

Multilabel Image Annotation using Multimodal Analysis

Pavithra S S¹, Chitrakala S²

¹Student, ²Professor,

^{1,2}Anna University - CEG Campus, Chennai, Tamil Nadu, India

ABSTRACT

Image Annotation is one of the most important powerful tools in the field of Computer Vision applications. It has potential application in Face recognition, Robotics, Text recognition, Image retrieval, Image analysis etc. Also, Neural network gains a massive attention in the field of computer science recently. In neural networks, Convolutional neural network (ConvNets or CNNs) is one of the main categories to do images recognition, images classifications, Objects detections, recognition faces etc., are some of the areas where CNNs are widely used. The existing approaches obtain the information cues needed for annotation from Input Images only. This results in lack of context understanding of the post. In order to overcome this issue, Multimodal Image Annotation using Deep Learning (MIADL) approach is proposed. This approach makes use of Multimodal data i.e. Image along with its textual description / content in Automatic Image Annotation. Incorporating Image along with its textual description / content (Multimodal data) gives the better understanding of the context of the post. This will also reduce irrelevant images in image retrieval systems. It is done by using Convolution Neural network to classify and assign multiple labels for the image. It is mainly is for multi-label classification problem that aims at associating a set of textual with an image that describe its semantics. Also using Multimodal data to annotate an Image significantly boost performance than the existing methods.

KEYWORDS: Neural network, Automatic Image Annotation, Convolution Neural Network (CNN), Part-of-Speech (POS) Tagging, NUS-WIDE dataset, Multimodal, Multilabel

INTRODUCTION

Automatic Image Annotation is the process by which a computer automatically assigns label to the image. It has wide range of Applications in areas like: Face recognition, Robotics, Text recognition, etc. Image annotation aims to describe (rather than merely recognize) an image by annotating all visual concepts that appear in the image. Image annotation is a multi-label multi-class classification problem, where as Image recognition is single-label multi-class classification problem. Typically, in Image recognition, from the set of targeted classes, each Image assigned with single label at a time. The set of possible output labels are referred as target classes. For Instance, the Target classes, $C = [\text{'apple'}, \text{'cat'}, \text{'dog'}]$. Given an Apple Image, Image Recognition model predicts which class it belongs. Here, $[1\ 0\ 0]$.

Multilabel Image Annotation is one of the most important challenges in computer vision with many real-world applications. It plays an important role in content-based image understanding. Multilabel Image annotation is annotating the objects from the image with more than two labels. In multi-label case each sample can belong to one or more than one class. It is efficient to assign relevant labels to an image to improve image retrieval accuracy. For Instance, the Target classes, $C = [\text{'apple'}, \text{'cat'}, \text{'dog'}]$. Given an Image, Image Annotation model assign multiple labels to a particular Image. If an image has both cat and dog, it assigns both labels to a particular Image. Here, $[0\ 1\ 1]$.

How to cite this paper: Pavithra S S | Chitrakala S "Multilabel Image Annotation using Multimodal Analysis" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-4 | Issue-5, August 2020, pp.1005-1012, URL: www.ijtsrd.com/papers/ijtsrd33002.pdf



IJTSRD33002

Copyright © 2020 by author(s) and International Journal of Trend in Scientific Research and Development Journal. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0) (<http://creativecommons.org/licenses/by/4.0>)



Existing methods mostly use only Image to do Annotation. In addition to the Image, use of its textual description (Multimodal data) gives better understanding of the context of the post. Existing performance is improved by using Multimodal feature learning incorporating both textual description of an image and the Visual Image. This will also reduce irrelevant images in image retrieval systems. Also using Multimodal data that is both text description of an image and Visual images to annotate significantly boost performance. The difference between Image Annotation Model with Multimodal data and Only Image is shown with an example. Consider the Image given in Figure 1,



Figure 1: Sample Image

If Annotation is done based on,

- Only Image - Labels: Clouds, Sky, Sun
- Multimodal (Image and Its Textual description) – Labels: Reflection, lake, Sun, Sky, Clouds, water

Multilabel Image Annotation on multimodal data provide a deep analysis through feature extraction and learning the feature maps, which aims to annotate the image with correct labels. Existing work usually make use of conventional visual features for multilabel annotation and classification. Deep Neural Networks Learning approaches shows the efficient performance than the existing work. The main Challenges while implementing Multilabel Image Annotation using multimodal data:

- Handling Noisy textual Content
- Label dependency: For each image there are multiple relevant labels. The model needs to assign closely related label to the image
- Higher dimensionality: Handling with larger number of labels. The algorithm needs to quite fast in handling with higher dimensional contents.

