

# Постановка задачи

Разработать агент на основе обучения с подкреплением (RL), который посредством активного управления фокусом внимания (позиция + масштаб окна) оптимально анализирует изображения для решения downstream-задачи (например, классификации, обнаружения объектов или распознавания сцен).

Преимущества подхода:

Близость к биологическому зрению;

Интерпретируемость: известно, куда и когда агент смотрел.

Эффективность: работает с малой долей полного изображения, экономя ресурсы.

Масштабируемость: легко адаптируется к разным разрешениям изображений.

## Previous works

### **Mnih et al., Recurrent Models of Visual Attention, NeurIPS 2014**

- RNN-агент последовательно выбирает координаты окна внимания и обрабатывает его патчи
- Обучение: REINFORCE + supervised loss
- Успешно на MNIST и простых задачах

### **Wu et al., Glance and Focus Networks for Dynamic Visual Recognition, 2022**

- Два этапа: «glance» – обзор всего изображения в низком разрешении; «focus» – детальный анализ выбранных регионов
- RL-политика решает, когда переключаться с coarse на fine

### **Zhu et al., Active Vision Reinforcement Learning under Limited Visual Observability, NeurIPS 2023**

- Агент видит только ограниченные фрагменты сцены
  - Actor–critic методы (PPO/A2C) планируют серию «сдвигов взгляда» для сбора информации

### **\* Zhu et al., Saccade Mechanisms for Image Classification, Object Detection and Tracking, arXiv 2022**

- Имитируют биологические саккады – быстрые, эпизодические перемещения взгляда
  - Стратегии выбора: численные (минимизация неопределённости) и эвристические (градиентные карты)

### **D’Ascoli et al., Adaptive patch selection to improve Vision Transformers through Reinforcement Learning, Knowledge-Based Systems 2025**

- RL-агент последовательно выбирает наиболее информативные патчи для подачи в Vision Transformer

- Снижает вычислительные затраты при сохранении (или улучшении) точности

## Метод

В данной работе применяется алгоритм Proximal Policy Optimization (PPO) для обучения агента, который решает задачу классификации изображений из набора данных CIFAR-10. Агент исследует различные участки изображений с использованием изменения масштаба и выполняет классификацию на основе локальных патчей изображений.

Среда обучения была реализована с использованием библиотеки Gym, а модель агента включает рекуррентную сеть актера-критика, которая обучается с применением метода PPO с улучшенной стабильностью за счет использования обрезанной целевой функции.

## Среда

Среда взаимодействия агента с данным набором данных является кастомизированной реализацией в рамках OpenAI Gym. Агент взаимодействует с изображениями из набора данных CIFAR-10 (60,000 цветных изображений размером 32x32 пикселя, разделённых на 10 классов). Для задач обучения агент получает наблюдения в виде локальных патчей изображений, определяемых случайным образом. Среда позволяет агенту:

- Перемещаться по изображению с возможностью изменения его области просмотра.
- Изменять уровень зума изображения для получения более детализированных или более общих представлений.
- Классифицировать изображение в одном из 10 классов.

Используется техника  $\epsilon$ -greedy, где с вероятностью агент выбирает случайное действие.

## Модель агента

Модель агента реализует рекуррентную нейронную сеть актер-критика:

- Сверточные слои для извлечения признаков из изображений.
- Рекуррентная сеть LSTM для обработки последовательных данных и поддержания скрытого состояния.
- Выходная сеть - политики, которая предсказывает вероятности действий агента (движения, приближение/удаление, классификация).
- MLP-ценности, которая предсказывает ожидаемое вознаграждение для каждого состояния.

## Алгоритм PPO

Алгоритм PPO используется для оптимизации политики агента с учетом ограничений на величину изменений в политике. Это достигается через введение обрезанной замещающей функции, которая ограничивает величину обновлений и способствует более

стабильному обучению. Целевая функция PPO для обновления параметров политики  $\theta$  вычисляется как:

$$L(\theta) = \mathbb{E}_t \left[ \min \left( r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t \right) \right]$$

$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$  - это отношение вероятностей действий, выполненных по старой и новой политике, а  $\hat{A}_t$  - оценка преимущества для текущего шага  $t$ .

### Обновление модели

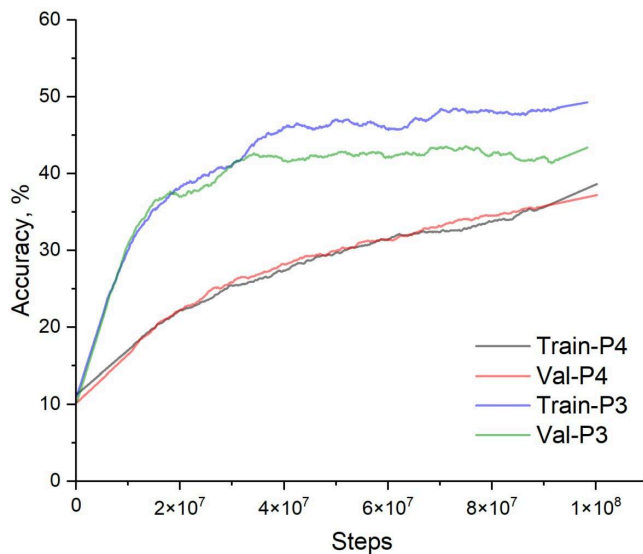
Для вычисления преимущества  $\hat{A}_t$  используется метод Generalized Advantage Estimation (GAE). Целевая функция, включающая потерю по политике, потерю по ценности и регуляризацию энтропии для поощрения исследования, формулируется как:

$$L_{\text{total}} = L_{\text{policy}} + L_{\text{value}} - \beta H[\pi(\cdot|s_t)]$$

