AI-powered Counterfeit Products Detection Solution for Consumer Brands

Anshul Garg[#1], Kumar Shubham[#2], Sanket Patil[#3], Karthik Bettadapura[#4]

[#]*Semantics Dept., Dataweave Pvt. Ltd., Singapore*

[1]anshul.garg@dataweave.com

[2]shubham@dataweave.com

[3]sanket@dataweave.com

[4]karthik@dataweave.com

***Abstract*— With the rise in e-commerce adoption by consumers worldwide, the sale of counterfeit products on open marketplaces is an increasingly challenging problem faced by shoppers, online retailers, and consumer brands. Recent surveys have reported that over a third of shoppers have had counterfeits delivered to them on placing an order online. These defective products can have a severe impact on the brand value of retailers and consumer brands, damaging consumer trust and potentially having an adverse impact on sales volumes. If one considers the large amount of products being sold over retailer websites, the problem of identifying counterfeits only magnifies.**

In this paper, we propose an efficient and robust method of detecting counterfeit products at scale using computer vision and deep learning applied on online product catalog data. The approach entails identifying small variations between two images.. Essentially, the original brand manufacturer's image on its own website is compared with images listed on online marketplaces by third party sellers to detect variations, which indicate whether a product listing is legitimate or not. We use pre-trained Convolutional Neural Network (CNN) based models to take advantage of transfer learning and further fine tune them on internal data to focus on fine grained image features. In addition, we use several image processing and matching techniques based on image signatures, key points, and descriptors to achieve better accuracy.

DataWeave's Counterfeit Products Detection solution is powered by datasets that we have built in house over the years, consisting of millions of products collected from thousands of retail websites across geographies. The dataset has hierarchical information pertaining to retail taxonomy.

Detecting counterfeits on e-commerce websites where the volume of data can be in the range of terabytes is a major challenge. Our technology platform efficiently stores data pertaining to millions of processed images in Internet Archive's ARC files and maintains the indexes in lucene search engines. The entire pipeline is automated and connected via big data technologies like Kafka and orchestrated across servers using Celery.

***Keywords*— Counterfeit detection, CNNs, Transfer learning, Kafka, Machine Learning, Image signatures, Solr**

I. Introduction

Open marketplaces online are filled with a variety of counterfeit products, ranging from adulterated versions of cosmetics and health products [1], and fake versions of products for brands [2]. The severity of the problem is plainly visible when you dig deep into the numbers. According to a report, over 70% of the products sold by third party sellers on Amazon.com are fake [3]. About one-third of shoppers have had counterfeit products delivered to them [4].

Fake products can have a significant impact on businesses [5] in the following ways :-

- Undermines and discourages investments
- Biggest unfair competitor
- Erodes consumer confidence
- Inhibits growth of enterprises
- Companies incur additional costs for conducting investigations and litigation to protect their IPR against infringement

Retailers like Amazon are trying to curb the problem but with little success [6]. Consumer brands like Ralph Lauren have started to address it as well [12] but, there is still lack of proper technology and solutions in the market to address this problem effectively.

In this paper, we propose a novel approach to identify fake products, unauthorized white labeling, and image theft across retailer websites and will also talk about how we scale our solution to millions of products to generate actionable output in real time.

In section II we will discuss general challenges and trickery used in counterfeiting. Section III is about the specific machine learning algorithms, image processing and text processing algorithms used to

solve this problem. Section IV talks about the scalability and infrastructure related aspects. We conclude our work in section V.

II. CHALLENGES

The type of counterfeiting varies with seller and category type. In case of categories like cosmetics, medical products, the products are often adulterated, while for products in electronics and accessories, poor quality products, with little consideration for safety standards, are sold.

To identify all these cases, brands need to keep a proper track of all the third party sellers selling their products. There are cases where sellers modify the textual information completely to protect it from backtracking but use the same product images or slightly modified images.

Within the image itself we have categorized different types of variations:

- color variation.
- logo variation.
- image modification with additional objects.

