International Journal of Engineering & Technology, 7 (2.27) (2018) 56-60

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Semantic image annotation using convolutional neural network and WordNet ontology

Jaison Saji Chacko[1]*, Tulasi B. [1]

[1] Department of Computer Science, CHRIST (Deemed to be University), Bengaluru
**Corresponding author E-mail: jaison.chacko@mca.christuniversity.in*

Abstract

Images are a major source of content on the web. The increase in mobile phones and digital cameras have led to huge amount of non-textual data being generated which is mostly images. Accurate annotation is critical for efficient image search and retrieval. Semantic image annotation refers to adding meaningful meta-data to an image which can be used to infer additional knowledge from an image. It enables users to perform complex queries and retrieve accurate image results. This paper proposes an image annotation technique that uses deep learning and semantic labeling. A convolutional neural network is used to classify images and the predicted class labels are mapped to semantic concepts. The results shows that combining semantic class labeling with image classification can help in polishing the results and finding common concepts and themes.

Keywords: *Convolutional Neural Networks; Deep Learning; Image Annotation; Semantic Labeling; WordNet Ontology.*

1. Introduction

Semantic image annotation is the addition of meaningful data to an image so that a machine can understand the contents of the image. It is an important problem in computer vision. Manual annotation of images is a very expensive and time consuming process, especially in scientific communities and large business organizations as they have huge amount of data to deal with. Moreover, manual annotation is subjective when it comes to data such as images and videos because people interpret visual data differently. This problem was first addressed in psychology and is called the 'semantic gap'. It is defined as the distance between information that can be extracted from the visual data and the interpretation that different users have for the same data. Overcoming the semantic gap can enable rapid development in the field of computer vision and artificial intelligence.

Content-based Image Retrieval (CBIR) techniques find relevance based on low-level visual features such as texture, shape and color. This is not very effective as it only checks for visual similarity, whereas semantic similarity also holds an important role in efficient image annotation.

In this paper, we propose an annotation technique for images using convolutional neural networks and semantic labeling. The problem of semantic image annotation can be divided into two main sub problems.

1) Multi-Label Image classification &
2) Semantic class labeling.

A convolutional neural network (CNN) is a type of deep neural network that is widely used for problems involving pattern recognition, such as computer vision and speech recognition. CNNs are inspired by the biological neurons and their inter-connections in the animal visual cortex. One of the main advantages of CNNs is that it requires very less preprocessing and is capable of learning features from the training data automatically. In this work, we trained a CNN to do multi-label image classification where an image can belong to multiple classes. A convolutional neural network usually consist of convolution layers and pooling layers. The convolution layer has learnable filters which are also known as kernels. Pooling layer is used for down-sampling of the data. The most common pooling technique used is max pooling.

Semantic class labeling is a technique in which the class labels are represented as concepts which are part of an ontology. Ontologies define concepts and categories in a particular domain, with their properties and relationships between them. Ontologies are usually equated with taxonomic hierarchies of concepts. It is very useful for knowledge representation and inference. For this experiment we use the WordNet ontology [1] for the semantic class labeling. Class labels represented using the WordNet ontology can be used to infer additional knowledge about the images.

We use the CNN to predict class labels for a given image. The labels are mapped to semantic concepts in WordNet. We find the distinct common ancestors among the labels using the WordNet hierarchy to find general concepts present in the image. The results show that finding the LCH (lowest common hypernyms) among the labels help in identifying abstract concepts of an image. For example, if the class labels are 'computer keyboard' and 'mouse', then their LCH is 'input devices'. In this way we can identify common concepts and themes of an image. The overall design of the proposed model is shown in Figure 1.

