When the Few Outweigh the Many: Illicit Content Recognition with Few-Shot Learning

G. Cascavilla[1], G. Catolino[2], M. Conti[3], D. Mellios[2] and D. A. Tamburri[1]

[1] *Eindhoven University of Technology, Jheronimus Academy of Data Science, The Netherlands*

[2] *Tilburg University, Jheronimus Academy of Data Science, The Netherlands*

[3] *University of Padova, Italy*

Keywords: Siamese Neural Network, Dark Web, One-Shot Learning, Few-Shot Learning, Cybersecurity.

Abstract: The anonymity and untraceability benefits of the Dark web account for the exponentially-increased potential of its popularity while creating a suitable womb for many illicit activities, to date. Hence, in collaboration with cybersecurity and law enforcement agencies, research has provided approaches for recognizing and classifying illicit activities with most exploiting textual dark web markets' content recognition; few such approaches use images that originated from dark web content. This paper investigates this alternative technique for recognizing illegal activities from images. In particular, we investigate label-agnostic learning techniques like One-Shot and Few-Shot learning featuring the use Siamese neural networks, a state-of-the-art approach in the field. Our solution manages to handle small-scale datasets with promising accuracy. In particular, Siamese neural networks reach 90.9% on 20-Shot experiments over a 10-class dataset; this leads us to conclude that such models are a promising and cheaper alternative to the definition of automated law-enforcing machinery over the dark web.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

The web as we know it today has two primary layers. On the one hand, the Surface web offers most if not all the web pages we normally use daily. On the other hand, the Deep web—or hidden web(Raghavan and García-Molina, 2001)—offers parts of the World Wide Web whose contents are not indexed by standard web search-engine programs. The latter remains unindexed because its content is considered either irrelevant or confidential, and for security purposes, it is intentionally concealed. The advantages above, however, create a suitable womb for many illicit activities concealed from regular search indexing. Such activities collectively form a relatively small fraction of the Deep web, called the Dark Web (Godawatte et al., 2019). The Dark Web uses the Tor—The Onion Routing[1]—network to access its content, and featuring a sensibly different architecture than the Surface web; for example, each request is redirected through various remote servers to reach the requested content and finally return to the user via a different server, thereby making that request untraceable.

Although the Dark Web amounts to circa 0.005% of the web (Juan Sanchez, 2019), only 48% of the Dark Web content is legal (Al-Nabki et al., 2019), with the rest being illicit, suspicious, or otherwise un-categorized but still within a grey-area of legality (e.g., Smart Drug Trafficking). Such illegal activities usually contain drug selling, counterfeit products, and child abuse content (Dalins et al., 2018). The majority of these illicit contents are sold through various Dark Web markets. Numerous Surface websites advertise these markets, providing the user with the onion link. Consequently, the Dark Web markets are gaining exponential popularity, endangering, in many cases, the unsuspected user who cannot identify the legality of each product or the truthfulness of the presented information. Lastly, the vulnerability in malicious and phishing code is deep in these markets, posing an additional threat to the everyday user.

In an attempt to shed light on the illicit activities on the Dark Web, the research community is either classifying images, text, or even the underlying code of the dark websites (Cascavilla et al., 2022a; Cascavilla et al., 2022b). Several studies implement machine learning algorithms and deep learning techniques for automatic taxonomy extraction and Deep

[1] https://www.torproject.org/

Cascavilla, G., Catolino, G., Conti, M., Mellios, D. and Tamburri, D.
When the Few Outweigh the Many: Illicit Content Recognition with Few-Shot Learning.
DOI: 10.5220/0012049400003555
In *Proceedings of the 20th International Conference on Security and Cryptography (SECRYPT 2023)*, pages 324-334
ISBN: 978-989-758-666-8; ISSN: 2184-7711

Copyright © 2023 by SCITEPRESS – Science and Technology Publications, Lda. Under CC license (CC BY-NC-ND 4.0)

and Dark Web content analysis. On the one hand, image categorization of the HTML pages in the various Dark Web content is researched in depth by (Hashemi and Hall, 2019). Specifically, authors in (Hashemi and Hall, 2019) identified and categorized dark propaganda based on visual content while using semantic segmentation with specifically designed filters. Finally, (Fidalgo et al., 2019), through specially designed masks and "bags of visual words," they classified illicit images from the Dark Web with high accuracy. On the other hand, the textual appearance is the main focus of the studies (Al Nabki et al., 2017) (Ghosh et al., 2017). In the latter research (Ghosh et al., 2017), they proposed an onion crawler to thematically categorize the content of Dark Web pages, e.g., drug-related, gun-related.

The above studies share one key element, the existence of significantly big datasets that are accurately labeled or, in the case of (Fidalgo et al., 2019), a dataset that can be considered "ideal". That is adequately cleaned images, lacking any noisy background that someone might encounter when scraping images from the Dark Web. Besides, the data used in the studies mentioned earlier are mostly well-balanced and categorized at a high level, which means that specific categories have not yet been investigated. However, "Reality is cruel," meaning relying on more data is not always possible. Law enforcement should have the possibility to react as soon as possible to detect illicit activities on the Dark Web using an approach that works with high accuracy even when the data collection is reduced. In the context of our research, we provided a novel approach for illicit image recognition, considering new Dark Web images–thus possibly implying small and noisy data. In particular, we investigated an alternative approach when handling small datasets using the ability of label agnostic learning techniques, i.e., One-Shot (Lake et al., 2011) and Few-Shot (Hilliard et al., 2018), when identifying illicit images, thus possibly improving the problem of handling unlabeled and few data. One/Few-shot learning requires fewer data to train a model, thus eliminating high data collection and labeling costs. Moreover, low training data means low dimensionality in the training dataset, which can significantly reduce computational costs. When new data are added, the model can recognize them without re-training. The Dark Web can benefit from these approaches since new illicit content images are arising daily, making their identification time-consuming and challenging. The approaches mentioned rely on using Siamese networks since it can be more robust to class imbalance and works well with images without losing their information. Moreover, it has not been studied on Dark Web content yet. Consequently, we formulated the following main research question:

RQ. ***To what extent can illicit Dark Web content be classified through a limited number of images?***

To answer our main research question, we need to address the following sub-questions:

SRQ1. *To what extent One-Shot technique using Siamese Neural Networks can identify illicit images from the Dark Web?*

SRQ2. *To what extent Few-Shot technique using Siamese Neural Networks can identify illicit images from the Dark Web?*

The goal is to investigate the ability of One-Shot and Few-Shot learning techniques to identify and separate illicit image embeddings using Siamese Neural Networks. We verify the results by evaluating the model's performance and focusing mainly on the accuracy metric.

The results of our study highlight that our approach peaked at 90.9 % testing accuracy on 943 unseen images of 10 different categories.

To sum up, the paper provides four key contributions:

1. A novel Dataset of Dark Web illicit contents consisted of 3750 images categorized in 55 different categories, e.g., drugs and weapons (Replication-Package, 2023).

2. A new approach that exploits the One-Shot Learning technique to identify illicit images from the Dark Web;

3. A new approach that exploits the Few-Shot One-Shot Learning technique to identify illicit images from the Dark Web;

4. An online available repository reporting the raw data in the context of the study for further research and new considerations by the community (Replication-Package, 2023).

The remainder of this paper is organized as follows. Section 3 provides an overview of the dataset and the related approach used to build, clean, and prepare it. Section 4 introduces and explains the methodology for classifying illicit images. In Section 5 are presented the results of our approach. Section 6 discusses the results of our research and the related limitations, while Section 8 draws the conclusions and sketches some possible future research.

2 RELATED WORK

Previous research on the Dark Web mainly focused on classifying the illicit activities in the Dark Market places based on their textual content. More specifically, (Al Nabki et al., 2017) created the well-known DUTA dataset, which consists of 5002 labeled Dark websites. Three supervised machine learning algorithms were tested: Support Vector Machines (SVM), Logistic Regression, and Naive Bayes. Using Term Frequency - Inverse Document Frequency (TF-IDF) and Bag Of Words (BOW) dictionaries tuned explicitly for their dataset, they achieved high accuracy when predicting illicit content. Similarly, (Ghosh et al., 2017) created an onion crawler to thematically categorize the content of Dark Web pages as drug-related, gun-related, etc., based on specific keywords. Authors in (Choshen et al., 2019), while following a similar approach, enriched their experiments with data originating from eBay product pages as well as Legal Onion websites in an attempt to identify the legal and illegal language used in the Dark Web. Lastly, (Ranade et al., 2018) collected data from the Twitter streaming API to generate a multilingual corpus based on keywords such as DDoS attacks, DNS, spam, malware, etc. The collected data was fed to a translating algorithm designed by the researchers, which achieved 97% semantic relevance compared to Google's translated output upon expert evaluation.

Even though the textual representations of the Dark Marketplaces are thoroughly investigated, more extensive research should be conducted on the images originating from these markets. One of the most influential studies regarding HTML classification based on the visual contents of Dark websites is (Hashemi and Hall, 2019). The researchers in (Hashemi and Hall, 2019) are identifying and categorizing dark propaganda based on the visual content of the investigated websites. They trained the well-known Convolutional Neural Network (CNN) Alex-Net on 120,000 images obtained from the Dark Web and finally tested on 1.2 million suspicious images concluding with an accuracy of 86%. On the other hand, the researchers in (Fidalgo et al., 2018) created a dataset (TOIC) of almost 700 images scraped from the Dark Web. They generated dictionaries representing this database by implementing K-Means and Nearest Neighbour algorithms. Edge-Shifting dense techniques were tested on a different radius, resulting in an 85.6% overall accuracy. Inspired by the promising results, the authors in (Fidalgo et al., 2019) introduced specifically designed masks, and through a similar "bag of visual words" BoVW classified illicit images. The accuracy of the pre-trained model when tested on the researchers' dataset TOIC while using BoVW reaches approximately 88%.

Label-agnostic techniques, such as One-Shot and Few-Shot, learn from the pixels of each image using the Siamese Networks produced embeddings. Therefore, re-training is optional. One of the main differences between One-Shot and Few-Shot Learning techniques is the volume of the input data, which means that the sample of data is used to classify the embeddings produced by the Siamese Networks. In particular, the model is trained on a few images (Li et al., 2017) (Wang et al., 2019), or one image per category (Shaban et al., 2017) (Vinyals et al., 2016). In (Chopra et al., 2005) are testing the ability of Few-Shot learning implementing Siamese Neural Networks on the AT&T dataset and the AR database of faces. The datasets, in combination, contain approximately 4000 images of faces photo-shoot in a period of 14 days. Their proposed networks recognize employee faces with an 80% accuracy. Siamese Neural Networks is one of the most popular choices for label-agnostic tasks. Its objective is to use twin embedding nets and generate representing vectors for each picture which are compared by calculating their euclidean distance. Studies like in (Schroff et al., 2015; Fei-Fei et al., 2006; Lake et al., 2011; Koch et al., 2015) used the Siamese networks' architecture, obtaining high accuracy in different domains. All the above studies share one key element, the existence of significantly big datasets that are accurately labeled or, in the case of (Fidalgo et al., 2019), a dataset that can be considered "ideal", where images are cleaned and lacked from noisy. Also, the data used are mostly well-balanced and categorized at a high level, which means that specific "in-depth" categories have not been investigated yet. Therefore, we investigated an alternative approach when handling small datasets using the ability of label agnostic learning techniques, i.e., One-Shot (Lake et al., 2011) and Few-Shot (Hilliard et al., 2018), when identifying illicit images, thus possibly improving the problem of handling unlabeled and few data. To the best of our knowledge, no previous research investigated the ability of label-agnostic techniques for illicit image recognition, which is the focus of our work.