We have found cases where third party sellers, just to make their product unique in the Amazon catalog perform some preprocessing over the thumbnail to make it look unique. These basically include color changes with some additional modifications over the image (Fig.1),

Fig.1 Same product sold with color modification and additional changes.

variation or removal of certified trademark from the image to sell it under a different brand's name (Fig.2), and selling the same product at more than 3x the price range by using the same brand's name. Just to make their product stand out, we have seen cases where they add or remove unnecessary objects within the thumbnail to make it unique for their use cases and also to beat multiple algorithms built to identify them. Fig.3 .

Fig.2 Product sold with logo variation

The problem with the current solutions in the industry is the scale and variation among the data. Different verticals have different complexities and one of the most challenging tasks is to make a generic system that can be used across categories and verticals.

Fig.3. addition or removal of person in image define counterfeit.

Let's consider the case of canvas art. Most of the time, products of canvas art share a similar studio image in the background. As a result, more than 60% of the overall image remains the same even for two completely different examples. Fig. 4. In the case of rings, images are too small to process easily.

Fig.4. product with similar background

We have proposed a solution which uses CNN features along with local image level features to identify similar products and then use the unstructured product details to decide whether a product listing is legitimate or not.

III. Proposed Approach

There are two major challenges which is associated in finding counterfeit products:

- Scaling the complex computational intensive algorithms to process hundreds of millions of data points
- Matching images based on fine grained features.

While dealing with products over such a big scale, there is always a concern about the time complexity, especially in case of machine learning and deep learning based systems. These algorithms are known for their complex matrix computations and take significant amount of time. Just to give a glimpse about the importance of this problem, consider a scenario where a normal image processing algorithm takes 0.01s for feature computation and comparison of a standard image. A O(n) computation across all products over a database of 100 million products to find a counterfeit will take 100 million X 0.01 sec. close to 1 million sec or approximately 12 days. For businesses, such time scales are impractical. Although there are multiple solutions and infrastructure to handle these cases, which are known to give good speed within a certain margin of error, such systems may not be suitable for this particular case as finding counterfeits require a more controlled approach over the algorithms.

To solve the given problem, we first reduced our search space by choosing a small bucket [13] of relevant products based on image signatures [14], a string of numbers depicting multiple image properties like neighboring pixel brightness changes, gradient changes across different kernels and hash based encoding of final feature layers [18] of CNN architecture like AlexNet [19], Inception [17], ResNet [16], etc. These image signatures are then indexed within a lucene based search engine to generate a bucket of relevant products in near O(1) time.

The next important task is to find counterfeits based on fine grained image features for which we use the bucket of promising candidates generated by the hash code to do some deeper algorithmic computation. The possible candidates are re-ranked based on multiple metric functions like chi-square [20], cosine similarity generated over CNN and other laplacian based features of seed products [15] (product for which we want to find counterfeits), and individual candidates. Once a product satisfies the necessary condition over individual metric scores, it is considered as a possible counterfeit product and sent for verification to our internal Quality Assurance team to achieve higher accuracy in our results. The feedback for a given product is used for fine tuning the output and also as a new instance of training data for further model improvement.

IV. Scalability and Infrastructure

One of the key challenges of applying algorithms to real world scenarios is the sheer scale of data. A system taking too much time or memory can prove to be completely useless while solving a business case. It becomes necessary to give equal weightage to the scalability of the system and the algorithm when you deal with such data, which can scale to millions or billions of data points.

At DataWeave, we have a platform that scales on demand to aggregate data from different retailers and consumer brands across geographies. All these data are normalized into a standard format to handle language variations and other noises.

We use the web ARChive [10] file format to store all the relevant text as well as the associated images across different machines. Information stored about each product is accessible via an API.

Once data is available for further processing, it is pushed to a Kafka [7] pipeline which has multiple consumers specifically designed to process data stored in WARC format. The consumers store the processed data in Solr [8]. We have a novel method of storing representations (features) wherein all these data along with associated deep learning features and image low bit signatures are stored separately in a distributed Solr [11] platform, which not only acts as a storage platform, but enables us to search over the meta data and other textual information associated with them. Fig. 5. depicts the storage architecture for storing images and associated features in solr. Every step is logged through an EFK (elastic + FluentD + Kibana) pipeline [9].