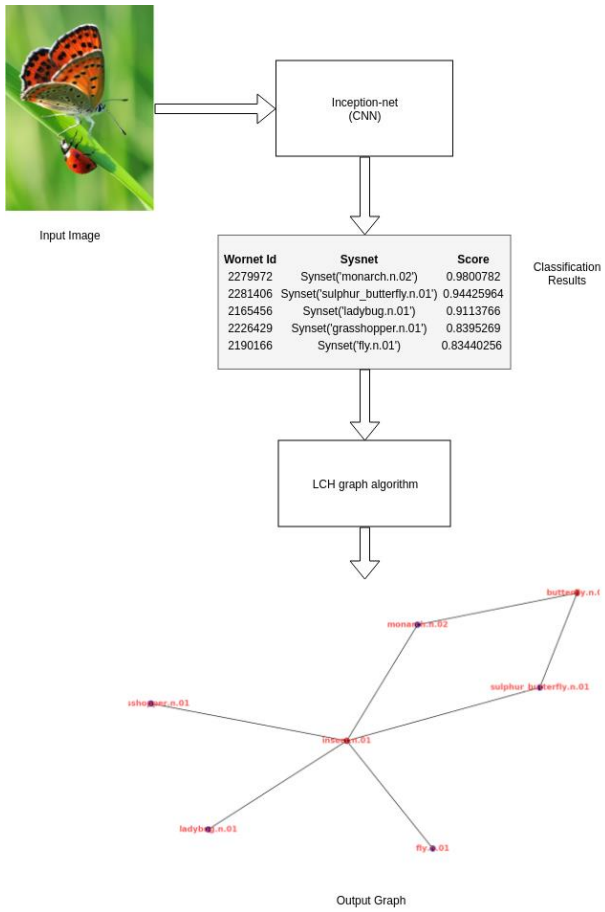


Fig. 1: Proposed Model.

We believe that this is the first attempt to improve semantic annotation of images using deep learning and WordNet ontology by creating abstractions using LCH. The remainder of this paper is organized as follows. In the next section, we review related works and approaches that has already been proposed. Section 3 describes the experimental setup. Section 4 describes multi-label image classification, algorithms and training results. In Section 5, we discuss how semantic class labeling is done with the WordNet vocabulary and in Section 6, we conclude with results and future work.

2. Related work

There have been several approaches to classify and annotate images. Some of the traditional approaches involve retrieving visual features from images and mapping them to semantic labels [2] [3] [4] [5]. Linked data technologies such as ontologies can be used for annotation of images [6]. Ontologies describe concepts and hierarchies. Manual annotation is mostly ontology based which gives accurate results. Another common approach is keyword based annotation which is less time consuming compared to ontology based methods but can give relatively less accuracy. Both the techniques are prone to bias and subjectivity as higher level interpretation of images can be different from person to person [7]. Semantic image annotation has been proved to be most useful in the field of medicine. Inferring knowledge from medical images is important for accurate diagnosis of diseases. Many methods [8] [9] [10] have been proposed for semantic annotation of medical images. In 2003, O. Marques et al., proposed a three-layer model [11] for image annotation and retrieval. The lowest layer is the visual plane which organizes visual features extracted from the raw image pixels. These features could be texture, colour or shapes. These features are mapped with semantically meaningful keywords in the next layer called the semantic plane. These labels are mapped on to the ontological plane which has domain specific schemas and ontologies to which the labels belong. To map the low level features of the images to high level semantics, the paper proposes a Bayesian network. The main drawback of this approach is that Bayesian models are not very good at tasks such as image classification as compared to the deep learning approach. In 2004, Gustavo Carneiro et al., proposed a model for automatic image annotation [12] where a set of latent variables are introduced, which encodes the hidden state of the world. During training, a set of labels are assigned to each image and the image is segmented into different regions and a unsupervised learning algorithm is run on this data to find the joint density of the high level semantic labels and the low level visual features. The proposed model using discrete cosine transform (DCT) for feature extraction. In 2016, Linan Feng proposed a system for image retrieval and annotation based on semantic concept co-occurrence [13]. In his paper, he describes a new approach to automatically generate descriptions for the images by considering concept co-occurrence patterns in the pre-labelled training dataset that makes it possible to create complex semantic descriptions for scene images. The main focus of the paper is on the hypothesis that multiple concepts co-occur frequently across images form patterns which could provide context and helps in inferring individual concepts from the image. The major drawback with the existing solutions for semantic image annotation is that, the feature extraction techniques used are very primitive. Using a fixed set of features such as colour density, texture or feature vectors generated from DCT cannot generalize well. Whereas, a more refined approach would be to use a deep learning algorithm to automatically learn useful features from images. Most current approaches for image classification uses convolution neural networks [14] [15] [16]. CNNs outperforms most of the traditional machine learning algorithms when it comes to image classification. In 2016 Christian Szedegy et al., described the Inception-v3 model [4] which is a deep convolutional neural network that was benchmarked on the ILSVRC 2012 classification challenge and demonstrated significant improvement over the state of the art. Another important work in the field of image recognition is transfer learning where mid-level image representations learned by a CNN can be reused. Maxime Oquab et al., designed a method of transfer learning [17] which shows that despite the difference in image statistics and tasks in two datasets, transferred representation leads to significant improvement in results.