RELATED WORKS

A. TEXTUAL FEATURE EXTRACTION

POS tagging is used as a preliminary linguistic text analysis in diverse natural language processing domains such as speech processing, information extraction, machine translation, and others. It is a task that first identifies appropriate syntactic categories for each word in running text and second assigns the predicted syntactic tag to all concerned words. Existing works make use of Conditional random field (CRF)-based POS tagger with both language dependent and independent feature set.

Wahab Khan [1] focused on the implementation of both machine and deep learning models for Urdu POS tagging task with well-balanced language-independent feature set.

Zhenghua Li [2] proposed an coupled sequence labeling model for exploiting multiple non-overlapping datasets with heterogeneous annotations. The key idea is to bundle two sets of POS tags together (e.g. "[NN, n]"), and build a conditional random field (CRF) based tagging model in the enlarged space of bundled tags with the help of ambiguous labeling. To solve the efficiency issue, proposed a context-aware online pruning approach for approximate gradient computation.

Part-of-Speech Tagging by Latent Analogy by Jerome R. Bellegarda [3] focused on two loosely coupled sub problems: 1) extract from the training corpus those sentences which are the most germane in a global sense, and 2) exploit the evidence thus gathered to assemble the POS sequence based on local constraints. Address by leveraging the latent topicality of every sentence, as uncovered by a global LSM analysis of the entire training corpus. Each input surface form thus leads to its own customized neighbourhood, comprising those training sentences which are most related to it. POS tagging then follows via locally optimal sequence alignment and maximum likelihood position scoring, in which the influence of the entire neighbourhood is implicitly and automatically taken into account.

B. VISUAL FEATURE LEARNING

Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, [4] proposed a unified deep convolutional neural network, denoted the MS-CNN, for fast multi-scale object detection. The detection is performed at various intermediate network layers, whose receptive fields match various object scales. This enables the detection of all object scales by feed forwarding a single

input image through the network, which results in a very fast detector. CNN feature approximation was also explored, as an alternative to input up sampling. It was shown to result in significant savings in memory and computation. Overall, the MS-CNN detector achieves high detection rates at speeds of up to 15 fps.

Z. Lu, P. Han, L. Wang, and J.-R. Wen [5] investigated the challenging problem of visual BOW representation refinement for image applications. To deal with the semantic gap and noise issues associated with the traditional visual BOW representation, Incorporated the annotations of images into visual BOW representation refinement and thus formulated it as semantic sparse recoding of the visual content. By developing an efficient algorithm, also generated more descriptive and robust visual BOW representation.

C. IMAGE ANNOTATION

A. Ulges, M. Worring, and T. Breuel [6] suggested a novel extension to image annotation that employs web-based user-driven category information like Flickr groups as an additional information source. This approach assumes images to come with a context of related pictures (e.g., taken over the same event). This context is matched with Flickr groups, and then a group-specific annotation is applied. Significant improvements of up to 100% and more have been validated on samples from the Corel dataset as well as real-world Flickr data. Also analyzed the validity of Flickr groups as a basis for our approach, and have shown two key characteristics they offer for learning visual contexts, namely a user-driven categorization and a rich group space, which aids in generalizing to novel categories.

Multi-label dictionary learning for image annotation by X. Y. Jing, F. Wu, Z. Li, R. Hu, and D. Zhang [7] proposed a novel image annotation approach named MLDL. It can conduct multi-label dictionary learning in input feature space and partial-identical label embedding in output label space, simultaneously. In the input feature space, MLDL incorporates the label consistency regularization term into multi-label dictionary learning to learn discriminative representation of features. In the output label space, MLDL learns the partial-identical label embedding, where samples with the exactly same label set can cluster together and samples with partial-identical label sets can collaboratively represent each other, to fully utilize the relationship between labels and visual features.