To check for the counterfeit of a product, a job is pushed across a distributed compute cluster using Celery. based on the product type and category of a product, necessary steps of processing and score generation is performed. The storage and compute architecture is horizontally scalable, which is achieved by adding more machines to improve the data storage or speed of processing.

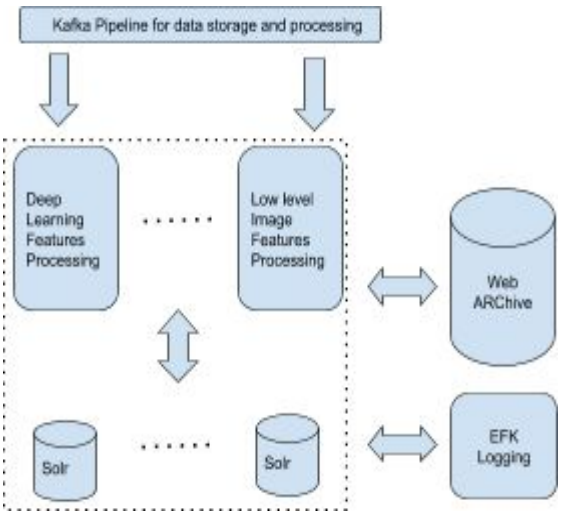

Fig.5. Storage of data and associated features in a distributed fashion

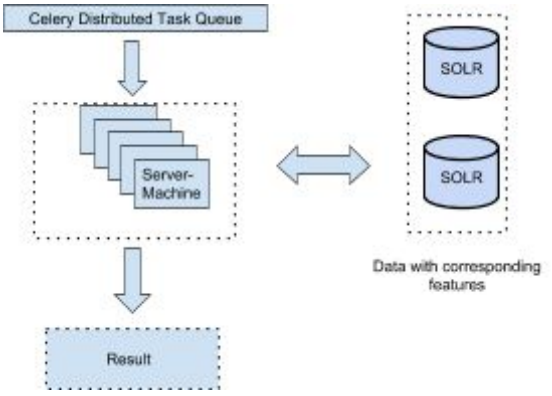

Fig.6. Counterfeit job scheduling over indexed data

V. Experimental Setup

We have used in-house datasets consisting of millions of products along with images collected from different geographies and from large retailer websites. The dataset has hierarchical information pertaining to retail taxonomy. At the root level, there is information such as category and subcategory, and at the leaf level, product details such as title, description and other <*attribute, value*> relationships. We have collected this data across various retail categories like apparel, furniture, electronics etc. We use our Quality Assurance team to annotate necessary information and label datasets using in-house annotation tools.

The process of indexing these datasets in our databases include the computation of several image features and storing them efficiently for swift access. Then, we collected product images and

descriptions for a big Home and Furniture Accessories brand. We detected the counterfeit products of this brand being sold at Amazon.com.

VI. Results

In our experiment, we considered 1,488 products of the aforementioned brand which had a total of 10,722 images. We searched counterfeits of these products among 233,105 Amazon.com products which in turn had 1 million images.

Of these 1,488 products, we found counterfeits for 515 at a precision of 86%. On further analysis, we found that 145 of these counterfeit products were sold under different brand names and the remaining 370 products were victims of image theft.

Fig.7(a). Original and their Counterfeit product

Fig.7(b). Original and their Counterfeit product

Fig.7(c). Original and their Counterfeit product

VII. Conclusion

In this work, we propose an efficient and robust method to detect image theft and counterfeit products on eCommerce marketplaces. We combine techniques from image processing and image hashing, and train a CNN model. We evaluated our model against product images of a manufacturer who makes furniture covers, boat covers etc. We found that nearly 35% of these products were counterfeits. Some of these counterfeit products were difficult even for humans to detect as shown in Fig 7. In this era of online shopping, duplicate products are a huge risk for the brand as well as the consumer.

Multiple types of Counterfeit products have been found to exist including exact copy and image theft. Brands are forged using advanced image processing tools. Fraudsters frequently update counterfeit product images. Proposed Counterfeit detector is designed to detect these kind of fraudulent image changes at a large scale.