3 DATASET OVERVIEW AND DATA ENGINEERING

This section reports the steps followed to extract new images from the Dark Web and create our datasets (Replication-Package, 2023).

3.1 Data Scraping - Collection

To collect data, we implemented a crawler using the Selenium Python library capable of automating steps to download HTML pages from the Dark Web using the Tor browser.

Since login into the website was mandatory to extract any information, we needed to deal with the security Captchas using Captcha-solving API. The script captured a screenshot of the website's login page, which was sent to the external server. After locating the input box, the resulting password was automatically typed into the appropriate field.

After logged in, the script crawled through the different product ads and collected the URLs of the images. Initially, the objective was to download the images of the products immediately after redirecting to the product page. However, this technique was identified as an attack and blocked. Hence, we built a list with external links of all the product images accompanied by the category these images belonged to. Lastly, a different script bypassed, in a similar manner, the security of the website and randomly downloaded the images from the servers, avoiding triggering any alarms. The data used in this research have been scraped in a period between January and March 2020 from various Dark Markets. More specifically, we scraped three popular Dark Web marketplaces Silk Road, BitBazaar, and Dark Market, resulting in 5500 images depicting drugs of all categories, credit cards, ID cards (IDs), and gift cards. Although the markets above broadly related to drugs, the sample of personal IDs and credit cards was relatively small, while the sample of passports was less than five images. These Dark markets also lacked images of weaponry, so we scraped additional random onion sites resulting in 210 high-quality images of guns and semi-automatic guns, 215 ID cards, 51 images of passports, and 118 additional credit cards. The dataset with all the data is currently stored in an encrypted hard drive and available under request. However, it is worth highlighting that the authors did not buy any item advertised in the markets cited above. All the images are publicly available from the crawled dark marketplaces as product advertisements.

3.2 Data Cleaning and Preparation

To prepare our data, we performed common steps like cleaning and removing duplicates. In particular, we tested for identical duplicates through hashing techniques[2], resulting in about 2000 matching images. For this reason, we removed them from the dataset. In the online appendix (Appendix, 2023) Figure 4 shows the distribution of the dataset.

Finally, based on the availability problem described in 3.1, we dealt with merging, removing, or relabeling specific sub-categories (the removed categories are marked with a red dot in Figure 4 (in online appendix (Appendix, 2023)), Moreover, we created a new category of counterfeit including passports, IDs, money bills, credit cards, gift cards, and documents. All the drug-related categories kept their initial labels. All these steps concluded in a dataset of 3570 images and 55 different classes.

3.3 Data Augmentation

Even after the various cleaning steps and precise categorization, the final dataset results unbalance. In the context of our paper, we experiment with our approach using One-Shot (Fei-Fei et al., 2006) and Few-Shot (Wang et al., 2019). Since previous studies ((O'Mahony et al., 2019; Ochal et al., 2021)) advise using them on balance data–with an identical sample of images for each category–avoiding a poor representation of specific categories, we performed data augmentation to balance the minority categories, e.g., type of drugs. In particular, the script calculates the final size after the possible augmentation steps and aids the user in a better sample decision. The code augments each image six times and saves it for later use while randomly removing excess images from the more significant categories to balance them with the remaining ones. The volume of images that needed to be removed was calculated based on the size of the smallest size category, and the images were deleted from each category. Pseudocode is available in Algorithm 1.

Algorithm 1: Augmentations.

```
Data: Images from categories $i$
Result: Augmented categories
$remove\_excess\_i \leftarrow 0$
$smallest\_category \leftarrow min(cat\_1, cat\_2, cat\_3)$
$smallest\_category \leftarrow smallest\_category * 6$ augmentations
for $i$ in $\#categories$ do
    $cat\_i \leftarrow cat\_i * 6$ augmentations
    $remove\_excess\_i \leftarrow cat\_i - smallest\_category$
    $cat\_i \leftarrow cat\_i - remove\_excess\_i$
```

The augmentations steps are: rotation by 30 degrees, horizontal flip, vertical flip, cropping by 30%-45%, change of contrast's gamma by 2.0 - 3.0, and

[2] Image Hashing - Python Documentation: https://pypi.org/project/ImageHash/

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

addition of Gaussian noise. The augmentations are illustrated in Fig. 1. We generated the augmentations using Imgaug Augmenters[3].

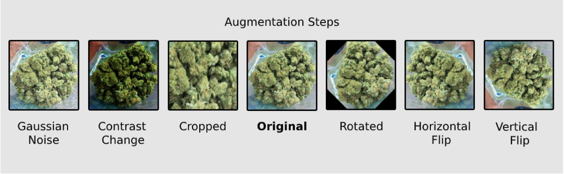


Figure 1: The 6 different augmentations implemented to this image from left to right are: Gaussian Noise of density 30, Contrast Change (gamma = 2.4), Zoom-Crop by 30%, Rotation by 40 degrees, Horizontal flip, and Vertical Flip.