3. Experimental setup

3.1. Tiny ImageNet

The dataset we used is Tiny ImageNet, which is a subset of the ImageNet dataset that is used as a benchmark for the ILSVRC competition, which is an annual image recognition challenge. The dataset consists of 200 classes with 500 images per class. Each image is 64x64 pixels in size.

3.2. Tools used

We used Google Tensorflow library for the CNN, the NLTK package for WordNet processing and NetworkX for visualization of the LCH graphs. Tensorflow is an open-source machine learning library which is primarily used to build and train deep neural networks. Natural Language Toolkit (NLTK) is an open-source library used for natural language processing. WordNet is available as a NLTK corpus reader. Using this package we can extract hypernyms, hyponyms and similarity between Synsets. NetworkX is a Python package used to build complex networks. It provides tools to create and visualize graphs. We used NetworkX to build the WordNet LCH graph for the class labels.

Fig. 2: Sample Images from the Tiny ImageNet Dataset.

4. Multi-label image classification

Convolutional neural networks have demonstrated state-of-the-art performance in multi-class image classification problem, which aims to assign a label to an image from a predefined set of classes. This problem has been studied extensively for past several years. However, multi-label image classification [18] is a more practical problem, since most images have more than one object in it. One of the main challenges of multi-label image classification is a lack of multi-labelled image datasets. Most of the benchmark datasets like ImageNet are single-labelled.

4.1. Inception-v3

We use the Inception-v3 [19] model for the image classification task. Inception-v3 is a deep convolutional neural network trained for the ImageNet Large Visual Recognition Challenge 2012. ImageNet has more than 1 Million images and 1000 classes. The model is used for multi-class classification whereas our task is to do multi-label classification. In order to do that, instead of calculating a probability distribution of classes, individual probability of classes should be calculated independently. As for the original Inception-v3 model, it uses a softmax activation function at the last layer which computes the probability distribution over K different possible outcomes. It converts a K-dimensional vector z of real values to a K-dimensional vector σ(z) of real values in the range [0, 1] that adds up to 1, so only the most significant class will be highlighted. The function is given by

$$\sigma(z) = \frac{e^{zj}}{\sum_{k=1}^{K} e^{zk}} \text{ For } j=1\ldots K.$$

Instead of the softmax activation function, we can use a sigmoid function S(x) which takes probability for individual classes.

$$S(x) = \frac{1}{1+e^{-x}}$$

4.2. Transfer learning

It is not always practical to train an entire convolutional neural network from scratch, because it takes a lot of computation and large amount of quality data. Instead, it is common to use a pre-trained model and fine-tune it for the task in hand [20]. From a CNN pre-trained on ImageNet, the last fully-connected layer can be removed and the remaining layers can be considered as a fixed feature extractor for the new dataset. This would compute a feature vector containing activations of the hidden layer immediately before the last layer. These features are called CNN codes. Once these codes are computed for each image in the new dataset, it can be trained on a linear classifier. Retraining Inception-v3 to fit to a custom image classification task is a fairly common practice and is frequently used to build highly accurate image classifiers with less data and computation [21].

