CS371-2025S: Final Project Proposal
"PhishClip: CLIP-based Prompt-tuned Phishing Site Screenshot Classifier"

Hankyul Jang

1. Introduction

Phishing websites visually imitate trusted login pages to steal user credentials, often targeting non-technical or digitally vulnerable individuals. Traditional detection methods, such as rule-based URL filters, struggle to identify sophisticated visual mimicry or redirection tactics.

To address this, we propose extbfPhishClip, a phishing screenshot classifier that leverages CLIP (Contrastive Language-Image Pretraining) and CoOp (Context Optimization). Our model uses soft prompt tuning to improve classification performance in few-shot scenarios. We evaluate our method on the Phishpedia benchmark dataset and compare against two baselines: zero-shot CLIP and the original Phishpedia model.

This project supports the United Nations Sustainable Development Goal 10 (Reduced Inequalities) by providing accessible phishing detection tools to protect digitally underserved populations and close security gaps across user groups.

2. Problem Definition & Challenges

Phishing website detection via screenshots presents a challenging problem where models must distinguish legitimate login pages from malicious lookalikes under limited supervision.

The key challenges in this task include:

3. Problem Definition & Challenges

Phishing website detection via screenshots presents a challenging problem where models must distinguish legitimate login pages from malicious lookalikes under limited supervision.

The key challenges in this task include:

- **Visual Mimicry:** Phishing sites are carefully crafted to appear nearly identical to real services, making traditional visual rules or template matching ineffective.
- **Domain Adaptation:** Attackers frequently change layouts, targets, and domains. Models must generalize to unseen styles and brands.
- **Data Scarcity:** It is difficult to maintain up-to-date labeled datasets of phishing sites due to the constantly evolving nature of attacks.
- **Prompt Robustness:** Zero-shot models like CLIP rely on manually crafted prompts, which may not fully capture phishing-specific semantics. Our CoOp-based approach learns task-adaptive soft prompts that better reflect the visual subtleties of phishing attempts.
- **Joint Multimodal Reasoning:** Combining visual and textual (URL-based) cues into a unified classifier structure is non-trivial and presents integration challenges.

4. Related Work

Phishpedia (Zhang et al., 2021) uses logo detection and a Siamese network for brand-specific visual matching, but requires explicit brand reference images. CLIP (Radford et al., 2021) introduces vision-language pretraining for zero-shot image classification. CoOp (Zhou et al., 2022) extends CLIP by learning soft prompts in few-shot settings, showing improved generalization over handcrafted prompts. Recent phishing detection systems also include URL-based methods, but they struggle with visual mimicry. Our work uniquely combines these paradigms with CoOp to enhance robustness against real-world phishing attempts.

5. Dataset

We use the public Phishpedia benchmark dataset (Zhang et al., 2021), which includes over 29,000 phishing pages and 30,000 legitimate pages. Each sample contains a rendered screenshot, full HTML, and its URL. The dataset includes labeled brand logos and metadata across 180+ target brands. We preprocess the screenshots to match CLIP's input resolution and tokenize domain URLs for use in prompt generation.

6. State-of-the-art Methods and Baselines

We compare our approach with two key baselines:

- **Phishpedia (Zhang et al.)**: Visual similarity using brand-specific logo detection and reference templates.
- **Zero-shot CLIP**: Using standard prompts without any fine-tuning.

Our method, CoOp-tuned CLIP, extends the latter by learning prompts tailored to phishing semantics. This enables generalization to unseen brands and supports few-shot learning.

7. Schedule

- Week 1: Dataset download, preprocessing, and zero-shot CLIP baseline.
- Week 2: CoOp implementation and prompt tuning.
- Week 3: Phishpedia baseline evaluation and comparative testing.
- Week 4: Final experiments, ablation studies, and report writing.

References

Radford, A., Kim, J. W., et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Zhang, Z., Lin, Z., Liu, X., and Tague, P. Phishpedia: A hybrid deep learning based approach to visually identify phishing webpages. In *USENIX Security Symposium*, 2021.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. In *CVPR*, 2022.