4 RESEARCH METHODOLOGY

This section explains the methodology implied to classify illicit images and carry out our experimental evaluation.

4.1 Experimental Setup

Figure 2 reports the experimental research pipeline followed in our study. The first step regards using the scraping tool that accesses the Dark Web through the Tor browser to scrape Onion websites. The images were randomly downloaded and stored locally in the appropriate folders based on their category, we removed duplicates, and some images were relabeled manually for better representation. We evaluated the most popular dimensions of images in the dataset and informed the user appropriately. We augmented the images and balanced the classes based on the users' dictation. Finally, we trained and evaluated the Siamese neural network using One-Shot and Few-Shot learning.

We implemented our script and model using Python using libraries like Selenium (Gojare et al., 2015), Glob[4], Shutil[5], Image PIL[6], ImGaug Augmenters[7], Tensorflow[8], Sklearn[9], and TSNE (Lin et al., 2017).

4.2 One-Shot and Few-Shot Learning

The One-Shot and Few-Shot techniques merely differ in the volume of the input data used for creating the embeddings and testing the models. In other words, one or a few images are used for each category for the One-Shot and the Few-Shot experiments. The number of categories is represented by k, hence, k-way datasets and N-shot where N is the number of images in each category.

4.3 Pair Generation for Siamese Networks

Before we move on to the models, it is essential to explain the pair generation procedure in detail. K-pairs of N images must be generated to test the Siamese Neural Networks on K-Shot experiments. Hence the higher the sample of data in each category, the more pairs can be generated. In particular, each image is randomly paired to 1,2,5(etc.) images, as described in (Varior et al., 2016) (Koch et al., 2015) (Qiao et al., 2017), and a binary label is assigned to the pair. If both images of the generated pair belong to the same category, the label is 1; otherwise is 0. However, randomly created pairs generally produce an imbalanced representation of the positive (1) and negative (0) pairs. For example, if there are 10 images in 10 different categories, and we choose one image from the first category with the goal being to find another image from the same category, the probability of achieving that is $\frac{9}{99}$ or 0.0909. That number is only decreasing $\frac{9}{99} * \frac{8}{98} * \frac{7}{97} * ..$ when we try to randomly select more images from the same category for a 5-Shot approach. Therefore, the positive pairs are disproportionately less than the negative.

In this paper, the generated pairs maintain the same number of negative and positive images similarly designed to (Shaban et al., 2017), even though the pairs were generated randomly. For each experiment, we first created the positive pairs, followed by the same number of negative ones, reassuring an accurate representation of the two labels [1,0] and eliminating any label bias. Based on this study's testing, the final accuracy can fluctuate drastically if the preparation of the pairs is not designed correctly. Meaning that the model will search for the easiest solution to produce the highest accuracy. That is, the output embeddings are always far away from the compared ones, and the model tends to predict a label 0 on every set because it cannot penalize the mistakes adequately. A solution to this issue follows the pair creation of (Varior et al., 2016). Instead of searching for an image from the same category, the authors

[3] Documentation and Examples: https://imgaug.readthedocs.io/en/latest/source/overview/arithmetic.html

[4] Glob Documentation: https://docs.python.org/3/library/glob.html

[5] Shutil Documentation: https://docs.python.org/3/library/shutil.html

[6] Image PIL: https://pillow.readthedocs.io/en/stable/reference/Image.html

[7] Augmentation Library: https://imgaug.readthedocs.io/en/latest/

[8] Tensorflow: https://www.tensorflow.org

[9] Sklearn: https://scikit-learn.org/stable/

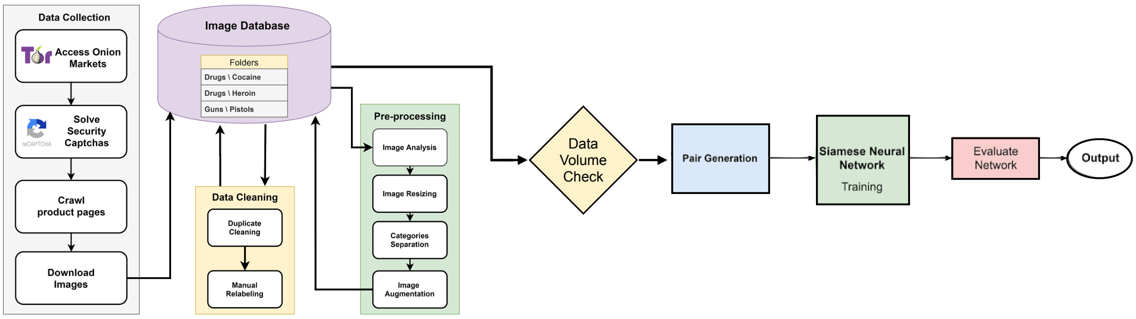


Figure 2: Pipeline of the study.

generated the positive pair by augmenting the initially chosen picture. Nonetheless, in this paper, the chosen images are always different from the comparing one because the augmentation is applied manually and in an earlier stage, as described in Section 3.3. This way, the compared embedding is rarely very close, constituting even harder One-Shot and Few-Shot tasks.