PROPOSED WORK

Image annotation is a multi-label multi-class classification problem. The Objective is to perform Multilabel Image Annotation using multimodal analysis that is both textual and visual image. Most Existing work focus on Multiclass classification that is each sample is mapped to one and only one label. The Proposed work assign multiple labels to a sample by using both visual content and textual data. The dataset is multimodal that consist of both textual and visual contents. The model is trained with NUS-WIDE dataset which consists of images and class labels from Flickr image metadata. Text Preprocessing module is used to remove numbers, stop words, punctuation and white spaces. Textual feature Extraction module is used to have verbal features done by Tokenization and POS Tagging. Image Preprocessing module includes Image Resizing and Gray Scale conversion techniques. Visual feature learning module focus on training

the images using CNN model to extract the visual features. Based on the training, Classification is done. Both Textual and Visual features concatenated to have the final annotation as the result.

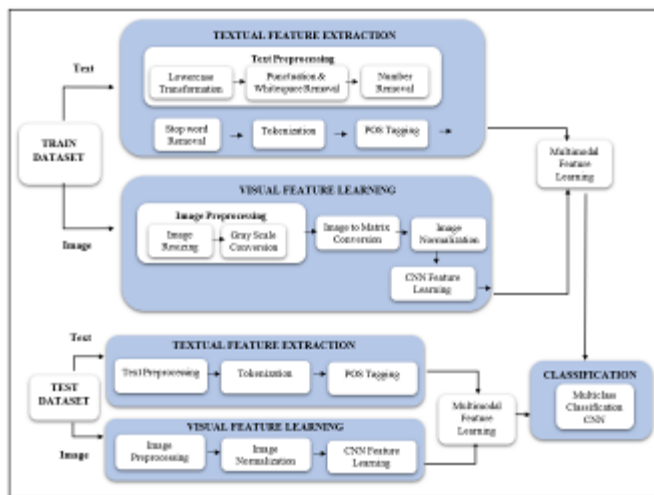


Figure 2. Overall System Architecture

The Detailed system architecture of the proposed system is shown in Figure 2. In Training phase, the model learns by using Train Dataset observations. In Testing Phase, the trained model working is tested by using Test Dataset observations. The Proposed work consist of five modules which are as follows:

1. Text preprocessing
2. Textual feature Extraction
3. Image preprocessing
4. Visual feature Learning
5. Multi class classification

A. TEXT PREPROCESSING

Cleaning and preprocessing the noisy text are essential for any kind of analysis to be performed. This module focuses on preprocessing the noisy textual description of the Images from the dataset. This is done to have meaningful contents on which the techniques are applied. Preprocessing includes the following: lowercase transformation, Number removal, Punctuation removal, Whitespace removal, Stop word removal.

Lowercase transformation: This step is done to convert all the textual tags to one unified case which will be easier for processing.

Number removal: This step is done to remove the numbers from the text. Since numbers are not relevant to the analyses it needs to be removed.

Punctuation removal: This step is done to remove the punctuations in the text which will be easier for further processing. Set of symbols like [!"#\$%&'()*+,-./:;<=>?@[\]^_`{|}~]: are removed.

Whitespace removal: This step is done to remove the leading and ending in the text.

Stop word removal: "Stop words" are the most common words in a language like "the", "a", "on", "is", "all". These words do not carry important meaning and are usually removed from texts. It is possible to remove stop words using

Natural Language Toolkit (NLTK), a suite of libraries and programs for symbolic and statistical natural language processing. The results after preprocessing is shown in Figure 3.

california ca usa museum photoshop losangeles los big nikon day arch dino dinosaur
angeles earth pillar used social chandelier kris to they d200 2000 roar natural history
museum roam rule hdr extinct trex tyrannosaurus the cs3 styrcosaurus first quality
photo matrix kros kriskros 5xp at finewartphotos platinumphoto flickrbestpics vision180

↓

california ca usa museum photoshop losangeles los big nikon day arch dino dinosaur
angeles earth pillar used social chandelier kris roar natural history museum roam rule
hdr extinct trex tyrannosaurus cs styrcosaurus first quality photo matrix kros kriskros
xp finewartphotos platinumphoto flickrbestpics vision

Figure 3: Text Preprocessing

B. TEXTUAL FEATURE EXTRACTION

This module extracts the features from the Textual contents by performing Tokenization and POS Tagging. Tokenization is the process of splitting the given text into smaller pieces called tokens. Part-of-speech tagging is applied on these tokens which aims to assign parts of speech to each word of a given text (such as nouns, verbs, adjectives, and others) based on its definition and its context. For better understanding of the post, the focus here is verbal features extractions.

