Advertisement Position Investigation

Ruofan Hu, Haolun Zheng, Weijing Guo, Yanwei Xu, Qi Wang

1 Summary

During our Internet surfing, almost every web page we browsing has some ads, generally recommending commodities, software products and so on. In common place, prime contents concentrate in the middle of a page, with both sides vacant. As a result, sites use these spaces to display ads. In this report, we investigated thousands of web pages and collected corresponding amount of advertising images. These web pages cover both popular and obscure sites abroad and domestic. We detected ads from those pages and trained a CNN-based dataset.

We divided those ads into four types in terms of contents, texts as well as our model's judgement and then analyzed their average position, from which we deduced that different kind of ads are located in different position likewise. More significantly, we also found some key takeaways of how to get an ad more possibly viewed.

2 Background

In this part, we state the definition of the ad and what can the position of ads impact on.

An advertisement is a notice, picture or film telling people about a product, job or service, which can be non-profit or commercial. But ads online are bound to try attracting attention and tempting viewers into a click to make money. In order to do it, a good eye-catching position together with not annoying user is of great significance. When contriving to monetize a website, whether through display advertisements, affiliate marketing, product sales, or some combination, one of the most important things you must do is to decide "what goes where". Notwithstanding the quality and type of products determine in a large extent the volume and caliber of your audience, the way one arranges those pieces can make a difference to how much one can benefit from the traffic.

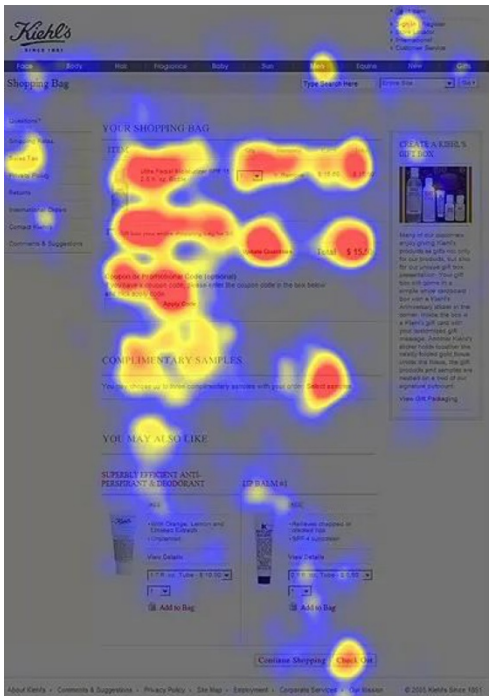


Figure 1: How one views a page.

It manifests how a viewer to a site consumes the content, where eyes apt to and which parts are most likely overlooked completely. It is clear that people mainly concern about middle of pages and seldom shoot a glance at corners and sides. We can also infer that the top of pages is neglected as well.

3 Data Collection Methodology

3.1 Data Description

First, let's provide a description of our dataset. An ad is usually an image with striking texts as well as some exaggerated pictures of what it points to. Our dataset includes the following information for such image: the source of the image, the coordinate position of the image, the size of the image, the brightness and saturation of the image, the type of advertisement, and the relatively clear text that can be extracted from the image.What's more, we divided ads to four types by their contents and areas pointed to, respectively non-ads, regular ads, gambling ads and pornographic ads.

Then, to ensure our data collected is reliable and valid from different links, we search for keywords of each type and extract their pages by a web crawler developed by the Selenium WebDriver framework with python programming. Search engine tends to sort search results by their popularity. In order to get unpopular sites' information and ads, we used page flips to gather backward pages.

In above process, we need to collect approximate 2000 pages together with their ad which are half domestic and half foreign. We know that some links may appear multiple times when searching for different keywords, which leads to conflicts with the 2,000 different URL links we asked to collect. Therefore, we use a list to record URL links at the time of collection, and appropriately expand the number of keywords, expect to collect up to 3000 links, and expect 2000 unique links

Now that we had raw materials, we processed them to get ads and recognize their kinds as well as positions.

3.2 Image Processing

There are two main types of image sources in our dataset: URL or base64 direct rendering. The images sourced from URLs are relatively easy to handle as they can be downloaded to the local system after retrieving the image, allowing for subsequent ad detection and processing. However, for the images that are encoded in base64 format, we need to first decode and synthesize the corresponding base64 code before we can effectively work with them.

During this process, we have observed that normal images and regular advertisements are mostly in the form of URL links. However, for non-standard ads such as explicit or gambling advertisements, a significant portion of them are encoded in Base64 format. With these Base64-encoded images, we often encounter challenges where we can only obtain a small, useless image resembling a cross mark, and there may also be issues with incomplete Base64 encoding. This highlights the limitations of extracting images from HTML and suggests the need to consider additional programming techniques such as JavaScript.

Fortunately, by expanding our data sources, we have minimized the impact of incomplete Base64 encodings, as they represent a very small proportion of the dataset.

3.3 Relative Position Processing

We accessed to the absolute coordinates of images mainly through the relevant API interface of Selenium. Then according to the resolution of the screen, we attained the relative position of each image.

In a broader sense, there are two issues with coordinates: pagination and invalid coordinates. Our approach is as follows. For each webpage, we use our screen resolution as a reference. For example, if our screen height is 1000 pixels and we obtain a coordinate of 1500, it means that the image should be positioned at 1.5 pages. Additionally, we have observed

that some collected coordinates are (0, 0), which indicates that the image is not fully expanded. These types of coordinates often relate to images that require certain clicks on the webpage to be visible. This further emphasizes the limitation of solely scraping images from HTML. Since the number of such images is small, we have chosen to omit secondary processing for them.

3.4 Ad Detection

We combined text extraction tools and model detection tools to detect ads. Our specific process is to first extract the text information from the images and match it with the text commonly found in advertisements. If there is a successful match, the image is considered an advertisement. However, the presence of multiple types of advertising features in the image based on the displayed words may require further determination using a recognition model to identify its final type.

In text extraction, there are two issues: how to perform matching and which words to match.

Regarding word matching resources, we have prepared 1000-1500 words for each category as keywords. We utilize ChatGPT for pre-generation, and then we undergo manual review to add and modify the keywords.

For the text word extraction and matching step, we have tested several open-source tools such as Cnocr and EasyOCR. After conducting preliminary tests on sample images, we have selected Cnocr as it provides accurate extraction for both horizontal and vertical text.

Cnocr offers multiple matching probabilities for a given text word. During the comparison process, we select the highest matching probability among those exceeding a predefined threshold. For example, let's consider a text word "Gam-ing" with matching probabilities of 40%, 45%, and 80% against the word "Gaming". If our threshold is set at 50%, since the matching probability of 80% is above the threshold, we consider it a successful match with "Gaming". Thus, we label the image accordingly under the "Gaming" category. We repeat this matching process for other categories' keywords as well.

As for the model, it’s thanks to dataset provided by our teaching assistant Yuxuan Shang, with which we trained a CNN-based model. This CNN model is very simple, yet it still achieves an accuracy of 72% on the test set. During training, we encountered the issue of data imbalance, as the number of images varied significantly across different categories. For example, we had 1000 images for the regular category but only 500 images for thess yellow category. To address this problem, we employed data augmentation techniques such as data duplication to achieve a balance in the dataset.

4 Dataset Description

When collecting data, we first gathered Chinese websites and then English websites. During the process of collecting images from Chinese websites, we encountered unexpected bugs, including issues with links. As a result, we had to divide the collection process into multiple stages, storing the data in different folders within 'Chinese_excels'. However, when collecting English data, we were able to resolve all the bugs and smoothly collected the content in batches, storing it in the 'pic' folder.

The directories of the data:

```
├──Chinese_excels
|    ├──1
|    ├──2
|    ├──3
|    └──4
├──pic
```

The data are all collected from the Internet, containing the Chinese and English urls. And Before data

washing, there are 9638 data in the dataset, and are separated into 4 types:

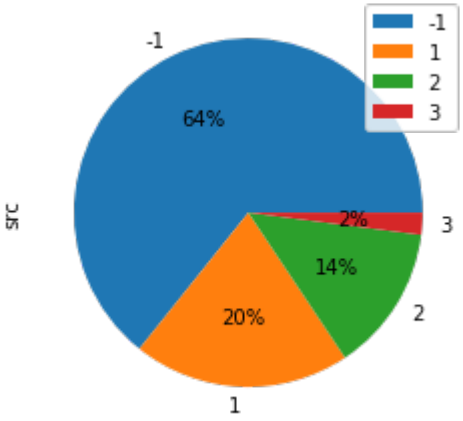
- not ads: -1

- gambling ads: 1

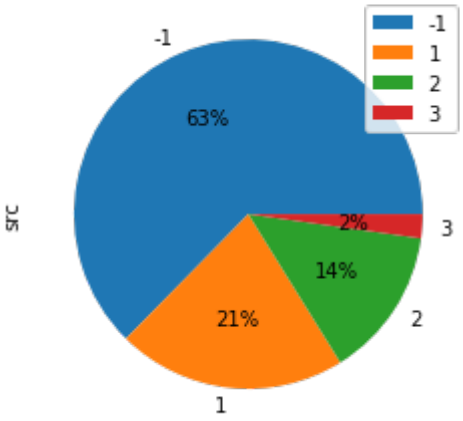
- regular ads: 2

- porn ads: 3

And their distribution in the dataset is:



After the data washing(drop the non-sense data and transform the data to the right format), the total number of the data is: 8864. And the distribution of the data is:



5 Data Analysis and Findings

And in the process of analyze the data, I met the following problems:

- The data is like a mess, the different formats and the useless(bad) data need to be dropped out.

- The data showed no regulation at the first glance. It need to be digged out carefully.

- The displacement of the different feature of the data, need to be thought about.

To solve the problems above, we find information on the Internet to do the data washing and the transformation. Then we thought about the useful feature among the data, which can show some regulation. We calculated the relevant size of x's and y's. To do further analysis, I even thought about, like the KMeans, such Machine Learning algorithms, to fix up the bad performance of the classification. However, by observing the data, we noticed that, the distribution of the different types of the ad pictures, gathering at the certain area. If use the mean value of the distribution of the center of the pictures and the mean value of the relevant size x's and y's, a rectangle can be drawn, to cover the distribution. This rectangle implicates the regulation of the different types of ad pictures.

We calculated the relative size of the images and plotted allocation maps of four types respectively.

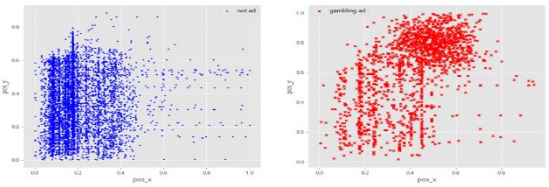


Figure 2: Allocation of non-ad(left) and gambling ad(right).

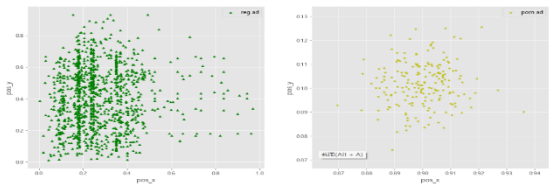


Figure 3: Allocation of regular ad(left) and porn-ad(right).

The non-ads are likely located on the left side of the page, and they don't reach to the top and become rare there. It is the same with regular advertisements. Yet, the situation changes when it turns to gambling ads, they not only hold left side, but also spread to the top middle space, occupying over 50 percent room of a page. And conversely, pornographic ads concentrate on the corner in the right side, just a tiny part, forming a great contrast with gambling ads, as they are both grey sectors.

Then we put these four maps in one map to see their relative relation and pattern.

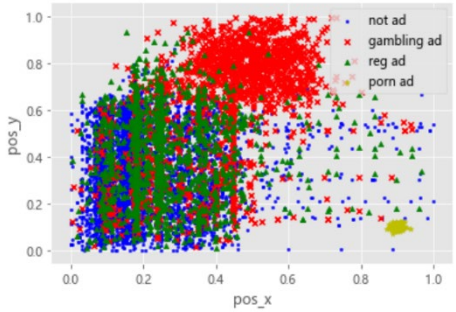


Figure 4: Allocation of all ads.

This image visually displays the distribution of all types of advertisements on a web page.

Furthermore, we drew the histogram of the distribution of all the types on the x direction, y direction, the width distribution and length in addition.

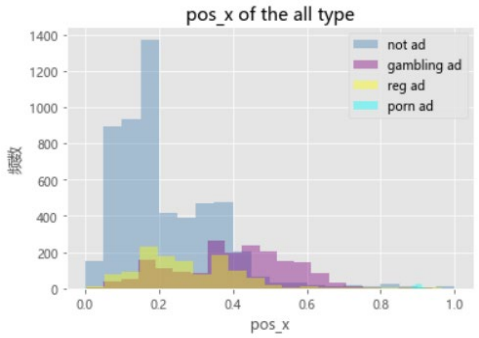


Figure 5: Distribution histogram of x-axis.

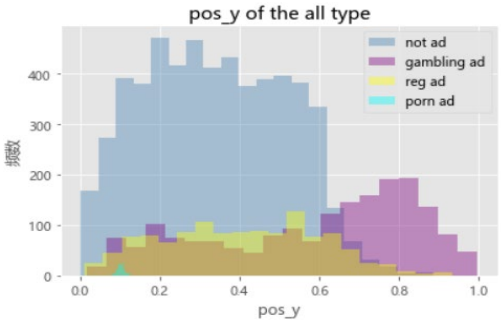


Figure 6: Distribution histogram of y-axis.

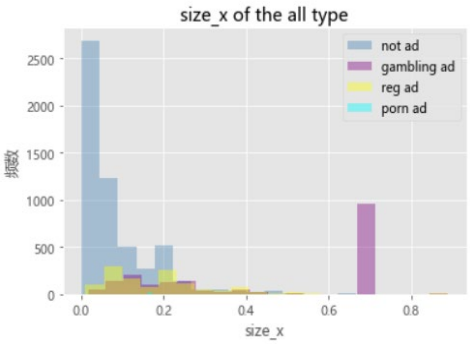


Figure 7: Distribution histogram of width.

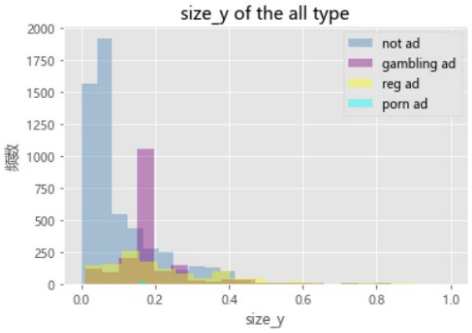


Figure 8: Distribution histogram of height.

We can infer from the figures that non-ads account for a big rate in images occurred on the surface of web pages, which commonly exist in the left side with a quite small width and height. The second most ad type is gambling, usually having a tremendous width and middle-sized height compared to the others. As for regular ads, they are regular as well, neither too big or too small, nor too many. Pornographic ads are not only few, but also small, seeming not so attractive.

Finally, we made an illustration of average position together with its occupation of each type of image, expressed as a rectangle in different colors related to the normalized web page.

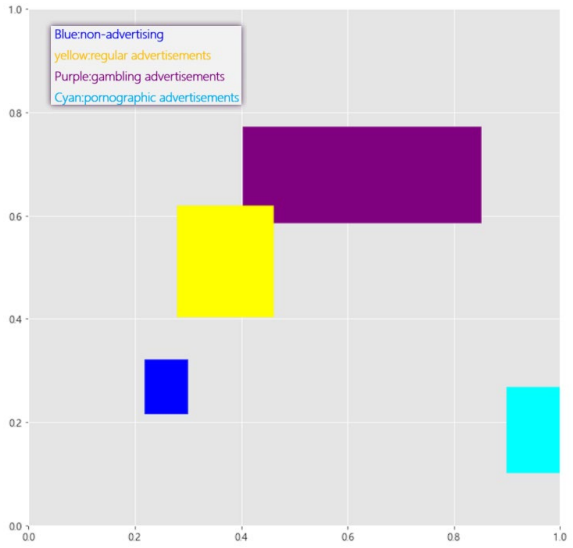


Figure 9: Average position of each type of image.