4.4 Siamese Neural Network Architecture

We created and tested different Siamese Neural Networks to identify the optimal number of hidden layers needed and the activation functions for the hidden and output layers. Also, we varied the last fully connected (Dense) layer and the in-between filter sizes to create a model able to extract the embedding of each input image pair as accurately as possible. The difficulty of the task highly correlates with the depth of the neural network. Our proposed model does not strictly follow any of the networks proposed in (Lake et al., 2011; Hilliard et al., 2018; Koch et al., 2015), but it is highly inspired by several related studies (Qiao et al., 2017; Li et al., 2017; Vinyals et al., 2016). We tuned the final architecture of the convolutional embedding neural networks on the datasets created. The proposed structure of the convolutional embedding neural network consists of six convolutional 2D layers. The first two layers have a filter size of 3 by 3, and the remaining of 2 by 2. The layers dimensions are increasing gradually, starting from 50 by 50 up to 220 by 220.

Between the convolutional 2D layers, we applied a max-pooling of (3x3) and (2x2) as depicted in (Varior et al., 2016). The filters of the hidden convolutional layers were tested with various sizes to eliminate overfitting effects. Additionally, we applied three lasso regularizations to the fully connected layer (512); a kernel regularization of 0.001, a bias regularization, and an activity regularization. The last Dense layer of the Embedding Network has a ReLu activation function which, compared to a linear one, produced more stable results. Lastly, the model implements the RMSprop optimizer with a 0.0001 learning rate and a decay of 0.7. The final structure of the Siamese network consists of two identical embedding neural networks and is illustrated in Fig. 5 (in online appendix (Appendix, 2023)).

Each embedding network is fed with one of the generated pairs' images. The twin embedding networks (Fig. 6 in the online appendix (Appendix, 2023)) are fully connected to the final dense layer that outputs either 1 if the pair is originated from the same category or 0 if it is from a different one. The Siamese Neural network predicts the label by calculating the Euclidean Distance between the two embeddings, as depicted in (Melekhov et al., 2016) and (Varior et al., 2016). Lastly, the network calculates the loss per pair of images via the Contrastive loss formula, as depicted in (Hadsell et al., 2006). In this research, similarly to previous studies, K-shot tests were performed. Therefore, we generated 1 pair and 5 pairs per image. The Siamese neural network is tested on 1-Shot and 5-Shot learning approaches with various samples of illicit pictures for each category. Besides, the networks were tested with more and fewer categories for each K-shot technique.

5 RESULTS

We performed two types of tests to evaluate the ability of the Siamese neural network regarding proper embedding creation and accurate separation of them. The models are tested on three different buckets of data. In particular, we experimented with our models considering the shape and the actual type of illicit content images. Following the experimentation techniques of (Garcia and Bruna, 2017) and (Vinyals et al., 2016), the tests are performed on gradually increased datasets. In particular, the first bucket consists of 55 categories (55-way) of illicit images; each category is represented by just one image. The sec-

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

ond and third buckets use the same dataset, but each category includes 5 and 20 images, respectively. Additionally, the number of classes in the buckets above was randomly reduced to identify the model's ability to separate a higher variety of embeddings. Hence, the model is also trained on 10, and 25 randomly selected categories. The tests were performed with 943 randomly chosen entirely new images, and the models were trained for 100 epochs.

Table 1 illustrates the accuracy of the various tests performed. Specifically, each model is tested on three category volumes, 10-way, 25-way and 55-way and for 1-Shot, 5-Shot, and 20-Shot tests. Looking at the table, the model was fed with 1 image per category, then 5 images per category, and finally 20 images. As expected, the model performs better when tested with a higher N-Shot since there are more trainable examples per category. Generally, 1-Shot tests were prone to overfitting, whereas the overfitting effect was drastically reduced in the 5-Shot and 20-Shot tests.

We can claim that the Siamese Neural Network results in higher training and validation accuracy when more data are present. In addition, the testing accuracy depicts an increase of almost 30% (from 70.1% to 99.9%) if 20 images are present (20-shot) in 10 classes (10-way), compared to just one image in each one of them. The same pattern is visible throughout the different category sizes, with approximately 20% (from 66.8% to 86.7%) increase in the 25 classes (25-way) test and 14.8% (from 71.4% to 86.2%) in the 55 classes test. These results imply that the model can generalize better due to the increased size of the trainable examples. Additionally, the model performs better when the categories are reduced from 55 to 25 and 25 to 10, but the difference never exceeds 4.2%. Meaning that the model is not affected by the number of classes if the volume of images in each class does not exceed the above sizes. It is worth noticing that in the case of 1-shot, the model under-performs on testing when the number of classes is reduced from 55 to 25. That occurs because the classes are randomly split and reduced; therefore, some classes might be recognized with higher precision in the 55-way bucket.

The ROC curves of the N-Shot tests conducted on the 55-way bucket of data in Fig. 3 are in charge of further justifying the increase in generalization when more data are present in each category. Looking at sub-figures 3a and 3b, it is safe to conclude that when additional images are present in each of the classes, while training, the Siamese Network can identify more tested positive pairs/labels. The above is visible in the ROC curve area of the last sub-figure 3c, which is equal to 86%, approximately 10% higher compared to the first 3a.

Table 1: The accuracy of the various tests with 10-way, 25-way and 55-way dataset on 1-shot, 5-shot, and up to 20-shot respectively.

