Phishing Detector

Tianhao Wu, Qian Liu, Mengjun Wang, Fujiao Ji

May 03, 2024

Outline:

- Introduction
- Motivation
- Model Overview

- **Model Design**
    - Data Collection
    - Faster-RCNN: used for crop the logo from screenshot
    - ViT: used for compare logo similarity with the target list
    - Involution: same as ViT
- Results

Introduction

Background:

A phishing website is a fraudulent site that mimics legitimate websites with the intent to deceive users into providing sensitive information such as usernames and passwords.

Motivation:

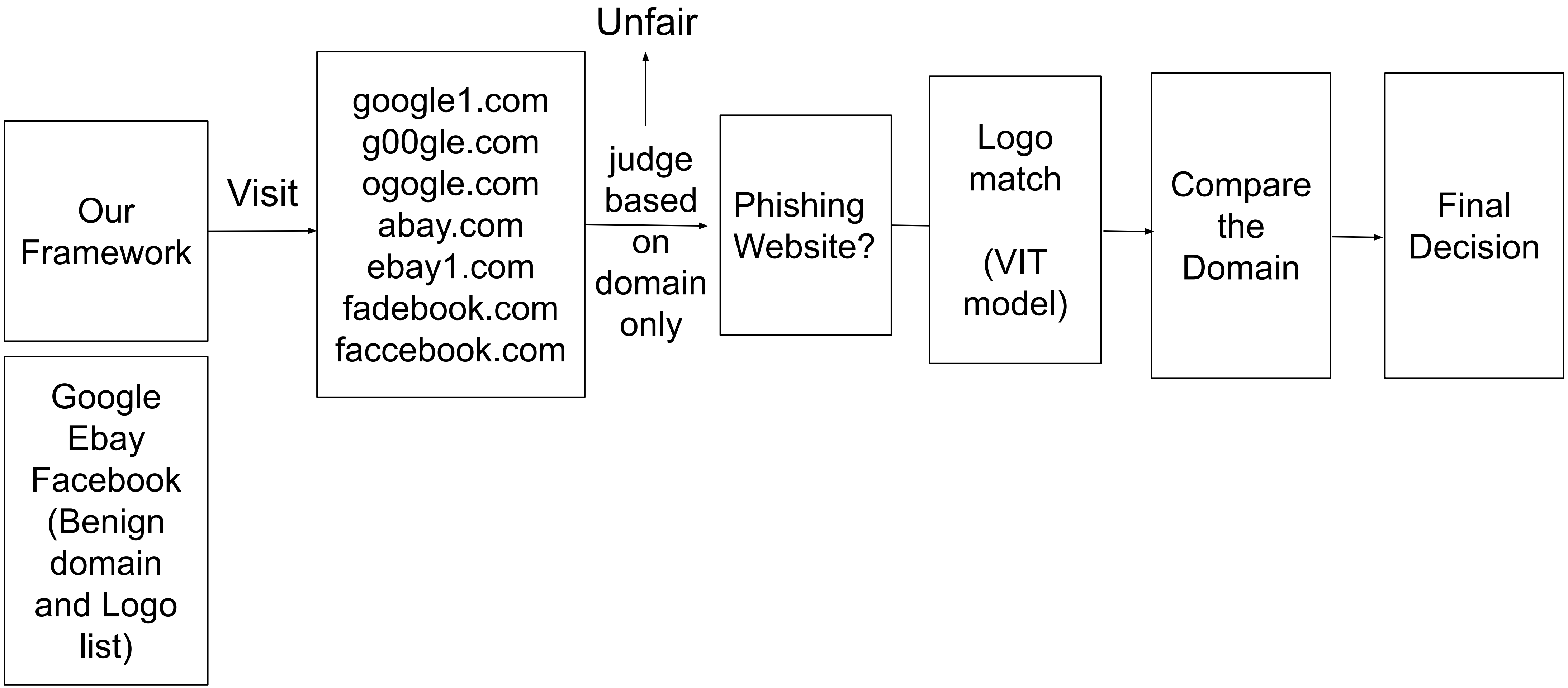
It is not fair to determine the website is a phishing website or not based only on domain information.

Research Goal:

Develop a phishing detector framework to recognize the phishing websites based on brands' logo and domain information.

Google	**google.com**
facebook	**facebook.com**
ebay	**ebay.com**

Overview of Our Framework



Our Framework

Google
Ebay
Facebook
(Benign domain and Logo list)

Visit

google1.com
g00gle.com
ogogle.com
abay.com
ebay1.com
fadebook.com
faccebook.com

Unfair

judge based on domain only

Phishing Website?