где  $L_{\text{policy}}$  — потеря по политике,  $L_{\text{value}}$  — потеря по ценности,

$H[\pi(\cdot|s_t)]$  — энтропия политики для предотвращения преждевременной сходимости к детерминированному решению, а  $\beta$  — коэффициент энтропии.

### Заключение



Итоговая модель показывает способность эффективно классифицировать изображения из набора CIFAR-10 с помощью сложных операций перемещения и зумирования изображения с точностью от 38 до 54% при сбалансированных гиперпараметрах. Предложенный метод позволяет агенту эффективно исследовать изображения с помощью обученной политики PPO и рекуррентной сети актера-критика. Он также демонстрирует, как алгоритм PPO может быть применен в контексте задачи классификации изображений, где агент не только принимает решения о классификации, но и активно исследует различные области изображения для

улучшения своей классификационной способности. Реализация, отражающая вариацию среды, может быть просмотрена в файлах кода PPO Greedy Map и PPO Multi Categorical.

## PPO + ViT

### Архитектура агента

Агент основан на Vision Transformer (ViT) с гибридным пространством действий. Модель состоит из патч-эмбединга, трансформера и специализированных головок для различных типов действий. Патч-эмбединг преобразует входные изображения в токены, а трансформер обрабатывает их с помощью механизма внимания. Специализированные головки генерируют различные компоненты действия: решение о классификации, предсказание класса, координаты движения и уровень масштабирования.

### Гибридное пространство действий

Пространство действий представляет собой словарь с четырьмя компонентами. Компонент decision определяет, будет ли агент продолжать исследование изображения или сделает классификацию. При решении продолжить исследование используются компоненты move и zoom для изменения позиции обзора и масштаба соответственно. Компонент class содержит предсказание одного из десяти классов CIFAR-10. Все непрерывные значения нормализованы в диапазоне [0, 1] для стабильности обучения.

### Окружение и процесс взаимодействия

Агент получает патч фиксированного размера (8×8 пикселей) из текущей позиции обзора. На каждом шаге агент может либо продолжить исследование, изменяя позицию и масштаб обзора, либо сделать классификацию текущего изображения. При исследовании агент выбирает новую позицию в нормализованных координатах и уровень масштабирования, что позволяет ему фокусироваться на различных частях изображения. Система масштабирования позволяет агенту изменять размер области обзора от минимального до максимального зума.

### Процесс принятия решений и обработка наблюдений

Агент получает патч изображения вместе с координатами центра и информацией о масштабе. Эти данные проходят через патч-эмбединг, который преобразует изображение

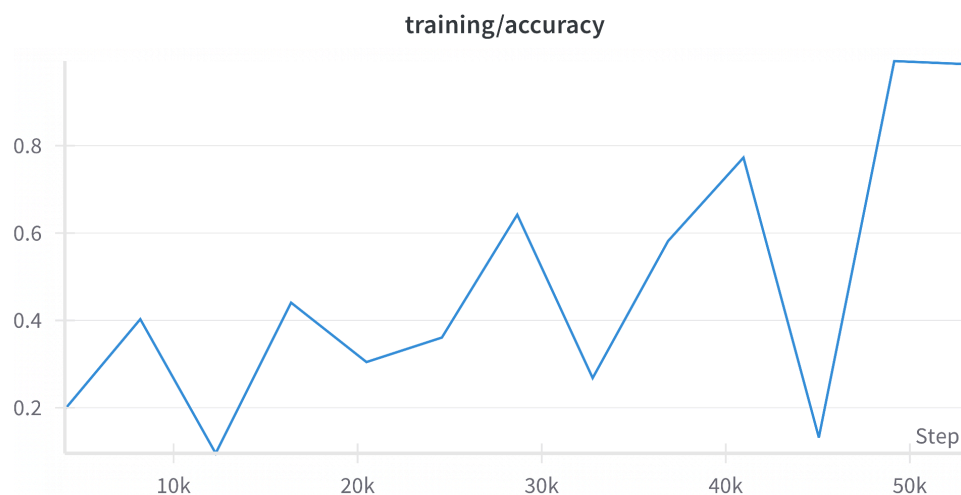
в токены и добавляет позиционное кодирование координат с помощью фурье-эмбеддингов. Трансформер обрабатывает эти токены, а специализированные головки генерируют все компоненты действия одновременно. В зависимости от предсказанного решения используются соответствующие компоненты действия: при решении продолжить исследование применяются координаты движения и масштаб, при решении классифицировать используется предсказание класса.

