Summer with AIRI 2025

Active Vision for Image Understanding via Reinforcement Learning

1 Постановка задачи

Разработаем агент на основе обучения с подкреплением, 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:

- **Биологическая правдоподобность**: 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.

2 Обзор существующих решений

В 2014 г. Mnih et al. [4] была предложена RNN-архитектура, последовательно выбирающая координаты окна внимания и обрабатывающая вырезанные патчи с помощью гибридного обучения REINFORCE + supervised loss. В 2022 г. в Glance-and-Focus Networks [7] была введена двухэтапная стратегия, при которой сначала выполняется «glance» — обзор всего изображения в низком разрешении, а затем RL-политика переключается на «focus» для детального анализа выбранных регионов. В 2023 г. подход Active Vision RL under Limited Observability [9] расширил область применения на сложные сцены с высокой неоднородностью и ограниченными наблюдениями, задействовав actor–critic методы (PPO/A2C) для планирования серии «сдвигов взгляда». Параллельно работы по имитации биологических саккад — Saccade Mechanisms [2] — предложили эпизодические перемещения взгляда, основанные как на численных критериях минимизации неопределённости, так и на эвристических градиентных картах. Наконец, в 2025 г. Adaptive Patch Selection for ViT [1] интегрировал RL-агента с Vision Transformer, позволяя последнему обрабатывать лишь наиболее информативные патчи и тем самым значительно снижая вычислительные затраты без потери качества предсказаний.

3 Метод: PPO + CNN + LSTM

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.

3.1 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 [3]). 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 ε-greedy, где с вероятностью агент выбирает случайное действие.

3.2 Модель агента

Агент реализует рекуррентную нейронную сеть актор-критика:

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

3.3 PPO

PPO [6] 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 θ вычисляется как:

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

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

3.4 Обновление модели

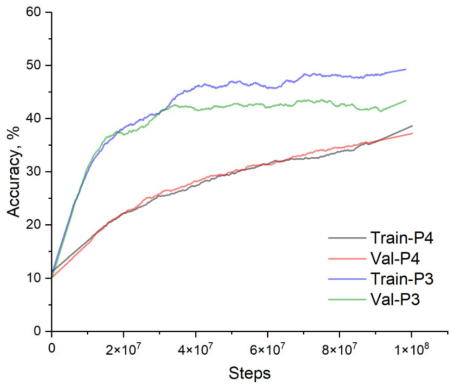
Для вычисления преимущества A_t используется метод Generalized Advantage Estimation (GAE) [5]. Целевая функция, включающая потерю по политике, потерю по ценности и регуляризацию энтропии для поощрения исследования, формулируется как:

$$L_{\text{total}} = L_{\text{policy}} + L_{\text{value}} - \beta H[\pi(\cdot|s_t)] \tag{2}$$

где L_{policy} – потеря по политике, L_{value} – потеря по ценности, – энтропия политики для предотвращения преждевременной сходимости к детерминированному решению, а β – коэффициент энтропии.

3.5 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 Categorial.



4 Метод: PPO + ViT

4.1 Архитектура агента

Агент основан на Vision Transformer (ViT) [8] с гибридным пространством действий. Модель состоит из патч-эмбеддинга, трансформера и специализированных головок для различных типов действий. Патч-эмбеддинг 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.

4.2 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.

4.3 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.

4.4 Процесс принятия решений и обработка наблюдений

Агент получает патч изображения вместе с координатами центра и информацией о масштабе. Эти данные проходят через патч-эмбеддинг, 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.

4.5 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.

4.6 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.

5 Результаты и выводы

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.

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.

6 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 отдаляясь от точки обзора. Таким образом мы потенциально можем сжать куда больше иерархической информации в вектор той же размерности.

Список литературы

[1] Francesco Cauteruccio, Marco Marchetti, Daniele Traini, Domenico Ursino, and Luca Virgili. Adaptive patch selection to improve vision transformers through reinforcement learning. *Applied Intelligence*, 55:607, 2025.

[2] Saurabh Farkya, Wenxi Li, Yifeng Zhang, Boxu Chen, and Le Yang. Saccade mechanisms for image classification, object detection and tracking. *arXiv preprint arXiv:2206.05102*, 2022. Presented at CVPR 2022 NeuroVision Workshop.

[3] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009. Original CIFAR-10 technical report.

[4] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27, pages 2204–2212, 2014. NeurIPS 2014.

[5] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015. Standard reference for GAE, not in search results.

[6] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. Standard reference for PPO, not directly in search results.

[7] Yulin Wang, Zanlin Ni, Shiji Song, Le Yang, and Gao Huang. Glance and focus networks for dynamic visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 45(8):1–15, 2022. Journal version of arXiv:2010.05300 (NeurIPS 2020).

[8] Gent Wu. Powerful design of small vision transformer on cifar10. *arXiv preprint arXiv:2501.06220*, 2025. Focuses on optimizing Tiny ViTs for small datasets like CIFAR-10.

[9] Yifan Zhu, Wenhao Li, Yifeng Zhang, and Sergey Levine. Active vision reinforcement learning under limited visual observability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. NeurIPS 2023.