C. IMAGE PREPROCESSING

This module focuses on preprocessing the images to proceed with feature learning.

1. Image Resizing:

When resizing the image, the graphic primitives that make up the image can be scaled using geometric transformations, with no loss of Image quality. The decrease in the pixel number (scaling down) usually results in a visible quality loss.

2. Gray Scale conversion:

Gray Scale conversion is done in which value of each pixel is a single sample representing only an amount of light, that is, it carries only intensity information. Grayscale images, a kind of black and white or Gray monochrome, are composed exclusively of shades of gray. The contrast ranges from black at the weakest intensity to white at the strongest. The Sample Image and its corresponding preprocessed image is given in Figure 4. The Gray Scale conversion is given in equation 1.

$$\text{Grayscale}(i,j) = 0.2989 * R(i,j) + 0.5870 * G(i,j) + 0.1140 * B(i,j); (1)$$

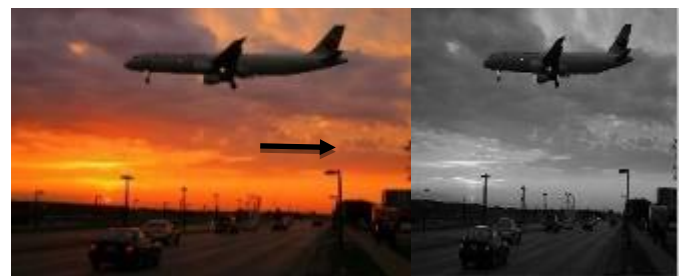


Figure 4: Image Preprocessing

D. VISUAL FEATURE LEARNING

This module focuses on learning the features from the images using Convolution neural network. The various steps in this module are Image to Matrix Conversion, Image Normalization, Feature extraction by CNN.

1. Image to Matrix Conversion

This is the first step done to understand the Image from which the features are extracted. Each Image has a corresponding matrix which consists of numbers denoting each pixel value. This pixel value depends on colour and intensity of the pixel. The value of the pixel ranges from 0 to 255 in an 8-bit gray scale image. The conversion result of an image is as follows:

```
[[[92., 73., 79.],
 [89., 70., 72.],
 [84., 72., 76.],
 ...,
 [171., 117., 91.],
 [167., 120., 90.],
 [170., 121., 91.]],

 [[90., 71., 75.],
 [89., 70., 74.],
 [84., 71., 78.],
 ...,
 [[29., 20., 15.],
 [24., 23., 18.],
 [26., 23., 18.],
 ...,
 [31., 24., 16.],
 [26., 26., 16.],
 [26., 26., 16.]]]
```

2. Image Normalization

Image Normalization is the process that changes the range of pixel values. It is mainly done to bring image to range that is normal to sense. Here, To have the range of pixel intensity from 0 to 1 the Matrix of the Image is divided by the value 255. Image Normalization equation is defined as in 2.

$$\text{Image Normalized} = \text{Image Matrix} / 255.0 \quad (2)$$

The normalized result of an image is as follows:

```
[[[0.36078432, 0.28627452, 0.30980393],
 [0.34901962, 0.27450982, 0.28235295],
 [0.32941177, 0.28235295, 0.29803923],
 ...,
 [0.67058825, 0.45882353, 0.35686275],
 [0.654902, 0.47058824, 0.3529412 ],
 [0.6666667, 0.4745098, 0.35686275]],

 [[0.3529412, 0.2784314, 0.29411766],
 [0.34901962, 0.27450982, 0.2901961 ],
 [0.32941177, 0.2784314, 0.30588236],
 ...,
 ...,
 ...]]]
```

```
[[0.11372549, 0.07843138, 0.05882353],
 [0.09411765, 0.09019608, 0.07058824],
 [0.10196079, 0.09019608, 0.07058824],
 ...,
 [0.12156863, 0.09411765, 0.0627451 ],
 [0.10196079, 0.10196079, 0.0627451 ],
 [0.10196079, 0.10196079, 0.0627451 ]]]
```