## Recurrent Memory Transformer и обработка истории

Для учета истории взаимодействий используется Recurrent Memory Transformer, который обрабатывает последовательность патчей и их координат. Это позволяет агенту учитывать предыдущие наблюдения при принятии решений. История патчей и их координат сохраняется и передается в модель – это обеспечивает контекстную информацию для принятия решений.

## Гибридное распределение действий

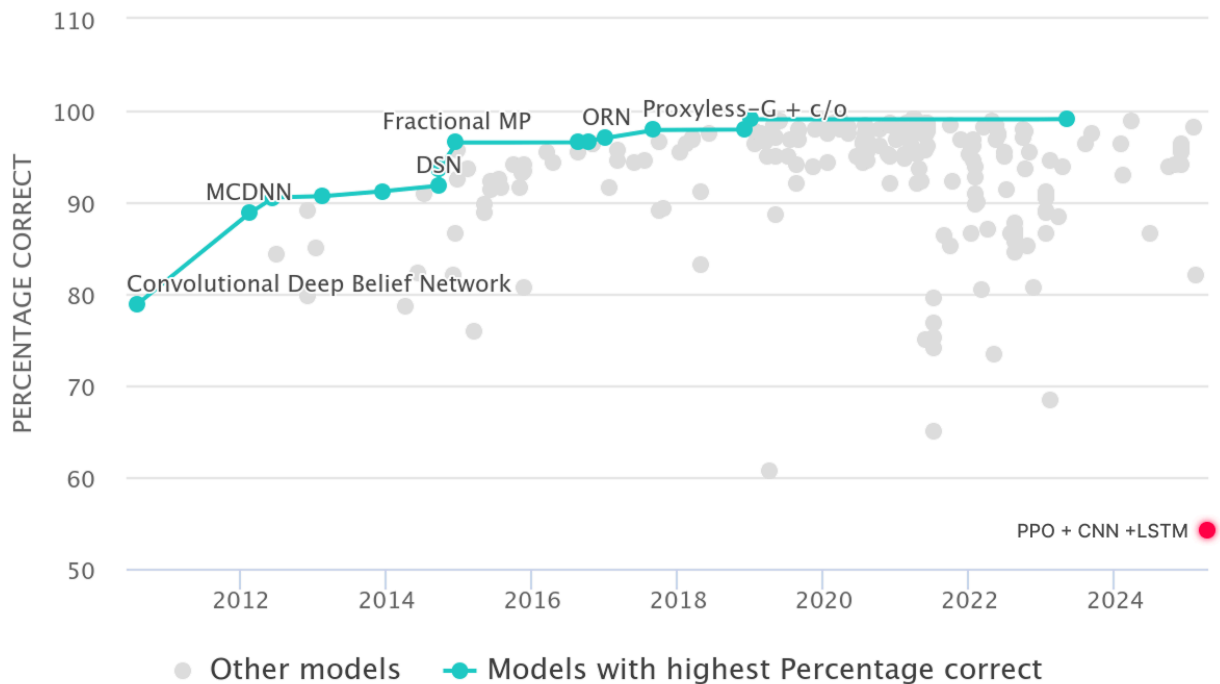
Модель использует комбинацию различных типов распределений для различных компонентов действия. Бернулли-распределение используется для бинарного решения о классификации, категориальное распределение - для предсказания класса. Непрерывные действия (движение и масштаб) обрабатываются детерминировано, т. к. это позволяет агенту более точно позиционировать вид в ограниченной среде. Это позволяет эффективно комбинировать дискретные и непрерывные действия в единой архитектуре.



Представлен график тестовой выборки быстрого обучения модели. Мы использовали тренировочную точность как основную метрику. Как видно сейчас она достаточно шумная,

однако модель способна к обучению и при более грамотном подборе гиперпараметров для избегания шумности обучения мы надеемся добиться достойного результата.

## Результаты



Как видно, наши модели не достигают SOTA результатов по точности. Однако динамика обучения положительная и в сравнении с традиционными методами мы смотрим на значительно меньшую часть изображения, что опережает другие решения в эффективности и скорости обработки. Наиболее заметными эти изменения, мы предполагаем, могут показаться на классификации изображений с более высоким разрешением.

## Выводы

Мы разработали интересные модели с исключительными особенностями, присущие человеческому зрению, которые были введены в первую очередь для эффективности. Мы верим, что при работе с изображениями большего качества и, особенно, при видеообработке эта архитектура сможет показать куда больше асимптотический выигрыш в эффективности распознавания. Пока что модели интересны исключительно с точки зрения реализации и требуют более тщательного подбора гиперпараметров и тренировки.

## Куда смотреть дальше?

Наиболее перспективным направлением дальнейшего развития мы считаем обработку видео. Эти модели внимания могут использовать как якорные элементы движущиеся части изображения, что позволит более эффективно понимать происходящее на видео, опираясь на самые важные элементы.

Следующей идеей является добавление настоящей фовеации, присущей человеческому зрению, при которой мы не жестко ограничиваем поле зрения, а лишь убираем информацию все больше и больше отдаваясь от точки обзора. Таким образом мы потенциально можем сжать куда больше иерархической информации в вектор той же размерности.