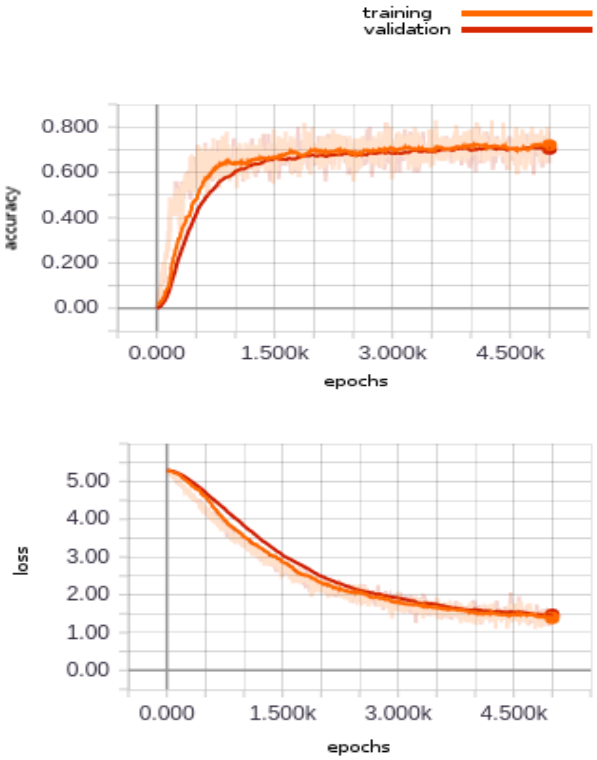


Fig. 3: Convnet Training Results.

As the Tiny ImageNet dataset we use is similar to the original ImageNet, using Inception-v3 model which is pre-trained on the ImageNet dataset is effective as the higher-level features in the ConvNet are relevant. We used Google TensorFlow library for training the CNN. For each image in the training dataset, a bottleneck file is created which contains the CNN codes for the particular image. The Inceptionv3 model was retrained for 5000 epochs on the Tiny ImageNet datasets. The final training results showed an error rate of 28 %. The accuracy is reasonable considering the smaller image dimensions (64 x 64 pixels) and large number of classes (200). As we used transfer learning, the training took significantly less time compared to training the model from scratch.

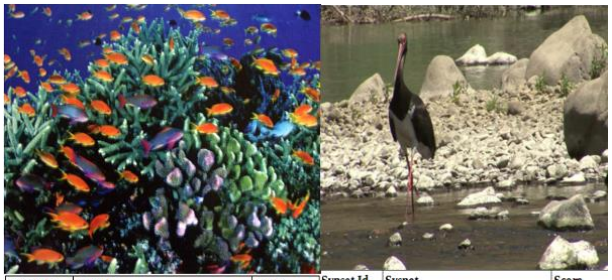
5. Semantic labelling

Consider a dataset T = {I_1 …, I_N} of images Ii and a semantic vocabulary W = {w_1 …, w_T} of semantic labels w_i. The goal of semantic image annotation is to, given an image I, assign a set of labels or captions from L that best describes I. The goal of semantic retrieval is to, given a semantic label w_i, extract the images in the database that contain the associated visual concept. ImageNet dataset is annotated with WordNet ontology which has in-depth information about the individual classes. WordNet is a large database that defines the lexical and semantic relations between English words. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms called Synset. In the tiny ImageNet dataset, instead of class labels, each class is represented using a unique Synset id. With the WordNet vocabulary we can retrieve additional semantic information about the class labels such as hyponyms, hypernyms and semantic similarity between them.