3. Feature Learning by CNN

In this step, Image features are learnt by Convolution Neural Network. It consists of an input and an output layer, as well as multiple hidden layers. The hidden layers of a CNN typically consist of Convolutional layers, ReLU layers i.e. activation functions, Pooling layers and Fully connected layers. **The Convolution layer** is the first layer in which the image (matrix with pixel values) is entered into it. The reading of the input matrix begins at the top left of image. Next the software selects a smaller matrix there, which is called a **filter** (or neuron, or core). Then the filter produces convolution, i.e. moves along the input image. **The nonlinear layer** is added after each convolution operation. It has an activation function, which brings nonlinear property. Without this property a network would not be sufficiently intense and will not be able to model the response variable (as a class label). **The pooling layer** follows the nonlinear layer. It works with width and height of the image and performs a down sampling operation on them. As a result, the image volume is reduced. This means that if some features have already been identified in the previous convolution operation, than a detailed image is no longer needed for further processing, and it is compressed to less detailed pictures.

E. MULTI CLASS CLASSIFICATION BY CNN

This module classifies and annotate the specified image with multiple labels. This Multiclass CNN Classification is done based on the features learnt in the previous step. After completion of series of convolutional, nonlinear and pooling layers, it is necessary to attach a **fully connected layer**. This layer takes the output information from convolutional networks. Attaching a fully connected layer to the end of the network results in an N dimensional vector, where N is the number of classes from which the model selects the desired class. The Final Annotation results for the given image is shown in Figure 6.



Figure 6. Final Annotation of an Image

The pseudocode for the proposed model is outlined in Table1.

Table 1. Pseudo Code of Multimodal Neural Network based Image Annotation (MIADL) Algorithm**INPUT: Multimodal Data – Textual Tags & Images from NUS WIDE Dataset****OUTPUT: Multiple Labels as Annotation****Begin***Extract_noisy_text(filename)***for each***If(text.contains(Uppercase))**Transform_to_Lowercase**If(text.contains(Numbers))**Remove_Numbers**If(text.contains(Punctuation))**Remove_Punctuation**If(textual_content.contains(Whitespace))**Remove Whitespace**If(textual_content.contains(Stopword))**For each word**Compare with stopwords**If match occur remove**Repeat until all stopwords are removed***end****for each Sentence***Tokens=tokenize(Sentence)**Append Tokens to Token_List[]**Define POS_Tag_Function:**Define is_verb:**If pos=VEB or VBZ or VBG**Return(Verbal_feature)***end***Read Set of Images**Specify Row & Column Size***for Each image***Resize to Specified dimension**Perform Gray Scale Conversion* **$Grayscale(I_{ij}) = 0.2989 * R(I_{ij}) + 0.5870 * G(I_{ij}) + 0.1140 * B(I_{ij});$** *Store the Preprocessed Image to Specified target location***end***Read Set of Preprocessed Images***for each image***Form an Image Matrix Flatten Array Perform Image Normalization* **$ImageNormalized = ImageMatrix/255.0$** *Specify for the model**batch_size**number of classes, number of epoch**number of img_channels**number of convolutional filters**size of pooling area for max pooling**convolution kernel size**Define the network model**Specify all layers**select an activation function**Specify the optimizer**Specify the metrics**Fit the model with**Training Data,**Number of Epochs,**Validation Data**Multimodal feature Learning**Final Annotation based on Multimodal learning**Textual Features + Visual Features***end****Table I – Pseudo Code for The Proposed System**

EXPERIMENTAL RESULTS:**DATASET DESCRIPTION**

The Benchmark dataset NUS-WIDE, which consists of images and class labels from Flickr image metadata is used. The dataset is multimodal that consist of both textual and visual contents. It consists of 269,648 images and the associated tags from Flickr, with a total number of 5,018 unique tags.