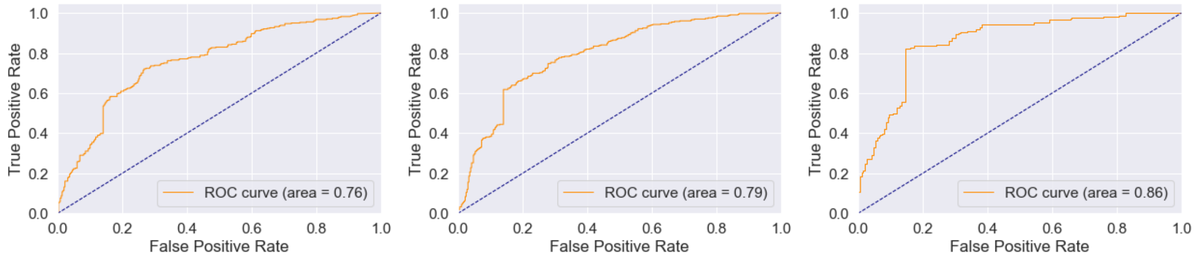
		1-Shot	5-Shot	20-Shot
10-way	Val Accuracy	98.9%	93.8%	96.4%
	Test Accuracy	70.1%	75.6%	**90.9%**
25-way	Val Accuracy	98.7%	92.7%	92.1%
	Test Accuracy	66.8%	76.2%	**86.7%**
55-way	Val Accuracy	97.6%	86.2%	87.7%
	Test Accuracy	71.4%	74.3%	**86.2%**

6 DISCUSSION

This section discusses the results and limitations of the study.

6.1 Research Questions

The main objective of this research was to investigate alternative approaches that can recognize illicit activities on the Dark Web. Specifically, this study aimed to bypass the burden of collecting supervised large-scale datasets using One-Shot and Few Shot learning. Therefore, the main question was related to the ability to detect illicit Dark Web content with a limited number of images. Our experiment showed promising results. Indeed, when considering the first subquestions related to using One-Shot learning techniques, we can claim that Siamese Neural Networks can recognize illicit images efficiently based on this study's experimentation. Indeed when testing the accuracy of the model, we reach a percentage around 70%. Moving the attention to Few-Shot learning techniques, i.e., SRQ2, the Siamese Neural Network presented promising generalization capabilities when the sample was increased by just four or up to 20 images per category. The testing accuracy reached 76.2% on the 5 images per category dataset with 25 categories and 90.9% on the 6 to 20 images per category dataset with ten categories. Lastly, it is worth noticing that the testing accuracy stayed under 86.2% regardless of the increased number of categories, 55, on the 6 to 20 images dataset. These techniques' usage can be promising compared to the previous study. In particular, Fidalgo et al. (Fidalgo et al., 2018) proposed an approach for detecting illicit contents on a small-scale dataset of approximately 700 images separated in 5 classes. The resulting accuracy in (Fidalgo et al., 2018) is 85.6% and in (Fidalgo et al., 2019) 87.98%. Our study outperforms the aforementioned by 3% with a 90.9% on 20-Shot tests. Finally, we



(a) ROC curve: 55-way on 1-shot (b) ROC curve: 55-way with 5-shot (c) ROC curve: 55-way on 20-shot

Figure 3: The ROC curves of the 5-Shot learning tests on 55-way with 1-shot, 5-shot, and up to 20-shot respectively.

scored similar results, 86.7%, with a dataset consisting of 25 classes, 20 classes more than the previous research. Our future agenda aims at comparing the methodology of their study.

6.2 Limitations

To the best of our knowledge, no previous study has investigated the above techniques on illicit content. Hence, a strait forward comparison of this study's results with previous studies was impossible. Finally, precise categorization and data cleaning were among the initial burdens. Random pair generation was the only possible solution, even though previous studies suggest a manual selection of them. Based on our study's results, higher precision of labeling yields superior accuracy. Although, due to the lack of expertise regarding illicit content labeling, the categories could not be further separated, and the pairs could not be generated manually. To avoid possible problems in the code, we developed our script and pipeline, we relied on stable Python Libraries.

7 SOCIETAL IMPACT

This study aims to investigate alternative approaches when handling small datasets. The expensive time procedure of collecting large-scale data (images) from Dark Web Markets, as well as the need for highly skilled personnel responsible for illicit content labeling, are some of the burdens this research is trying to bypass. We showed the ability of label-agnostic models handling unlabeled data to identify illicit images from the Dark Web. Law enforcement agencies can benefit from our suggested approach and could conduct faster investigations with fewer resources and capabilities. Moreover, the recognition speed of new illegal substances from the Dark Web represents a key factor in intercepting new illegal trends and persecuting illicit behaviors. Our proposed approach poses the basis for a less time-consuming system to assist law enforcement agencies during their activities.

8 CONCLUSION

This study presents a novel approach to recognizing illicit images from the Dark Web through a relatively small sample of images. We generated a new dataset consisting of 3570 images spreading over 55 sub-classes. Then, we investigated the Siamese neural network classification methods on One-Shot and Few-Shot experiments. Results show that Siamese network peaked at 90.9% testing accuracy on 943 unseen images of 10 different categories. To conclude, this study provided a new contribution to illicit image recognition through label-agnostic networks on One-Shot and Few-Shot experiments. These techniques could help law enforcement agencies effortlessly identify illicit activities in the Dark Web through small data samples. The future agenda includes the comparison of other different techniques and setting parameters.

REFERENCES

Al Nabki, M. W., Fidalgo, E., Alegre, E., and de Paz, I. (2017). Classifying illegal activities on tor network based on web textual contents. In *European Chapter of the Association for Computational Linguistics*, volume 1, pages 35–43.

Al-Nabki, M. W., Fidalgo, E., Alegre, E., and Fernández-Robles, L. (2019). Torank: Identifying the most influential suspicious domains in the tor network. *Expert Systems with Applications*, 123:212 – 226.

Appendix (2023). When the few outweigh the many: Illicit content recognition with few-shot learning https://doi.org/10.6084/m9.figshare.22726745.

