

Image Contextual Bandits: A Visual Transformer Approach (Proposal)

Ryan W.
University of Maryland, College Park

1 Background

The Multi-Armed Bandit (MAB) problem provides a formal framework for sequential decision-making under uncertainty. In this setting, an agent repeatedly chooses from a set of actions (termed "arms") with initially unknown reward distributions. After each selection, the agent receives a stochastic reward and updates its strategy. The core challenge is the exploration-exploitation trade-off: the agent must balance exploiting arms that have yielded high rewards historically with exploring potentially superior but less-tried arms to refine its knowledge.

Contextual Bandits extend the classic MAB by incorporating side information, or context, available at each decision point. Here, each arm is associated with a context vector $x_{t,a}$ at time t . The agent's goal is to learn a mapping from contexts to arms that maximizes cumulative reward. This is highly relevant to real-world applications like personalized news article recommendation, where the context might include user demographics and article features.

This project focuses on the specific and challenging case of Image Contextual Bandits, where the context for each arm is a high-dimensional image. A prime application is visual-aware recommendation systems. For instance, on a movie streaming platform, each movie (arm) has a thumbnail (image context). A user's click or watch time provides a reward signal. The agent must learn to associate visual features from the thumbnail with the likelihood of user engagement.

2 Literature Review

The evolution of contextual bandit algorithms closely mirrors advances in function approximation techniques, particularly with the rise of deep learning. This section reviews the key developments that form the foundation for our proposed Visual Transformer-based approach.

2.1 Linear Contextual Bandits: LinUCB

The first major contextual bandit algorithm, **LinUCB** [2], addressed the problem under the assumption of a linear relationship between context vectors and expected rewards. Formally, for an arm a at time t with context vector $x_{t,a} \in \mathbb{R}^d$, the expected reward is given by:

$$\mathbb{E}[r_{t,a}] = \theta^* \top x_{t,a}$$

where θ^* is an unknown parameter vector. LinUCB maintains an estimate $\hat{\theta}_t$ using ridge regression and computes an Upper Confidence Bound (UCB) for each arm:

$$U_{t,a} = \hat{\theta}_t \top x_{t,a} + \alpha \sqrt{x_{t,a} \top A_t^{-1} x_{t,a}}$$

where $A_t = \lambda I + \sum_{s=1}^{t-1} x_{s,a_s} x_{s,a_s} \top$ is the regularized covariance matrix. The first term represents exploitation (predicted reward), while the second term encourages exploration based on estimation uncertainty.

While computationally efficient and providing theoretical guarantees, the linearity assumption severely limits LinUCB’s applicability to complex, high-dimensional contexts like images, where reward functions are inherently non-linear.

2.2 Non-Linear Generalizations: Neural Bandits

NeuralUCB [4] and related approaches overcome the linearity limitation by using neural networks to model the unknown reward function. The exploit term becomes $f(x_{t,a}; \theta_t)$, the output of a neural network with parameters θ_t . The key innovation lies in the exploration term, which leverages neural tangent kernel (NTK) theory to approximate uncertainty:

$$U_{t,a} = f(x_{t,a}; \theta_t) + \alpha \sqrt{g(x_{t,a}; \theta_t)^\top (\lambda I + G_t)^{-1} g(x_{t,a}; \theta_t)}$$

where $g(x_{t,a}; \theta_t) = \nabla_{\theta} f(x_{t,a}; \theta_t)$ is the gradient of the network’s output with respect to its parameters, and $G_t = \sum_{s=1}^{t-1} g(x_{s,a_s}; \theta_s) g(x_{s,a_s}; \theta_s)^\top$ accumulates past gradients.

This approach enables handling of non-linear reward functions but introduces computational challenges due to the need for gradient computations and maintaining the matrix G_t .

2.3 Image-Aware Bandits: CNN-UCB

CNN-UCB [1] specifically adapts the NeuralUCB framework for image contexts by employing a Convolutional Neural Network (CNN) as the backbone architecture.