METRICS USED

The metrics used for evaluating the performance of the proposed model are defined as follows:

- Precision
- Recall
- F1-Score

True Positives (TP) - These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes.

True Negatives (TN) - These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no.

False positives and false negatives, these values occur when the actual class contradicts with the predicted class.

False Positives (FP) – When actual class is no and predicted class is yes.

False Negatives (FN) – When actual class is yes but predicted class in no.

Precision - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. It is computed as in Equation 3.

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (3)$$

Recall - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes. It is computed as in Equation 4.

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (4)$$

F1 score - F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall. It is computed as in Equation 5

$$\text{F1 Score} = 2 * (\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision}) \quad (5)$$

OUTPUT: The multimodal data are pre-processed, the necessary features are extracted both from text and Images and fed to CNN for training. The Table 1 shows the Input image and the corresponding final annotation as the output.

Table 2. Output of the Proposed Work




Input Noisy Text	Input Image	Final Annotation
Category: Scenery blue sea sky clouds plane philippines peak aerial mindanao planeview tawitawi bongao bongaopeak		'cloud', 'peak', 'sky', 'bridge'
Category: Sports show camera sky blackandwhite bw cloud 6 hat tattoo photoshop plane photographer baseball display military watch crowd tshirt aeroplane formation jeans cap shade denim spectators six raf fairford		'sky', 'tattoo', 'crowd', 'cap', 'person', 'airplane', 'cloud'
Category: Airport blue storm rain clouds airplane interestingness nikon escape flight explore bolt lightning thunder lightningbolt rain cloud rain clouds nearmiss 18200vr d80 nikonstunninggallery abigfave nikond80 300preset 300v1 cellformation explore interestiness		airplane, 'lightning', 'abigfave', 'clouds', 'bridge', 'sky'

Table 3 shows the calculation of category wise performance Score of Precision, Recall and F-Score value.

Table 3. Category wise Performance Score

CATEGORY	PRECISION (%)	RECALL(%)	F-SCORE(%)
Airport	79	80	80
Flora	78	75	77
Vehicles	81	78	80
Sports	76	77	77
Scenery	85	82	84

Figure 7. shows the graph of the Category wise performance of Precision, Recall and F-Score. X-axis denotes the Performance score of each Category and Y-axis denotes the Various Categories used.

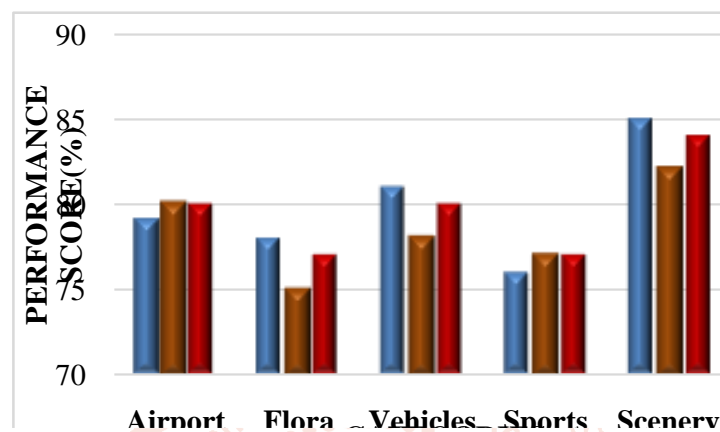


Figure 7. Category wise Performance

Table 4 shows the Overall performance score of Precision, Recall and F-Score Values.

Table 4. Performance Score

PRECISION (%)	RECALL (%)	F-SCORE (%)
80	78	80

Figure 8. shows the graph of Overall performance of Precision, Recall and F-Score. X-axis denotes the Metrics and Y-axis denotes the Performance score.

