Using Text and Visual Cues for Fine-Grained Classification

Zaryab Shaker
School of Computer Science and Engineering
Xian Technological University
Xian, China
E-mail: zaryabkhan0346@gmail.com

Muhammad Adeel Ahmed Tahir
School of Computer Science and Engineering
Xian Technological University
Xian, China
E-mail: adikhan0313@gmail.com

Feng Xiao
School of Computer Science and Engineering
Xian Technological University
Xian, China
E-mail: xffriends@163.com

***Abstract*—Text is an important invention of humanity, which plays a key role in human life, so far from dark ages. Text in image is closely related to the scene or a product and is widely used in vision based application. In this paper we are addressing the problem of visual understanding with text. The main focus is combining textual cues and visual cues in deep neural network. First the text is recognized and classified from the image. Then we combine the attended word embedding and visual feature vector which are then optimized by CNN for Fine-grained image classification. We carried out the experiments on soft drink dataset in Pakistan. The results shows that the system achieves significant performance which can be potentially beneficial for real world application e.g. product search.**

Keywords-Scene Text; Product Text; Fine-Grained Classification; Convolution Neural Network; Attention; Product Search

I. INTRODUCTION

Fine-Grained image classification [1, 28] is a real-world emerging problem and it has received great attention from research communities around the globe. In computer vision, fine grained [1] image involves the problem of assigning images to classes where different instances of different classes differ slightly in their appearances e.g., flower species, animal species, product/place types. Therefore, fine-grained image classification [28] is a challenging assignment due to the slight variations among highly-confused categories of instances belonging to various classes of objects, which are hard to distinguish. Further, in some of the cases, human intervention or specific knowledge of a particular domain is also required to perform precise fine-grained image classification [1].

For some years, fine-grained image classification [1] is also being applied for natural scene classification. This involves the natural images of wide and diverse nature. It is witnessed that classification of shops, variety of products in shops, etc. are the considered areas where fine-grained image classification is being used [4]. Using fine-grained image classification [28] techniques on the soft drink dataset is an area that has received limited attention from the researchers, whereas this area has also exciting applications in restaurants and shops, where automated orders could be places once a specific brand of soft drink is going to be out of stock.

In this work, we have exploited text and visual cues in form of features to be used with Convolutional Neural Network for attaining good performance in the fine-grained classification [1, 28] of soft drinks. To the best of our knowledge, it

DOI: 10.21307/ijanmc-2021-026

is a unique application of a classification technique in the domain of fine-grained image classification of soft drink datasets.

The second section of this paper discusses the proposed classification technique, the third section involves results and discussion and conclusions are drawn in the last section.

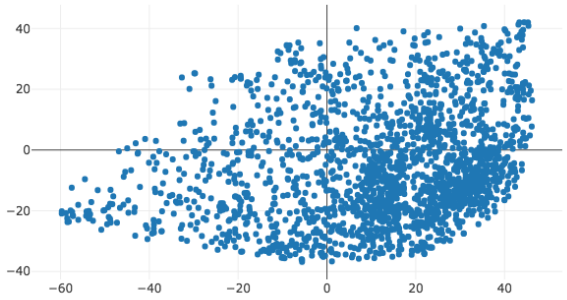

Figure 1. T-SNE Plot for word Embedding

II. Literature Survey

A. Text detection and recognition

Text in images carries a high level of information which makes this property of images very rich in computer vision as well as for humans. The information encoded in the text can be very beneficial for many computer vision applications. Text detection and recognition [7] face some challenges like the diversity of nature, the complexity of background, interference factors, etc. The first novel approach of a real end-to-end model for text detection and recognition in a scene was proposed in 2010 by Neumann et al [8]. It achieved a highly significant increase in recognition rate from 53% to 72% on the Char74k dataset (de Campos et al.) but there is a weakness in this system is that it was only applicable to horizontal texts. Later on Coates et al. [9] 2011 apply large-scale algorithms to build highly effective classifying model for both detection and recognition. Their end-to-end system has high accuracy and performs well on complex natural images but the drawback is that it requires a relatively large volume of training data.