2.4 Visual Transformers for Bandit Problems

Visual Transformers (ViTs) [3] process images as sequences of patches and use self-attention mechanisms to model relationships between all patches simultaneously. This provides several advantages for contextual bandits:

- **Global context understanding:** Self-attention captures long-range dependencies across the entire image
- **Adaptive feature weighting:** Dynamic importance assignment to different image regions
- **Scalability to high-resolution images:** Computational complexity grows linearly with sequence length

To our knowledge, no existing work has integrated ViTs into the contextual bandit framework. Our proposed **ViT-UCB** algorithm aims to bridge this gap by leveraging ViTs’ superior global feature extraction capabilities while maintaining the theoretical foundations of neural bandit algorithms.

3 Project Goals, Validation, and Deliverables

My goal is to develop and implement a ViT based contextual bandit algorithm where the context is an image. There will be a paper detailing the algorithm (and hopefully theoretical regret bounds).

To verify that the algorithm is effective, an experiment would be conducted to show that the ViT based approach accumulates more reward and suffers lower regret than existing algorithms such as LinUCB and CNUCB. This experiment should generate graphs that support the result, and its code should be released on GitHub

4 Approach

Since most related works utilized a UCB-based approach, this project will most likely do so as well, where the exploit term would be an estimate of the reward using a mixture of a ViT and a traditional feed-forward NN (FFNN). Currently, there are 2 possibilities for updating the estimate: freezing the ViT and just update the NN, or update both the ViT (using LoRA based approach) and NN. See algorithm outline below:

1. Observe arm 1’s image context
2. Compute its upper confidence bound (Sum of ViT-FFNN and an explore term)
3. Repeat Steps 1-2 for each arm

4. Choose and play arm with highest UCB and observe reward
5. Update FFNN (and ViT)
6. Repeat steps 1-5 for each time step

The experiment would be synthetic but with a meaningful structure. Specifically, each arm would be associated with a movie thumbnail (the context) and generate rewards from a Gaussian distribution centered on that movie’s Rotten Tomatoes score. This setup ensures that image features are somewhat correlated with the expected reward, creating a learnable relation without requiring a real deployment environment.

5 Documentation

The following documentations would be made:

- Literature review
 - Detailed annotations on each paper read
 - Summary of all papers read
- GitHub
 - Implementation of algorithm
 - Script to generate data
 - Experiment code and visualizations
- Note down every modification made to the algorithm
- (If possible) derivation of theoretical bounds of algorithm

6 Project Timeline

Table 1: Project Timeline Overview

Time Period	Major Milestones
Weeks 1-2	Literature review completion Synthetic environment setup Baseline implementation started
Weeks 3-4	All baseline algorithms implemented CNN-UCB validation completed Evaluation framework established
Weeks 5-6	ViT-UCB core implementation Two training strategies developed Initial hyperparameter tuning
Weeks 7-8	Ablation studies completed Full comparative experiments run All results collected and analyzed
Weeks 9-10	Theoretical analysis (time permitting) Paper writing and documentation Code repository preparation
Week 11	Final revisions and polishing Project submission

References

- [1] BAN, Y., AND HE, J. Convolutional neural bandit for visual-aware recommendation, 2022.
- [2] CHU, W., LI, L., REYZIN, L., AND SCHAPIRE, R. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (Fort Lauderdale, FL, USA, 11–13 Apr 2011), G. Gordon, D. Dunson, and M. Dudík, Eds., vol. 15 of *Proceedings of Machine Learning Research*, PMLR, pp. 208–214.
- [3] DOSOVITSKIY, A., BEYER, L., KOLESNIKOV, A., WEISSENBERN, D., ZHAI, X., UNTERTHINER, T., DEHGHANI, M., MINDERER, M., HEIGOLD, G., GELLY, S., USZKOREIT, J., AND HOULSBY, N. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.
- [4] ZHOU, D., LI, L., AND GU, Q. Neural contextual bandits with UCB-based exploration. In *Proceedings of the 37th International Conference on Machine Learning* (13–18 Jul 2020), H. D. III and A. Singh, Eds., vol. 119 of *Proceedings of Machine Learning Research*, PMLR, pp. 11492–11502.