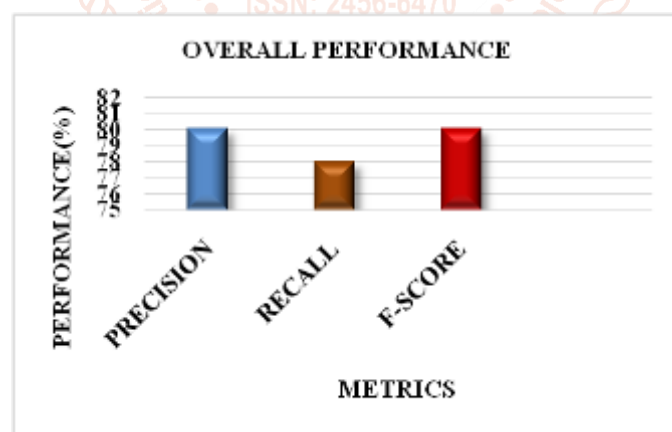


Figure 8. Overall Performance

The system after training recognizes for the test images given. Based on the results from the test images the analysis is done. The System works well for the real images. The Problem in the dataset can be minimized by sampling the dataset and training them. The overall performance of the system is Precision 80%, Recall 78% and F-score is calculated as 80% and Fig 5. Shows the results.

CONCLUSION

Thus, the proposed system for Multilabel Image Annotation using Multimodal analysis is successfully developed. This system is able to annotate an Image based on considering both Image and its noisy textual tags. The System make use of multimodal features to classify and predict the correct labels to an image. On Comparison, annotating a post by

considering its Multimodal data (Image along with its textual content) gives better understanding than having Only Image. The trained model performance is evaluated by using Precision, Recall metrics which in turn used to calculate the F-score. In Future, this can be extended and implied in Social Media Sites like Twitter, Facebook where the post has multimodal data (both text and images). Other Parameters

like location, timestamp, in addition text and Images may considered which will be useful in Detecting an event, Analysing an event and so on.

REFERENCES

- [1] W. Khan et al., "Part of Speech Tagging in Urdu: Comparison of Machine and Deep Learning Approaches," in IEEE Access, vol. 7, pp. 38918-38936, 2019, doi: 10.1109/ACCESS.2019.2897327.
- [2] Z. Li, J. Chao, M. Zhang, W. Chen, M. Zhang and G. Fu, "Coupled POS Tagging on Heterogeneous Annotations," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 3, pp. 557-571, March 2017, doi: 10.1109/TASLP.2016.2644262.
- [3] J. R. Bellegarda, "Part-of-Speech Tagging by Latent Analogy," in IEEE Journal of Selected Topics in Signal Processing, vol. 4, no. 6, pp. 985-993, Dec. 2010, doi: 10.1109/JSTSP.2010.2075970.
- [4] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in ECCV, 2016, pp. 354-370.
- [5] Z. Lu, P. Han, L. Wang, and J.-R. Wen, "Semantic sparse recoding of visual content for image applications," IEEE Transactions on Image Processing, vol. 24, no. 1, pp. 176-188, 2015.
- [6] Ulges A, M. Worring, and T. Breuel, "Learning visual contexts for image annotation from Flickr groups," IEEE Transactions on Multimedia, vol. 13, no. 2, pp. 330-341, 2011.
- [7] X. Jing, F. Wu, Z. Li, R. Hu and D. Zhang, "Multi-Label Dictionary Learning for Image Annotation," in IEEE Transactions on Image Processing, vol. 25, no. 6, pp. 2712-2725, June 2016, doi: 10.1109/TIP.2016.2549459.
- [8] Yulei Niu, Zhiwu Lu, Ji-Rong Wen, Tao Xiang, and Shih-Fu Chang, "Multi-Modal Multi-Scale Deep Learning for Large-Scale Image Annotation", IEEE Transactions On Image Processing, Vol. 28, NO. 4, April 2019
- [9] Z. Lu, Z. Fu, T. Xiang, P. Han, L. Wang, and X. Gao, "Learning from weak and noisy labels for semantic segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 3, pp. 486-500, 2017.
- [10] Z. Lu, P. Han, L. Wang, and J.-R. Wen, "Semantic sparse recoding of visual content for image applications," IEEE Transactions on Image Processing, vol. 24, no. 1, pp. 176-188, 2015.
- [11] X. Yu, T. Liu, M. Gong, and D. Tao, "Learning with biased complementary labels," arXiv preprint arXiv: 1711.09535, 2017.
- [12] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 652-663, 2017.