Cascavilla, G., Catolino, G., Ebert, F., Tamburri, D., and van den Heuvel, W. (2022a). "When the Code becomes a Crime Scene" Towards Dark Web Threat Intelligence with Software Quality Metrics. In *2022*

SCIENCE AND TECHNOLOGY PUBLICATIONS

IEEE International Conference on Software Maintenance and Evolution (ICSME), pages 439–443. IEEE.

Cascavilla, G., Catolino, G., and Sangiovanni, M. (2022b). Illicit darkweb classification via natural-language processing: Classifying illicit content of webpages based on textual information. In *Proceedings of the 19th International Conference on Security and Cryptography - Volume 1: SECRYPT,*, pages 620–626. INSTICC, SciTePress.

Chopra, S., Hadsell, R., and LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 539–546.

Choshen, L., Eldad, D., Hershcovich, D., Sulem, E., and Abend, O. (2019). The language of legal and illegal activity on the Darknet. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4271–4279.

Dalins, J., Tyshetskiy, Y., Wilson, C., Carman, M. J., and Boudry, D. (2018). Laying foundations for effective machine learning in law enforcement. majura – a labelling schema for child exploitation materials. *Digital Investigation*, 26:40 – 54.

Fei-Fei, L., Fergus, R., and Perona, P. (2006). One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28:594–611.

Fidalgo, E., Alegre, E., Fernández-Robles, L., and González-Castro, V. (2019). Classifying suspicious content in tor darknet through semantic attention keypoint filtering. *Digital Investigation*, 30:12 – 22.

Fidalgo, E., Alegre, E., González-Castro, V., and Fernández-Robles, L. (2018). Illegal activity categorisation in darknet based on image classification using creic method. pages 600–609.

Garcia, V. and Bruna, J. (2017). Few-shot learning with graph neural networks.

Ghosh, S., Das, A., Porras, P., Yegneswaran, V., and Gehani, A. (2017). Automated categorization of onion sites for analyzing the darkweb ecosystem. pages 1793–1802.

Godawatte, K., Raza, M., Murtaza, M., and Saeed, A. (2019). Dark web along with the dark web marketing and surveillance. In *PDCAT*, pages 483–485. IEEE.

Gojare, S., Joshi, R., and Gaigaware, D. (2015). Analysis and design of selenium webdriver automation testing framework. *Procedia Computer Science*, 50:341 – 346. Big Data, Cloud and Computing Challenges.

Hadsell, R., Chopra, S., and LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

Hashemi, M. and Hall, M. (2019). Detecting and classifying online dark visual propaganda. *Image and Vision Computing*, 89:95 – 105.

Hilliard, N., Phillips, L., Howland, S., Yankov, A., Corley, C. D., and Hodas, N. O. (2018). Few-shot learning with metric-agnostic conditional embeddings.

Juan Sanchez, G. G. (2019). Who's afraid of the dark? hype versus reality on the dark web. https://www.recordedfuture.com/dark-web-reality.

Koch, G., Zemel, R., and Salakhutdinov, R. (2015). Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2.

Lake, B. M., Salakhutdinov, R., Gross, J., and Tenenbaum, J. B. (2011). One shot learning of simple visual concepts. *Cognitive Science*, 33.

Li, Z., Zhou, F., Chen, F., and Li, H. (2017). Meta-sgd: Learning to learn quickly for few shot learning. *CoRR*, abs/1707.09835.

Lin, X., Wang, H., Li, Z., Zhang, Y., Yuille, A., and Lee, T. S. (2017). Transfer of view-manifold learning to similarity perception of novel objects.

Melekhov, I., Kannala, J., and Rahtu, E. (2016). Siamese network features for image matching. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 378–383.

Ochal, M. et al. (2021). Class imbalance in few-shot learning.

O'Mahony, N., Campbell, S., Carvalho, A., Krpalkova, L., Hernandez, G. V., Harapanahalli, S., Riordan, D., and Walsh, J. (2019). One-shot learning for custom identification tasks; a review. *Procedia Manufacturing*, 38:186–193.

Qiao, S., Liu, C., et al. (2017). Few-shot image recognition by predicting parameters from activations.

Raghavan, S. and García-Molina, H. (2001). Crawling the hidden web. In *Proceedings of the 27th International Conference on Very Large Databases (VLDB 2001)*, pages 129–138.

Ranade, P., Mittal, S., Joshi, A., and Joshi, K. (2018). Using deep neural networks to translate multi-lingual threat intelligence.

Replication-Package (2023). When the few outweigh the many: Illicit content recognition with few-shot learning. https://doi.org/10.5281/zenodo.7657482.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