Further in 2014 Jaderberg et al [10] addresses the problem of text detection and recognition by generating text proposals with CNN and provides an end-to-end system for reading text in natural scene images. That system was capable of both text spotting and image retrieval and perform excellently on complex natural images. Yao et al [11] present a unified framework for detecting and recognizing the text in images to handle the texts of different orientations. They also provide a method for 'search dictionary' to correct the recognition errors. The system achieves highly competitive performance, especially on multi oriented texts. Also, the works of Shiva kumara et al. [12] and C Yao et al. [13] realize the significance of multi-oriented text detection and recognition to the research community. In the paper of Yingying Zhu et al. [14], they discussed in detail the recent advances and future trends for scene text detection and recognition.

B. Fine-Grained Classification:

Fine-Grained classification [28] aims for the deep insight into image that is why this problem got a lot of attention from researchers around the globe. Many approaches have been developed to address the particular problem till now with the margin of improvement in the future.

Existing deep learning-based fine-grained image classification approaches [15] could be sub-classified into the following according to the use of additional information or human inference:

1) Approaches that directly use the general deep CNNs for image fine-grained classification [16].

2) Part detection and alignment-based approaches [17].

3) Ensemble of network-based approaches [18].

4) Approaches based on attention mechanisms [19].

The prior work in fine-grained classification [28] can be simply divided into two paths. The first is to detect the discriminative object parts in the image to compensate for nuisance variations such as pose. Many parts-based methods with geometric constraints have been proposed for bird classification [16] and dogs [20].

The second track is to derive discriminative and robust features. Classic hand-crafted feature

descriptors such as the Scale Invariant Feature Transform (SIFT) [21], Histogram of Oriented Gradients (HoG) [22], and Color Histogram [23] other methods such as the Part-based One-vs-One Features (POOFs) [24] focus on modeling corresponding parts activation. Deep convolutional neural network (DCNN) approaches for general object classification achieve state-of-the-art performance for fine-grained classification by applying transfer learning [25].

C. Attention Mechanism

The idea of attention is one of the most influential ideas in deep learning allows the network to focus on specific aspects of a complex input. The main idea of the attention t mechanism is to allow the decoder to "look back" at the original input and extracts only the significant information that is important for decoding [27].

Consider we are attempting machine translation on the following sentence: "The cat is beautiful." If you can ask someone to pick out the keywords of the sentence, i.e. which ones describe the most meaning, they would likely say "cat" and "beautiful." Articles like "the" and "is" are not as relevant in translation as the previous words (though they aren't completely insignificant). Therefore, we focus our attention on important words. We use the attention mechanism on texts to get our most relevant words from text features and we put attention on the visual features to get our most relevant features like edges, color, size of object, etc.

The attention mechanism [27] scores each input word (via dot product with attention weights), then to create a distribution scores are passed through softmax function. An attention vector is produce by multiplying distribution with the context vector and then passed to the decoder. The advantage of attention are its ability to identify the information in an input most pertinent to accomplishing a task, increasing performance especially in natural language processing but it increases computations unlike to human.[30].

D. Multimodal Fusion:

Multimodal processing [35] significantly enhances the understanding, modeling, and performance of human-computer interaction. In multimodal fusion [31], user interaction with system is through various input modalities like speech, gesture, and eye gaze. In our context, different multimedia researchers presented different fusion strategies used for combining multiple modalities in order to an encoder-decoder architecture accomplish various tasks [28, 29, 33].

The literature on multimodal fusion [31] research is presented through several classifications based on the fusion methodology. The methods can be described from their advantages, weaknesses, basic concept, and their usage in various analysis tasks but multimodal fusion has several issues that influence the process such as contextual information, confidence level, synchronization between different modalities, etc. In 2016 [32] uses multilayer and multimodal fusion of deep neural networks for video classification. In 2017 [33] uses weakly paired multimodal fusion for object recognition. [34] Uses Multimodal deep networks for image-based document and text classification by introduce an end-to-end learnable multimodal deep network that jointly learns text and image features and performs the final classification based on a fused heterogeneous representation of the document. They validated their approach on the Tobacco3482 and RVL-CDIP datasets. In 2020 [28] did Fine-grained Classification by the Combination of Visual and Locally Pooled Textual Features.

III. PROPOSED METHODOLOGY

Our approach is to classify soft drinks images into their respective classes with the help of text and visual features. We extract textual and visual features from the input image with the help of different models [37, 40] and treat those features as input for our multimodal [31] which combines both the inputs to anticipate the classification of the given image. Textual cues play a key role in fine-grained classification [28], especially in the classification of business places such as bakery, café, bookstore, and daily use products. Multiple models have been developed to address the particular problem. These models assist us to extract the textual information that is highly useful for image classification. We adopt word2vec [40].

Visual features are the second input to our modal. The visual feature is the information about the contents of an image which describes its specific structures such as shapes, edges, objects, patterns, and colors i.e. properties. In our case, we extract the visual features (works as second input) with the help of VGG pertained modal [37] on ImageNet [36] by fine-tuning the modal with our dataset. These visual features work as building blocks with the texture input to give us the desired results.

The 224x224 Input images are transferred to VGG model [37] for visual features extraction. We extract visual features by importing the VGG [37] model from tensor flow keras with the pretrained weights. The top layer is set to false when loading the model. Further, we unfreeze the last five layers to fine tune the modal. After defining the model, PIL object of the image has to be converted in a pixel data NumPy array where we only have one sample and the values are then appropriately scaled to get the features. We get the feature vector 'y_f' from the last max-pooling layer as a 4096 dimensional features. Feature extraction part is from the input layer to the last max pooling layer.

At the same time, the input is send to OCR for character recognition. OCR gives us the classification of text by localizing and recognizing the text. Then OCR saves the recognized text in a file which is then passed to word2vec [40] model for textual representation 'x_f'. Word2Vec [40] is a two-layer neural networks which has been trained for the reconstruction of linguistic contexts of words. It takes a large corpus as its input and produces a vector space, of given dimensions, with each unique word in the corpus being assigned a corresponding vector in the space. The purpose and usefulness of Word2vec [40] is to group similar words vectors together in vector space. That is, it detects similarity between vectors by using the 'cosine similarity' function. Let 'a' be the first vector and 'b' be the second vector,

$$\text{Cos(a, b)} = \frac{a.b}{\|a\| * \|b\|}$$

$$x_f = \cos(a, b) \quad (1)$$

After that, we put an attention mechanism [30] on both inputs of multimodal [31] i.e. visual features and textual features. There can be some recognized text that is more relevant than others at the moment of discriminating similar classes. So we need to capture the inner correlation between the textual and visual features. The attention mechanism learns a tensor of weight that is used between the visual features and textual features. Let 'X' be the extracted textual features and 'Y' is visual features and weight is 'W'. We compute the attention mechanism by:

$$w_a = \text{Softmax}\ (\tan h\ (y_{fa}^t . W.x_f))$$

$$x_{fa} = w_a . x_f \quad (2)$$

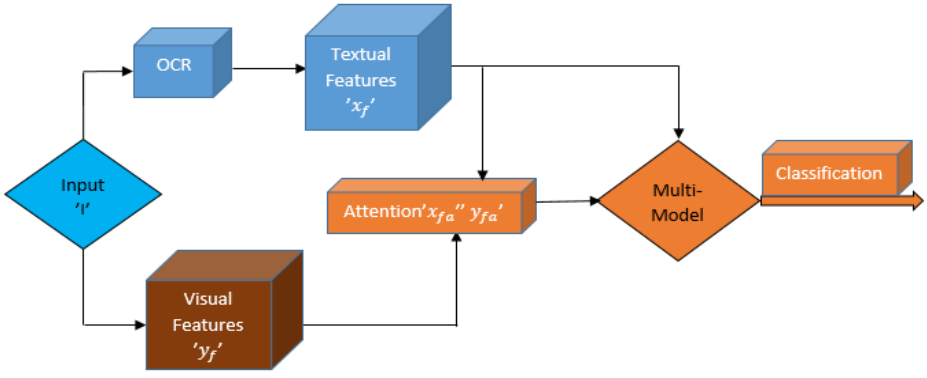

Figure 2. Proposed Model Pipeline_ Text and Visual features are attended and Combined for Fine Grained Classification.

The resulting normalized attended vector$'w_a'$, is multiplied with the textual features $'x_f'$ to obtain the final attended textual features $'x_{fa}$ '. The obtained textual features $'x_{fa}$ ' and the visual features$'y_{fa}$ ' are concatenated in the multimodal to form the final features by

$$Z = [^{x_{fa}} +^{y_{fa}}] \quad (3)$$

Finally, the resulting vector serves as input to a final classification layer that outputs the probability of a given class based on low-rank bilinear pooling operation.

IV. Experiments and Results

A. Dataset:

We have collected the new dataset of soft drink bottles of 10 classes in Pakistan with 375 original images. The dataset contains several occluded, rotated, low quality and blurred text instances which increases the difficulty of performing successful text recognition. Due to limited resources, our dataset is not fully organized and has many limitations. The images are divided into 2 sets having 200 training images, 175 test images.

B. Implementation Details:

We start by augmentation the training images flipped vertically and horizontally to make 2 more images of every image to avoid the problem of overfitting for feature extraction mode. So the total number of training images is 600. We load the image and convert it to array using keras preprocessing. We also expand the dimensions and use a preprocess function in keras to fit the image according to our model. Then the predicted output is sent to for attention.

We extract text from all of the images by using the combination of tesseract [39] and easyOCR [38]. EasyOCR is used to localize a text and tesseract is used for the recognition of the text. The extracted text is saved in two files each set of training set text and test set text. Then these sets are loaded for word embedding [40].To recognize context meanings we use two layer shallow neural network to describe word embedding. We import word2vec [40] from genism library.

We put the attention on both textual and visual features for the fine-grained classification [28] of our dataset. The network is trained for 5 epochs with Adam optimizer. The batch size employed in all our experiments is pre-defined from a library 'config', with a learning rate of 0.0001, momentum of 0.9.

These experiments were implemented by using tensor flow deep learning framework on a simple laptop of 8 GB ram and 2.7 GHz (i7).

C. Results:

Random results of some classes are shown below.

		Nestle juice	Nestle juice (0.896555)
		Sprite	Sprite (0.984277)

Bad results are because of lot of noise or the quality of the image.

D. Comparison Graphs:

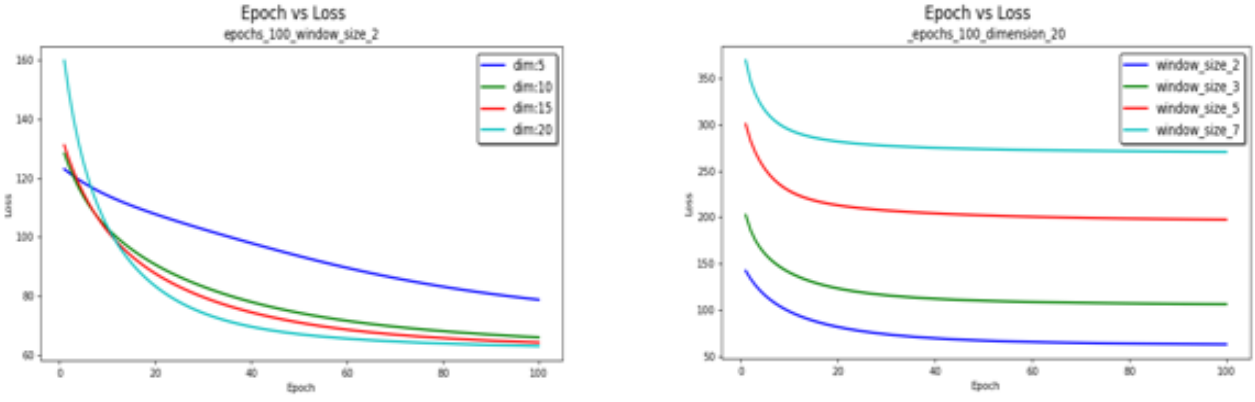

Figure 3. Visualization of Epoch vs Loss with different window size and dimension

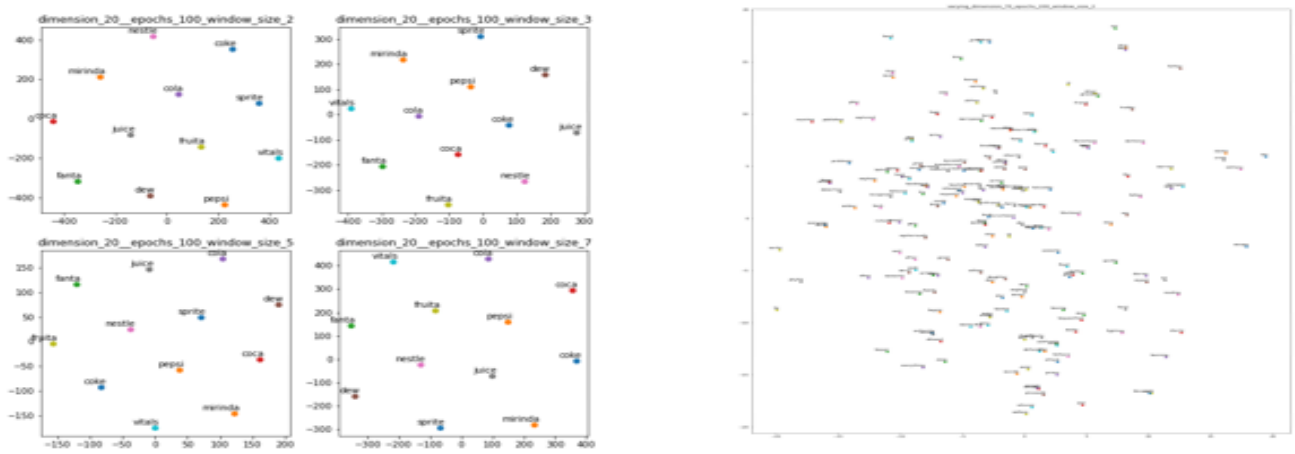

Figure 4. Parameters visualization with different window size and dimension

V. Conclusion

In this paper we demonstrated the importance of textual and visual cues for fine grained classification. We developed a frame work for precise classification of soft drink bottles by combining pre-trained models. The results show that the textual features plays more important role then visual features for classification of real life products.

Furthermore, as the system is created with low resources it can be modified and enhanced for better performance on large datasets and real world classification applications.

References

[1] Z Akata, S Reed, D Walter, H Lee, "Evaluation of output embeddings for fine-grained image classification, " pattern recognition," 2015.

[2] X He, Y Peng, "Fine-grained image classification via combining vision and language," Computer Vision and Pattern Recognition, 2017.

[3] Maron, AL Ratan," Multiple-instance learning for natural scene classification," ICML, 1998.

[4] W Geng, F Han, J Lin, L Zhu, J Bai, S Wang, "Fine-grained grocery product recognition by one-shot learning," Proceedings of the 26th ACM international conference on Multimedia," 2018.

[5] S Albawi, TA Mohammed, "Understanding of a convolutional neural network," IEEE, 2017.

[6] SE Umbaugh," Digital image processing and analysis: human and compute vision applications with CVIPtools," Amazon book, 2010.

[7] Q Ye, D Doermann, "Text detection and recognition in imagery: A survey," IEEE transactions on pattern analysis, 2014.

[8] L Neumann, J Matas, "A method for text localization and recognition in real-world images," Asian conference on computer vision, 2010.

[9] A Coates, B Carpenter, C Case, "Text detection and character recognition in scene images with unsupervised feature learning," IEEE, 2011.

[10] M Jaderberg, A Vedaldi, A Zisserman, "Deep features for text spotting," European conference on computer, 2014.