5.1. Test results

Synset Id	Synset	Score
3085013	computer_keyboard.n.01	0.97067
4399382	teddy.n.01	0.93394
3584254	ipod.n.01	0.83247
3179701	desk.n.01	0.82618
4265275	space_heater.n.01	0.79119

Synset Id	Synset	Score
1644900	tailed_frog.n.01	0.98266
1641577	bullfrog.n.01	0.97945
2281406	sulphur_butterfly.n.01	0.91
2226429	grasshopper.n.01	0.90318
2279972	monarch.n.02	0.88190



Synset Id	Sysnet	Score
9256479	Synset('coral_reef.n.01')	0.981776
1443537	Synset('goldfish.n.01')	0.976247
1917289	Synset('brain_coral.n.01')	0.946066
2281406	Synset('sulphur_butterfly.n.01')	0.8876
2206856	Synset('bee.n.01')	0.86737

Synset Id	Sysnet	Score
2002724	Synset('black_stork.n.01')	0.992844
1855672	Synset('goose.n.01')	0.881375
2058221	Synset('albatross.n.02')	0.744415
2423022	Synset('gazelle.n.01')	0.715505
4597913	Synset('wooden_spoon.n.02')	0.710858

Fig. 4: Test Results.

Using the NLTK tool for Python, we can retrieve the hypernyms of the predicted class labels. For example in order to find the hypernym of computer keyboard, we can use wn.synset ('computer keyboard.n.01').hypernyms ()
Which outputs: 'input device' and 'keyboard'
We can also find the semantic similarity between the predicted class labels. As there is high probability for semantically similar concepts to co-occur in an image [13], for example 'cars' and 'trucks'.
The semantic similarity score can help in accurately labelling images. In order to generate further abstraction or generalization for the predictions, we can find the LCH (lowest common hypernyms) among the predicted Synsets. The LCH is the lowest single hypernym that is shared by two given words. Once we find the distinct lowest common hypernyms among the labels, we can plot it as a graph. Generalization of the prediction using LCH can help identifying the common themes or categories present in an image, for example when the predicted labels are 'tailed frog' and 'bullfrog', their LCH is 'frog'.

Algorithm 1 LCH graph

```
procedure GENERATE_LCH_GRAPH
    L_i: predicted synsets
    L_i : {L_1....L_n}
    Initialize graph G
    for i → 1 to n do
        for j → 1 to n do
            if L_i not equal to L_j then
                //Lowest Common Hypernym
                lch= LCH(L_i, L_j)
                if L_i not in G then
                    G.add_node(L_i)
                if L_j not in G then
                    G.add_node(L_j)
                if lch not in lch_list then
                    G.add_node(lch)
                // Map concepts to their LCH
                G.add_edge(L_i,lch)
                G.add_edge(L_j,lch)
```

The LCH graph shows the relationship between the class labels using their common ancestors in the Wordnet vocabulary.

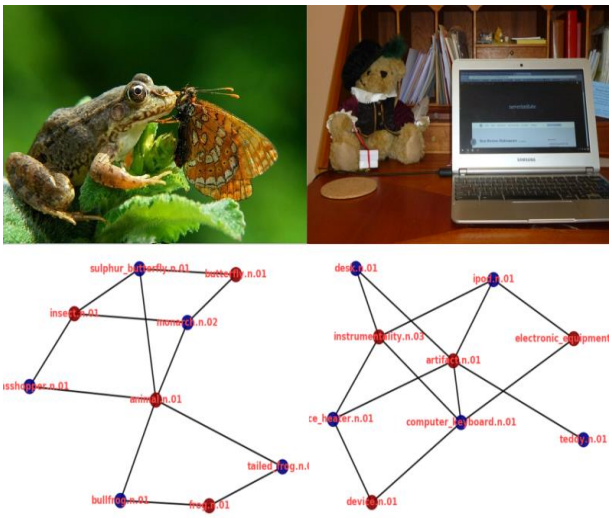


Fig. 5: Example for LCH graphs.

Figure 5 shows graphs depicting relationship between the class labels by mapping them to their LCH. From the first graph we can infer that the labels 'tailed frog' and 'bullfrog' share the same hypernym 'frog'. Similarly the labels 'sulphur butterfly' and 'monarch' share the same hypernym 'butterfly'. The common hypernyms among the class labels are 'frog', 'butterfly', 'insects' and 'animal'. The objects in the image can be correctly identified under these categories. Similarly in the second graph, using LCH we correctly identified 'electronic equipment', 'artefact', 'device' and 'instrumentality'. So the abstraction is useful to find common concepts and themes. One of the drawbacks of the approach is that we are taking top 5 predictions and finding the LCHs among them, so if the wrong predictions are very different from the correct labels, the LCH graph will not give accurate abstractions. As the predicted labels, 'sulphur butterfly' and 'bee' in figure 5 are incorrect, the LCH graph does not produce a good abstraction. If the predicted label is wrong but similar to the correct class label, finding the LCH can help in generalizing the result but if the prediction is completely wrong, the LCH graph will not help in creating accurate abstraction.

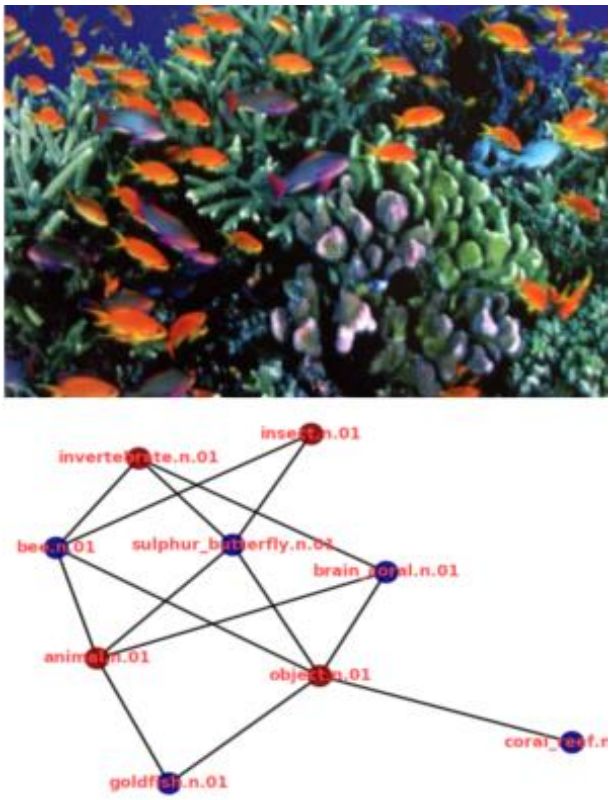


Fig. 6: Example for Bad LCH Graph.

6. Conclusion

Automatic image classification using CNNs have shown significant improvement over traditional machine learning approaches and requires very less pre-processing. But the knowledge inference is only limited to the class labels. Automatic image classification combined with semantic class labelling can enable more useful image annotations from which additional knowledge can be inferred. Manual annotation requires a lot of time and man power. Using deep learning we are able to semi-automate the annotation process. The proposed method uses a CNN to do multi-label image classification where the class labels are mapped to the labels in the WordNet vocabulary. By exploiting the fact that semantically similar concepts occur together frequently, we can infer common patterns or themes in an image by finding the common ancestor between these concepts. In future, we would like to extend this research in the following directions. Firstly, we would like to do further experiments on semantic class labelling with different datasets. Next, along with image classification, we would like to experiment with image localization and segmentation.

References

[1] Millers, George A"WordNet: a lexical database for English." Communications of the ACM 38.11 (1995): 39-41. https://doi.org/10.1145/219717.219748.

[2] Stamou, Giorgos, et al., "*Multimedia annotations on the semantic web.*" IEEE Multimedia 13.1 (2006): 86-90 https://doi.org/10.1109/MMUL.2006.15.

[3] Little, Suzanne, Ovidio Salvetti, and Petra Perner. "*Semi-automatic semantic annotation of images.*" Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007). IEEE, 2007. https://doi.org/10.1109/ICDMW.2007.22.

[4] Verma, Yashaswi, and C. V. Jawahar. "*Image Annotation by Propagating Labels from Semantic Neighborhoods.*" International Journal of Computer Vision (2016): 1-23.

[5] Dureja, Aman, & Payal Pahwa. "*Image retrieval techniques: a survey.*" International Journal of Engineering & Technology [Online], 7.1.2 (2018): 215-219. Web. 7 May. 2018.

[6] Berners-Lee, Tim, James Hendler, and Ora Lassila. "*The semantic web.*" Scientific american 284.5 (2001): 34-43. https://doi.org/10.1038/scientificamerican0501-34.

[7] Reena Pagare and Anita Shinde, "*A Study on Image Annotation Techniques*", Harlow, England: International Journal of Computer applications, Volume 37- No.6, January 2012.

[8] Luque, Edson F., Daniel L. Rubin, and Dilvan A. Moreira. "*Automatic Classification of Cancer Tumors using Image Annotations and Ontologies.*" 2015 IEEE 28th International Symposium on Computer-Based Medical Systems. IEEE, 2015.

[9] Seifert, Sascha, et al., "*Semantic annotation of medical images.*" SPIE medical imaging. International Society for Optics and Photonics, 2010.

[10] Rubin, D. L., Rodriguez, C., Shah, P., and Beaulieu, C. "*Semantic Annotation and Markup of Radiological Images.*" AMIA Annual Symposium Proceedings (2008), Volume 2008, p. 626.

[11] O. Marques, N. Barman, "*Semi-Automatic Semantic Annotation of Images Using Machine Learning Techniques*", Proc. of ISWC, pp. 550565, 2003.

[12] Carneiro, Gustavo, and Nuno Vasconcelos. "*Formulating semantic image annotation as a supervised learning problem.*" Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Vol. 2. IEEE, 2005.

[13] Feng, Linan, and Bir Bhanu. "*Semantic Concept Co-Occurrence Patterns for Image Annotation and Retrieval.*" IEEE transactions on pattern analysis and machine intelligence 38.4 (2016): 785-799. https://doi.org/10.1109/TPAMI.2015.2469281.

[14] LeCun, Yann, et al., "*Gradient-based learning applied to document recognition.*" Proceedings of the IEEE 86.11 (1998): 2278-2324. https://doi.org/10.1109/5.726791.

[15] LeCun, Yann, Koray Kavukcuoglu, and Cl´ement Farabet. "*Convolutional networks and applications in vision.*" Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on. IEEE, 2010. https://doi.org/10.1109/ISCAS.2010.5537907.

[16] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "*Imagenet classification with deep convolutional neural networks.*" Advances in neural information processing systems. 2012.

[17] Maxime Oquab, L´eon Bottou, Ivan Laptev, Josef Sivic. "*Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks.*" IEEE Conference on Computer Vision and Pattern Recognition, Jun 2014, Columbus, OH, United States. https://doi.org/10.1109/CVPR.2014.222.

[18] Boutell, Matthew R., et al., "*Learning multi-label scene classification.*" Pattern recognition 37.9 (2004): 1757-1771. https://doi.org/10.1016/j.patcog.2004.03.009.

[19] Szegedy, Christian, et al., "*Rethinking the inception architecture for computer vision.*" Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[20] Pan, Sinno Jialin, and Qiang Yang. "*A survey on transfer learning.*" IEEE Transactions on knowledge and data engineering 22.10 (2010): 1345-1359. https://doi.org/10.1109/TKDE.2009.191.

[21] Sai V, Bhavya, Narasimha Rao G, Ramya M, Sujana Sree Y, & Anuradha T. "*Classification of skin cancer images using TensorFlow and inception v3.*" International Journal of Engineering & Technology [Online], 7.2.7 (2018): 717-721. Web. 9 May. 2018.