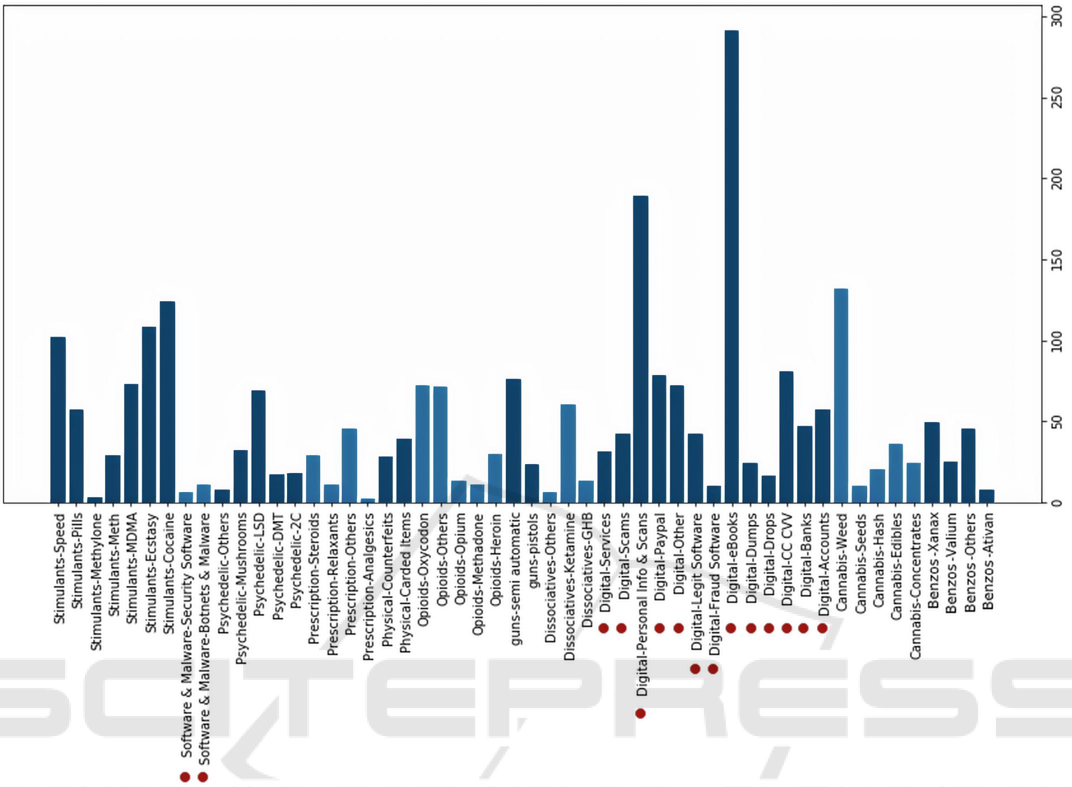
Shaban, A., Bansal, S., Liu, Z., Essa, I., and Boots, B. (2017). One-shot learning for semantic segmentation.

Varior, R. R., Haloi, M., and Wang, G. (2016). Gated siamese convolutional neural network architecture for human re-identification.

Vinyals, O., Blundell, C., Lillicrap, T. P., Kavukcuoglu, K., and Wierstra, D. (2016). Matching networks for one shot learning. In *NIPS*.

Wang, Y., Yao, Q., Kwok, J., and Ni, L. M. (2019). Generalizing from a few examples: A survey on few-shot learning.

APPENDIX



SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

Figure 4: The distribution of the initial dataset. The red dots represent the categories that are excluded from the final experiments.

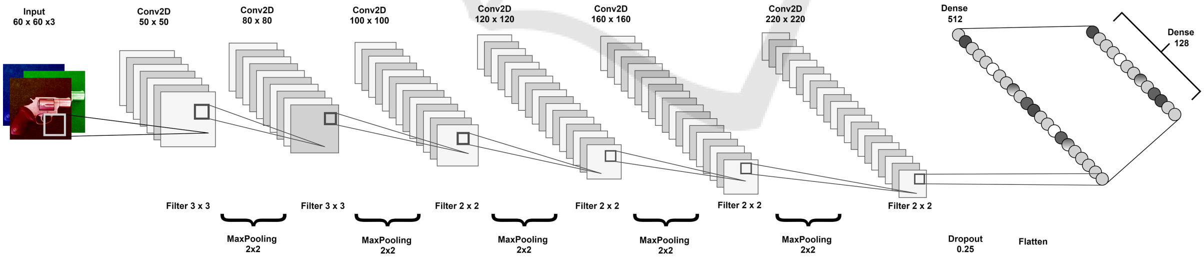


Figure 5: The proposed embedding convolutional neural network. A Siamese network consists of two identical embedding nets.

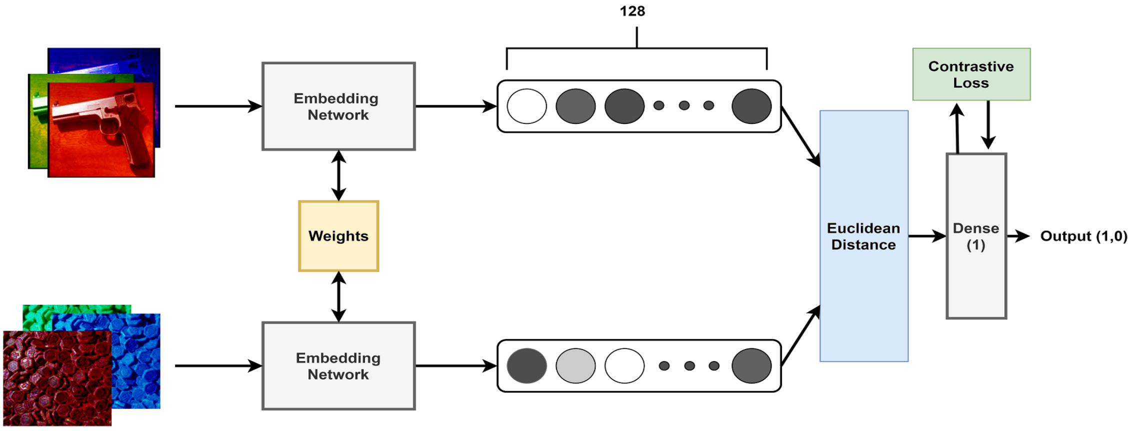


Figure 6: The twin embedding networks (Siamese Neural Network). The weights are shareable between the twin networks at the last fully connected layer. The output size for each embedding is 128. A fully connected layer is outputting 1 or 0 based on the error calculated from the Contrastive loss.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS