперименте несмотря на высокую точность предсказаний, вероятности ответов модели низкие.

Размер свертки больше характерного размера маски

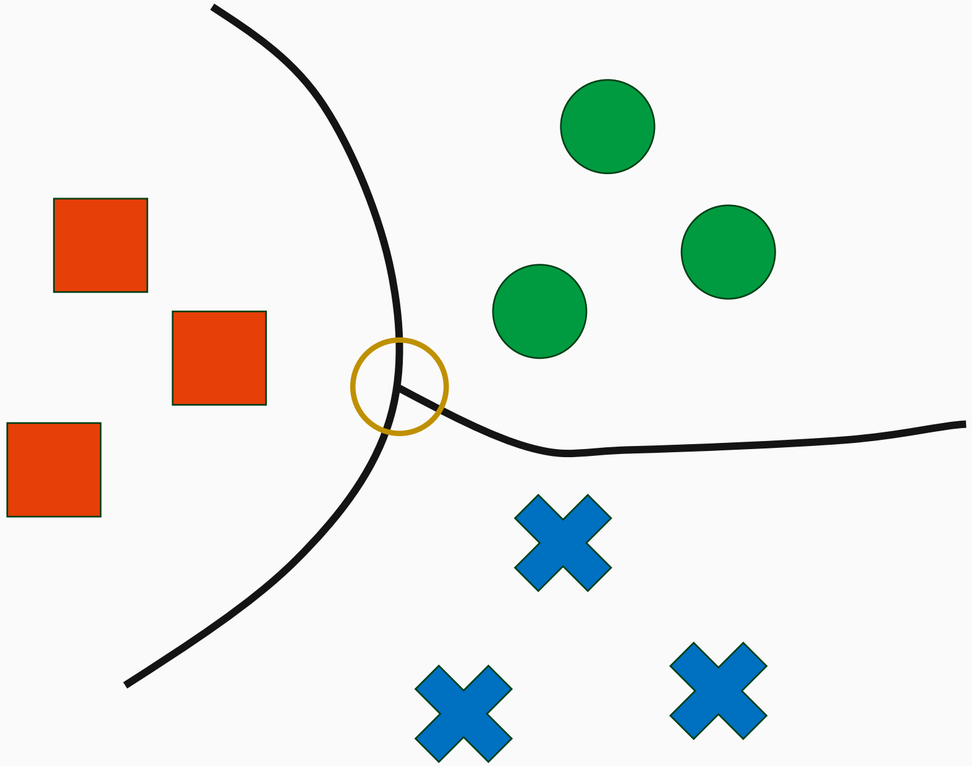


Теоретическая интерпретация

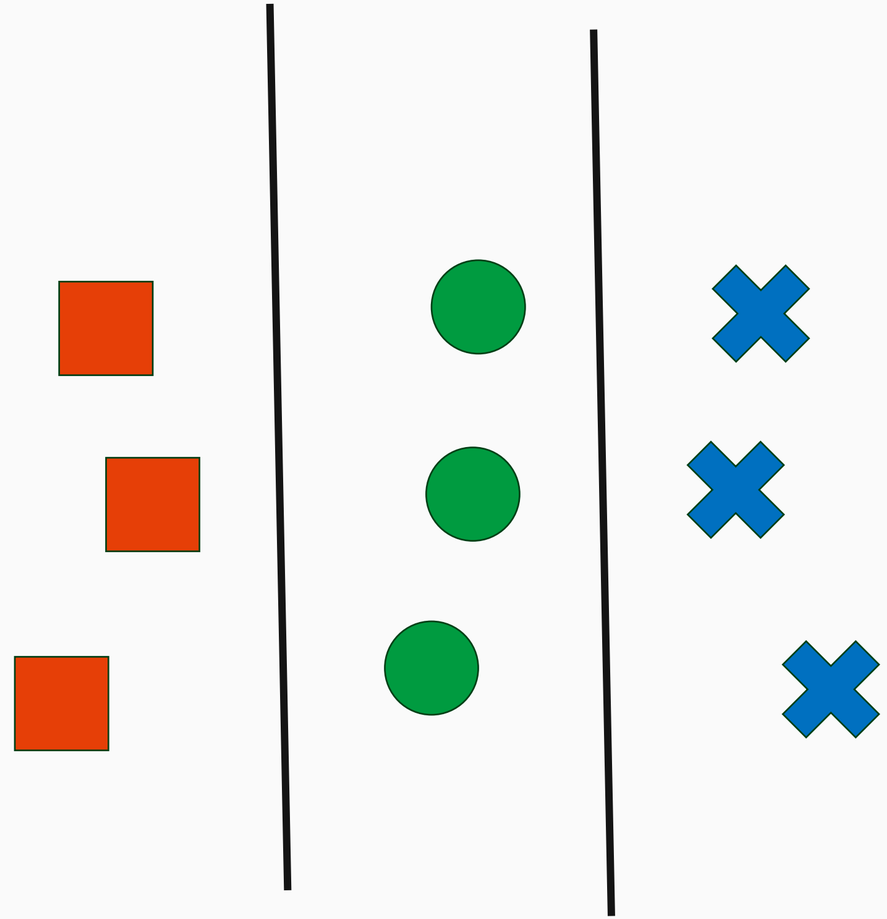
- Представим изображение как объект в линейном пространстве (все множество изображений – гиперкуб со стороной 1)
- При применении масок мы получаем объекты из некоторой окрестности универсального изображения
- Внутри этой окрестности содержатся объекты, которые модель переводит в любой класс из заданного набора

Пример – при применении горизонтальной маски с количеством классов = 64, относительная разность норм для маскированных изображений не более 1.6%. И в этом шаре содержатся объекты всех 64 классов.

Иллюстрация



В данном случае разделяющие гиперплоскости пересекаются в некоторой точке – в её окрестности есть любой класс



Здесь разделяющие гиперплоскости НЕ пересекаются – нет окрестности, в которой есть объекты любого класса

Основные итоги работы

Разработан алгоритм для генерации универсальных изображений, которые при применении различных масок имеют разные классы на выходе модели

Получены и исследованы изображения, полученные этим алгоритмом.

Такие исследования могут помочь в теоретических исследованиях работы нейронных сетей

В дальнейшем планируется исследовать данную задачу на бóльшем числе масок и моделей.

Вопросы