We had considered the main distribution of each advertising location. And in order to differentiate, we highlighted the part with the least intersection with other types of advertising distribution in the figure. It pretty clearly shows that regular ads are scattered in the lower left. Non-ads are above regular ads. Gambling ads are a little higher and in a medium position in x direction, which take over a large range. Pornographic ads are distributed in the lower right corner.

Besides, we extracted the position and size information from each type of image and made a covariance table to check the correlation.

	not_x	not_y	gam_x	gam_y	reg_x	reg_y
not_x	1	0.067	0.073	0.069	-0.065	0.024
not_y	0.067	1	0.023	0.041	0.028	0.035
gam_x	0.073	0.023	1	0.44	-0.014	-0.02
gam_y	0.069	0.041	0.44	1	-0.0022	-0.024
reg_x	-0.065	0.028	-0.014	-0.0022	1	0.032
reg_y	0.024	0.035	-0.02	-0.024	0.032	1

Table 1: Covariance table of each types' position.

The correlation map of the different features collected before, the relation between the different features is quite weak, meaning they don't disturb the distribution of each other, among the different types of ad pictures.

Not all positions are created equal. After multiple decades of Internet access (and millennia of reading), humans have become creatures of habit when it comes to digesting content. As a result, certain areas of your site will be inherently more popular than others, not because of what they contain but because of where they’re positioned.

Left is right. It’s nature of us to dip from left to right. Our gaze starts left and shifts right, but our focus wanes in the process. So, here’s the key takeaway: the left side of the page will get much more attention than the right.

Height disadvantage. When we click on a web page and enter it, we spontaneously scroll down at the first time, because we know useful content scarcely located on the top. So, the top of a page is often overlooked.

Hot spots. We all know useful contents are concentrated in the midst of a page, as a result, we only look through them and ignore corner and side. So, more sites are starting to mingle ads with contents, ensuring it get increasing views and clicks.

6 Conclusion

In brief, non-advertising and regular ads mainly situated on the left side of a page, but advertisements are usually larger than non-ads. And pornographic ads take over the other side. Gambling related images often occur above, with a large span in x direction, while the others more in y axis.

And among these three types of advertising, the width of gambling ads is very large, generally more than half of the screen, while the size of porn ads is generally relatively small. The regular ads and non-ads have the normal size.

Besides, the relation between different features is weak, which means the distribution of different kinds of ads don't disturb each other. This phenomenon shows that it is possible to analyze the various advertisements separately

In conclusion, the size of gambling advertisements is generally the largest of all advertisements, and the position is mostly distributed directly above the screen, it can be said that the attraction is relatively large, and porn advertisements do not have strong attraction in terms of size and position, which can also show that the investment in gambling advertising is much greater than porn. And the normal ads and non-ads is normal both size and position.

7 Miscellaneous

When we crawled for web pages at first time, we sent requests, finding the code we got didn't have ads next to them. After checking the detail information, it turned out that ads might be rendered in post with JavaScript at a larger stage resulting in. Resultantly, we used the selenium automation library to simulate our behavior toward browsing and then went in and crawled all the images on the page.

When we dealt with base64 code, we found some codes unavailable to dispose. We assumed that it's too long because the code was not complete, so we could not recompose these pictures.

When we disposed of relative position, we found some images' coordinates are (0, 0), which is obviously impossible. After checking the detail information, it's clear that these pictures are not straightforward displayed, meaning that it may need a click on something to appear to the surface. So, it hides in the position (0, 0) until it is triggered. Moreover, most of these hidden images are normal images, consequently we abandoned them when we processed images' positions.

When we trained the model, we had encountered data augmentation due to the very different amount of data for each type of advertisement image, and data enhancement for some data processing such as copying and flipping to balance the amount of each image. But we need to say, the advertisement model is a little difficult to train, and the model we use is just use simple CNN, which is not adapted to difficult task.

It was also found difficult to collect URLs with **s**pecific pornographic game advertisements, most of the pages are still normal pages. But we hadn't managed to settle it down. We considered making a pre-judgement to decide whether to keep a page as data source.

Besides, we once used the ad block, which is a Python interface to check the picture link is advertisements or not, but it can not distinguish the most pictures we found in the website, although we have used different text to help it. So we give up the tool. But the TA once said the ad block is used widely in the practice, we thought it maybe have some other things we do not take into consideration.