[11] C Yao, X Bai, W Liu, "A unified framework for multioriented text detection and recognition,"IEEE Transactions on Image Processing, 2014

[12] P Shivakumara, A Dutta, CL Tan, U Pal, "Multi-oriented scene text detection in video based on wavelet and angle projection boundary growing," Multimedia tools and applications, 2014.

[13] Z Zhang, C Zhang, W Shen, C Yao, "Multi-oriented text detection with fully convolutional networks,"pattern recognition, 2016.

[14] Y Zhu, C Yao, X Bai, "Scene text detection and recognition: Recent advances and future trends,"Frontiers of Computer Science, 2016.

[15] B Zhao, J Feng, X Wu, S Yan, "segmentation," International Journal of Automation, 2017.

[16] N Zhang, J Donahue, R Girshick, T Darrell, "Part-based R-CNNs for fine-grained category detection," European conference, 2014.

[17] E Gavves, B Fernando, CGM Snoek, "Fine-grained categorization by alignments," IEEE 2013.

[18] P Baraldi, M Compare, S Sauco, E Zio, "Ensemble neural network-based particle filtering for prognostics,"Mechanical Systems and Signal, 2013.

[19] F Fan, Y Feng, "D Zhao Multi-grained attention network for aspect-level sentiment classification," conference on empirical methods, 2018.

[20] OM Parkhi, A Vedaldi, A Zisserman, "Cats and dogs," IEEE conference, 2012.

[21] G Lowe, "Sift-the scale invariant feature transform," Int. J 2004.

[22] N Dalal, B Triggs, “Histograms of oriented gradients for human detection,” IEEE computer society conference, 2005.

[23] J Van De Weijer, C Schmid, J Verbeek, “Learning color names for real-world applications,” IEEE Transactions, 2009.

[24] T Berg, PN Belhumeur,“Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation,” Proceedings of the IEEE, 2013.

[25] KC Kamal, Z Yin, B Li, B Ma, “Transfer learning for fine-grained crop disease classification based on leaf images,” IEEE, 2019.

[26] V Badrinarayanan, A Kendall, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” IEEE transactions on, 2017.

[27] P Rodr ́guez, D Velazquez, G Cucurull,” Pay attention to the activations: a modular attention mechanism for fine-grained image recognition,” IEEE Transactions, 2019.

[28] A Mafla, S Dey, AF Biten, L Gomez, “Fine-grained image classification and retrieval by combining visual and locally pooled textual features,” WACV, 2020.

[29] X Bai, M Yang, P Lyu, Y Xu, J Luo, “Integrating scene text and visual appearance for fine-grained image classification,” IEEE Access, 2018.

[30] K Cho, A Courville, Y Bengio, “Describing multimedia content using attention-based encoder-decoder networks,” IEEE Transactions on Multimedia, 2015.

[31] PK Atrey, MA Hossain, A El Saddik, MS Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," Multimedia systems, 2010.

[32] X Yang, P Molchanov, J Kautz, “Multilayer and multimodal fusion of deep neural networks for video classification,” Proceedings of the 24th ACM, 2016.

[33] H Liu, Y Wu, F Sun, B Fang, “Weakly paired multimodal fusion for object recognition,”IEEE, 2017.

[34] N Audebert, C Herold, K Slimani, C Vidal, “Multimodal deep networks for text and image-based document classification,” Joint European Conference, 2019.

[35] P Maragos, A Potamianos, P Gros, “Multimodal processing and interaction: audio, video, text,” IEEE 2008.

[36] J Deng, W Dong, R Socher, LJ Li, K Li, “ImageNet,” IEEE, 2009.

[37] Karen Simonyan, Andrew Zisserman, “Very deep convolutional networks for large-scale image recognition,” Department of Engineering Science, University of Oxford, 2015.

[38] A Karnawat, K More, T Rade, B Rane, M Mulik, “A Survey on Easy OCR Techniques used to build Systems for Visually Impaired People,” ITB, 2016.

[39] R Smith, “An overview of the Tesseract OCR engine,” Ninth international conference on document analysis, 2007.

[40] KW Church, “Word2Vec,” Natural Language Engineering, 2017.