Logo match

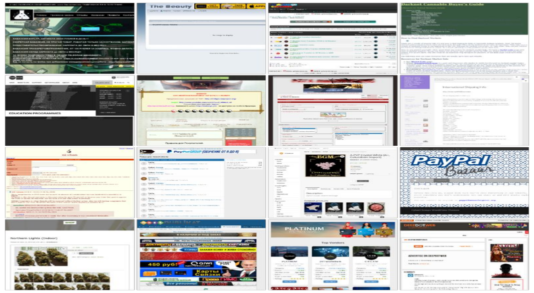
(VIT model)

Compare the Domain

Final Decision

Dataset

1. Faster RCNN: benign30k



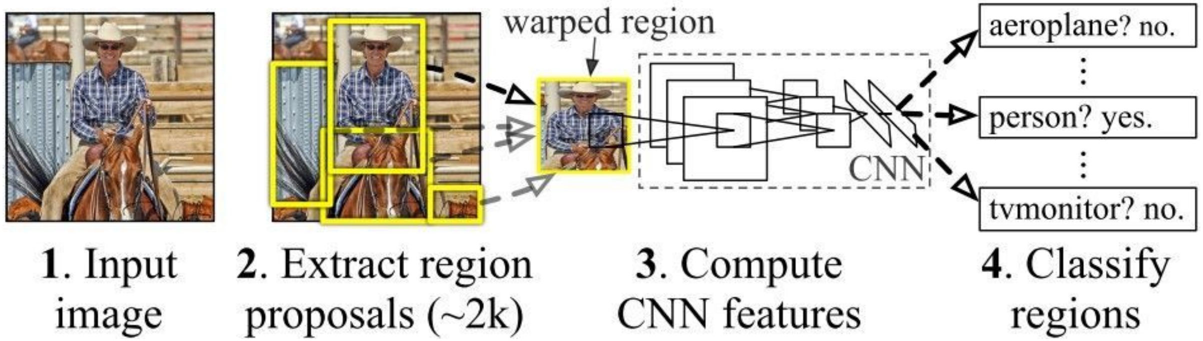
2. ViT & Involution: Logo2K+

4. Threshold: benign25k and phish25k

4. Evaluation: Web Archive

Logo Clip: Faster R-CNN

Faster R-CNN (Region-based Convolutional Neural Network) is an advanced deep learning model that builds on the foundations of previous models like R-CNN and Fast R-CNN to efficiently and accurately **detect objects** within images.

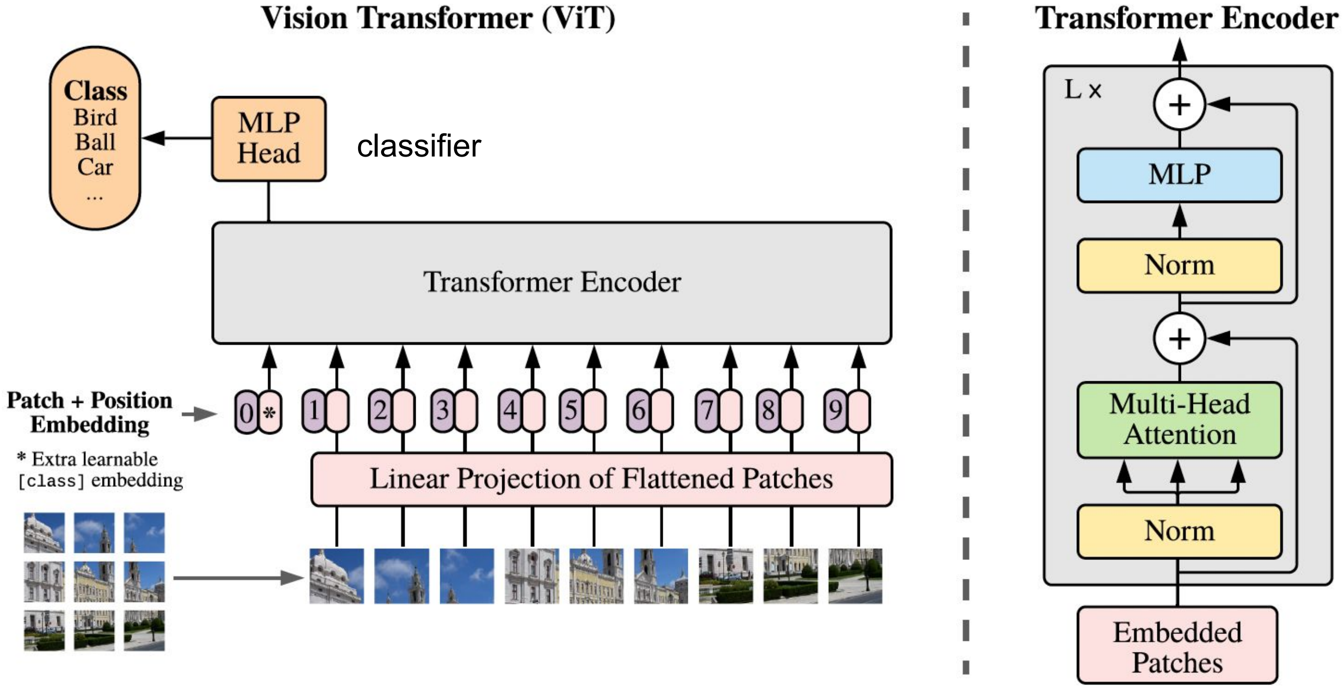


Method — ViT

ViT is used to **compare logo similarity** by encoding logos as sequences of image patches and then comparing their embeddings to a target list.

Model Structure:

- ❖ The input image is divided into fixed-size patches.
- ❖ These patches are linearly embedded into a higher-dimensional space.
- ❖ A sequence of positional embeddings is added to these patch embeddings to retain positional information.
- ❖ The resulting sequence of vectors is fed into a standard transformer encoder that uses self-attention mechanisms.
- ❖ The output of the transformer can be used for various tasks, including classification and similarity comparison.

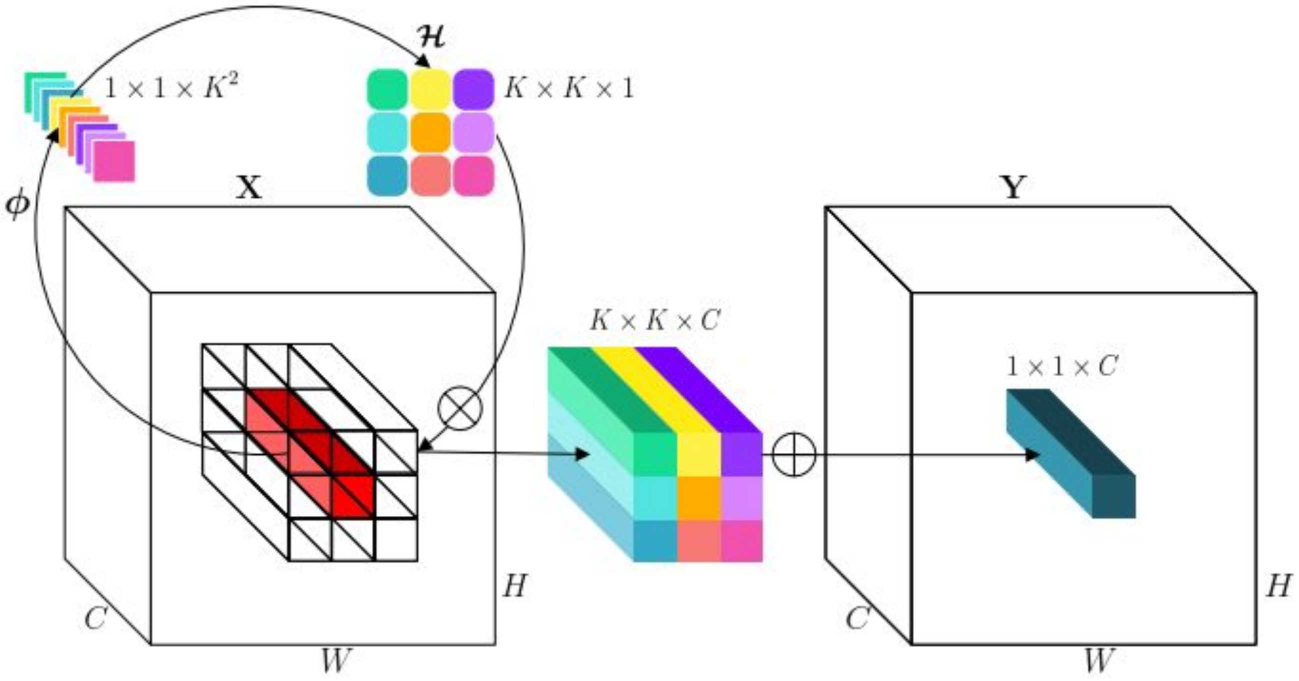


Method — Involution

Similar to ViT, involution can be adapted for tasks like **comparing visual similarities** by generating dynamic kernels that adapt to the specific features of each logo.

Model Structure:

- Instead of using fixed kernels, involution generates kernels based on the input feature map itself.
- This generation process involves a kernel-generating function that maps the input features to kernel weights.
- These dynamic kernels are then applied spatially across the input feature map.
- The output feature map reflects more adaptive and input-specific processing, which can be beneficial for detailed tasks like logo comparison.



Method — Novelty

- **ViT** introduces the application of transformers, previously used primarily in natural language processing, to the domain of image recognition. It treats images as sequences of patches and processes them using a transformer encoder, which allows it to *capture complex spatial hierarchies*.

- **Involution** is designed as an alternative to convolution, aiming to reduce computational complexity and increase model efficiency. Unlike convolution, which aggregates features over a spatial area using shared weights, involution uses spatial-specific weights. This means the weights are generated dynamically based on the input, allowing for *more adaptable and data-dependent processing*.

Evaluation Dataset: collected from Archive.org for recent 7 years change

- Total: 1,941 with 3 brands
  - Benign: 501
  - Phishing: 1440

Results

Brand	Total	ViT Detection Rate	Involution Detection Rate
Ebay	390	0.8051282051	0.9923076923
Google	393	0.2875318066	0.8727735369
Facebook	1158	0.2815198618	0.9343696028
Total	1941	0.3879443586	0.9335394127

Any Question?