VIII. Future Work

Our goal is to test this system against various product types especially Apparel and Footwear category. In general, these categories suffered more from fake products.

Also, we will be exploring the ways to reduce the dimension of our feature vectors to enhance storage efficiency and fast execution. In addition to that, we will extend our work to use neural net based approaches for learning patterns from text descriptions to detect the fraud.

IX. References

1. Booth, Stephanie. “We Love Amazon, but Here Are 6 Health and Beauty Products You Should Buy Elsewhere”. health.com. 26 May 2017. (https://www.health.com/mind-body/amazon-buy-counterfeit-dangerous-makeup-sunscreen)

2. Liao, shannon. “Amazon still sells counterfeit goods despite efforts to clean up”. theverge.com. 30 Apr 2018. (https://www.theverge.com/2018/4/30/17301714/amazon-counterfeit-goods-crackdown)

3. Todd, “Does Amazon Sell Fake Products? You Better Belive it!”. learntogrowwealthonline.com . April 2018. (https://learntogrowwealthonline.com/does-amazon-sell-fake-products-2/)

4. Variyar, Mugdha. “A third of ecommerce buyers get counterfeit products”. economictimes.indiatimes.com. 24 Apr 2018. (https://economictimes.indiatimes.com/small-biz/startups/newsbuzz/a-third-of-ecommerce-buyers-get-counterfeit-products/articleshow/63889378.cms)

5. Anurag, “negative effects counterfeiting brands”. tradevigil.com. 11 July 2017. (https://www.tradevigil.com/negative-effects-counterfeiting-brands/)

6. Hern, Alex. “Amazon site awash with counterfeit goods despite crackdown”. theguardian.com. 27 Apr 2018. (https://www.theguardian.com/technology/2018/apr/27/amazon-site-awash-with-counterfeit-goods-despite-crackdown)

7. Garg, Nishant. Apache Kafka. Packt Publishing Ltd, 2013.

8. Smiley, David, and David Eric Pugh. Apache Solr 3 Enterprise Search Server. Packt Publishing Ltd, 2011.

9. Yang, Kaichuang. "Aggregated Containerized Logging Solution with Fluentd, Elasticsearch and Kibana." International Journal of Computer Applications 150.3 (2016).

10. Gomes, Daniel, João Miranda, and Miguel Costa. "A survey on web archiving initiatives." International Conference on Theory and Practice of Digital Libraries. Springer, Berlin, Heidelberg, 2011

11. shubham, kumar. “Video: Using Product Images to Achieve Over 90% Accuracy in Matching E-Commerce Products”. medium.com/dataweave/. 9 Aug 2018. (https://medium.com/dataweave/-e5eb8934704a)

12. James W. Gentry, Sanjay Putrevu, Clifford Shultz II, and Suraj Commuri (2001) ,"How Now Ralph Lauren? the Separation of Brand and Product in a Counterfeit Culture", in NA - Advances in Consumer Research Volume 28, eds. Mary C. Gilly and Joan Meyers-Levy, Valdosta, GA : Association for Consumer Research, Pages: 258-265.

13. Jing, Y., Liu, D., Kislyuk, D., Zhai, A., Xu, J., Donahue, J. and Tavel, S., 2015, August. Visual search at pinterest. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1889-1898). ACM.

14. Xie, Liang, et al. "Cross-modal self-taught hashing for large-scale image retrieval." Signal Processing 124 (2016): 81-92.

15. Kong, Hui, Hatice Cinar Akakin, and Sanjay E. Sarma. "A generalized Laplacian of Gaussian filter for blob detection and its applications." IEEE transactions on cybernetics 43.6 (2013): 1719-1733.

16. Tai, Ying, Jian Yang, and Xiaoming Liu. "Image super-resolution via deep recursive residual network." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Vol. 1. No. 2. 2017.

17. Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

18. Zhang, Ruimao, et al. "Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification." IEEE Transactions on Image

Processing 24.12 (2015): 4766-4779.

19. A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012

20. Lancaster, Henry Oliver, and Eugene Seneta. "Chi-square distribution." Encyclopedia of biostatistics 2 (2005).