RESEARCH ARTICLE OPEN ACCESS

Towards Describing Visual Explanation Using Machine Learning

Miss. Dhanashree Sh.Vispute, Prof.Amit H.Palve
PG Student, SPPU
Computer Engineering Department,
SITRC, Nashik-422213,India
Assistant Professor, SPPU
Computer Engineering Department,
SITRC, Nashik-422213, India

ABSTRACT

Existing visible explanation generating systems research to easily justify a class prediction. However, they may additionally point out visual parameters attribute which replicate a strong category prior, though the evidence may additionally not clearly be in the image. This is specifically regarding as alternatively such marketers fail in constructing have confidence with human users. We proposed an our very own model which focuses on the discriminating residences of the visible object, jointly predicts a category label, and explains why the predicted label is suitable for the image. The proposed machine robotically an-notates the images the use of hidden Markov model. To annotate images, principles are represented as states through the usage of Hidden Markov model. The parameters of the model are estimated from a set of manually annotated (training) images. Each image in a large check collection is then automatically annotated with the a posteriori chance of concepts present in it.

Keywords :- Visual explanation, Image Description, LSTM , HMM, Sentence generation.

I. INTRODUCTION

As the topic of neural networks is address, it is convenient to indicate the effectiveness of these machines. We are now capable to create software that can classify pictures to specific patterns in videos, high accuracy, notice specific patterns in videos, and examine to play games and a good deal more. Especially the mission of classification in the discipline of visible focus is a very great success story, albeit arguably amongst the easier of the supervised tasks. However, the query of how such a gadget comes to its determination is a ways from understood, for this reason they deficiency the whole lot needed trustworthy. We stay hesitant to follow these fantastically new structures in touchy areas, without stated credibility, - any army equipment, clinical service comes to mind, and possibly even more futuristic functions to softer sciences such as judicial sentencing - where lack of sensing ,incorrect labelling, wrong photograph segmentation, and of the underlying trouble can have forcefully, even fatal, consequences. It is confirm that, any such device that can grant explanations, whilst also performing outstandingly, is preferable to inscrutable systems.

It is necessary to understand, what expression the term systematization encapsulates, as there are one-of-a-kind forms. The common difference chosen for this setup is the division into the two components of introspection and justification. To provide an explanation for outputs by referring to the unique nation the network used to be in and subsequently how the enter traversed the community in terms of its layer activations is provided by introspection. For example, for the classification of an image as 'car' may read: 'The input collective to the fee x, the activation of layer 1 equated to y, and the easiest category which indicated most probability in the output layer was determined for the classification 'car'. So, it is clear that such explanations tackle only human beings with technical cognition. On the other side, a justification tries to connect the visual proof with the output, thereby additionally permitting laymen to apprehend the explanation. An example of this, once more with the 'car' classification, might read:' The image showed the characteristic of a bonnet, four wheels, a steerage wheel, and windows. It's as a result most probable a 'car'..

Our invented technology loss impose that generated sequences fulfill a positive world property, such as class specificity.

II. LITERATURE SURVEY

A. Related Work

Within the artificial brain community Automatic reasoning and rationalization has a long history

[1,13,14,15,16,17,18,19] .Explanation systems content a variety of purposes which consist of robot movements [17],explaining clinical analysis [13], simulator moves [14,15,16,19], and. Many of these structures are rule-based [13] or totally subspitible on filling in a predetermined template [16]. Methods such as [13] require expert -level explanations as well as choice processes. In opposite to, our visual rationalization method is discovered explicitly from data with the benefit of optimizing explanations to fulfill our two proposed visual explanation criteria.

We examine explanations as ciphering out why a sure choice is constant with visible evidence, and differentiate between introspection clarifications systems which explain how a mode determines its last output as properly as justification explanation systems which accountable to produce sentences describing about how visual indication is reliable with output. We listen on justification clarification systems due to the fact such structures may also be greater useful to non-experts who do not have distinctive know-how about contemporary laptop vision systems [1].

We contend that visual explanations must satisfy two criteria: they need to each be type discriminative and precisely describe a precise photo instance. As explanations are awesome from descriptions, which supply a sentence based totally solely on visible information, and definitions, which provide a sentence based only on classification information. Unlike descriptions and definitions, visible explanations element why a certain category is appropriate for a given image while only citing photograph relevant features. As an example, let us reflect on consideration on an photo classification system that predicts a sure image belongs to the classification "western grebe". A standard captioning device provide a description like "This is a giant chook with a white neck and black returned in the water." However, as this description does not mention any discriminative features, it should also be applied to a "laysan albatross" .In contrast, we propose to grant explanations, such as This is a western grebe because this hen has a long white neck, one pointy yellow beak, as nicely as a purple eye." The explanation includes the \red eye" property, e.g., when essential for distinguishing between "western grebe" and "laysanalbatross".In this way our device explains why the predicted class is the most excellent for the image

. ***B. Visual Description***

Early photograph description methods construct on first detecting visual principles in a scene (like subject, verb, and object) before generating a sentence with both a easy language mannequin or sentence guidance [23,24]. A long way exceed such systems and are successful of producing fluent

correct descriptions of images by Recent deep models [7,8,9,10,11,25,26] . Many of these systems study to map from pix to sentences explicitly, with no guidance on intermediate facets .Likewise, our model strive to analyze a visible rationalization given solely an picture and estimated label with no carnal guidance, such as object attributes or phase locations.

C. Fine-grained Classification

Object classification, and fine-grained classification in particular, is engaging to show justification structures due to the fact describing photo content is no longer agreeable for an explanation. on condition that are each class -specific and characterized in the image Explanation models ought to goal. Most fine-grained zero-shot and few-shot picture classification systems use attributes[26] as auxiliary records which can assist visual information. Attributes can be thinking of as a channel conveniently interpretable selection statements which can act as an justification. to distinct a high dimensional feature area into a sequence of easy .

III. PROPOSED METHODOLOGY

A. Problem Definition And Motivation

Many imaginative and prescient methods focus on discovering visual aspects which can assist "justify" an photo classification selection [3,16,6]. These models do now not associate determined discriminative features to herbal language expressions. The techniques discovering discriminative visible facets are complementary to our proposed system. In fact, discriminative visible points may want to be used as more inputs to our model to produce more advanced explanations.

B. System Overview

Following are the details of the proposed work as shown in Fig. 1. Initially modules of the gadget are mentioned and later their detail working is explained.

Modules:

1. Image Preprocessing
2. Feature Extraction
3. Prediction of Class
4. Discriminative Loss
5. LSTM
6. HMM

1. Image preprocessing:

Image preprocessing will be used to preprocess the image to grayscale and extract the pixel values for further processing.

2. Feature Extraction:

Visual features are useful to "Justify" an image classification decision. These models do not associate discovered discriminative features to natural language expressions .Our proposed system are complementary to the methods discovering discriminative visual features .In fact, discriminative visual features could be used as additional inputs to our model to produce better explanations. Workable model of ABAC.

A feature vector is then composed by concatenating the three channel histograms into one vector. For image retrieval, histogram of query image is then matched against histogram of all images in the database using some similarity metric.

A color histogram H for a given image is determine as a vector H = {h[1], h[2], . . . h[i], . . . , h[N]} where i represents a color in the color histogram, h[i] is the number of pixels in color i in that image, and N is the number of bins in the color histogram, i.e., the number of colors in the selected color model.

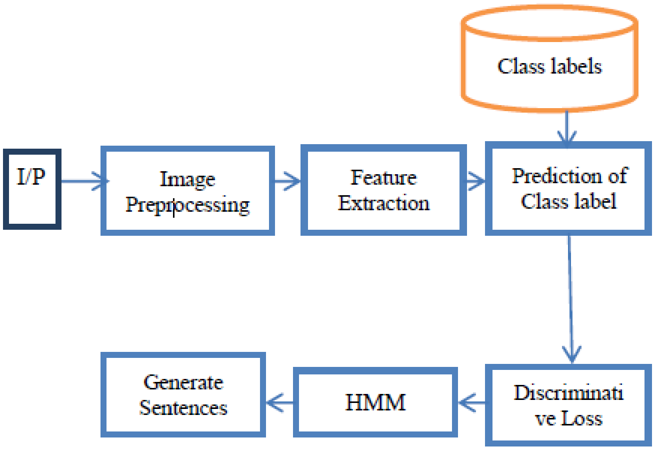


Fig. 1: System Architecture

3. Prediction of class label:

Features are trained with a discriminative loss & enforces that generated sentences contain class specific information. To demonstrate that both class information and the discriminative loss are important, we compare our explanation model to an explanation label mode which is not trained with the discriminative loss, and to an justify discriminative model which is not conditioned on the predicted class. The label predictions are based on class similarity.

Class Similarity: If a sentence fits the definition of a class well, it would have to score high when matched with the target sentences belonging to its label. Therefore, the CIDEr score of this sentence computed against each target sentence in its class and then added together will provide a measure for the similarity with respect to it shown class.

4. Discriminative Loss:

During training A novel discriminative loss acts on sampled word sequences Our loss enables us to enforce global sentence constraints on sentences. We ensure that the final output of our system fulfils, By assigning our loss to sampled sentences .Each training instance contains an category label, image and a ground truth sentence. At the time of training, the model receives the ground truth word for each time step t ∈ T. We define the relevance loss as:

$$L_R = \frac{1}{N}\sum_{n=0}^{N-1}\sum_{t=0}^{T-1}\log p(w_{t+1}|w_{0:t}, I, C)$$

wherewt is a ground truth word, I is the image, C is the category, and N isthe batch size. By training the model to adumbrate each word in a ground truth sentence, the model is trained to create sentences which correspond to image content. However, this loss does not explicitly boost generated sentences to discuss discerning visual properties. A discriminative loss is generated to target sentence generation on discriminative visual properties of an image which are both image relevant and category specific. All references used in the reference list are not cited in the paper .Experiment results are not complete.

5. HMM

Hidden Markov Model (HMM) where the hidden states are related to the (simplified) sentence structure we seek: T = {n1, n2, s, v, p}, and the emissions are related to the observed detections: {n1, n2, s} in the image if they exist.

Proposed HMM is suitable for generating sentences that contain the core components defined in T which produces a sentence of the form NP-VP-PP, which we will show in sec. 4 is sufficient for the task of generating sentences for describing images.

6. LSTM

LSTM (Long Short-Term Memory) are very good for evaluating sequences of values and predicting the adjacent one. For example, LSTM could be a preferred one if you want to predict the very next point of a given time series. Considering about sentences in texts; the **phrases are primary sequences of words**. So, it is natural to consider LSTM could be useful togenerate the next word of a given sentence.

c. Algorithm

Viterbi algorithm for finding optimal sequence of hidden states. Given an observation sequence and an HMM λ = (A,B), the algorithm returns the state path through the HMM that assigns maximum likelihood to the observation sequence. Note that states 0 and qF are non-emitting

```
function VITERBI(observations of len T, state-graph of len N) returns best-path
create a path probability matrix viterbi[N+2,T]
for each state s from 1 to N do ; initialization step
viterbi[s,1]←a0,s ∗ bs(o1)
backpointer[s,1]←0
for each time step t from 2 to T do ; recursion step
for each state s from 1 to N do
viterbi[s,t]← max viterbi[s0,t −1]s
backpointer[qF ,T]← argmax s=1 viterbi[s,T] ; termination step
  return the backtrace path by following backpointers to states back in time from backpointer[qF ,T]
```

D. Data Flow Diagram

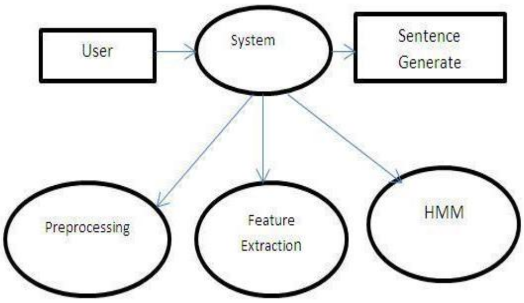
a graphical representation of the ”Stream” of data through an information system, modeling its process aspects is called as adata flow diagram (DFD). Often they are a preliminary step used to create an analysis of the system which can later be elaborated. DFDs can also be used for the visualization of data prepration (structured design). A DFD shows what type of information will be input to and output from the system, where the data will come from and go to, and where the data will be stored. It does not show information about the timing of processes or information about in case processes will operate in sequence or in parallel (which is shown on a flowchart).

If a sentence fits the definition of a class well, it would have to score high when matched with the target sentences belonging to its label. Therefore, the CIDEr score of this sentence computed against each target sentence in its class and then added together will provide a measure for the similarity with respect to its own class.

- ***DFD 0****:* A data flow diagram is graphical representation of flow of data through an information system where modeling its process aspects. Often they are a preliminary step used to create overview of the system.

  DFDs can also user for the visualization of data processing. It shows what kind of information will be input to and output from system.

- ***DFD 1****:* DFD level 1 diagram is the additional information about the major functions of the system. The Level 1 DFD shows how the system is divided into subsystems or processes, that each deals with one or more of the data owns to or from an external agent, and which in sync provide all of the functionality of the system as a whole.

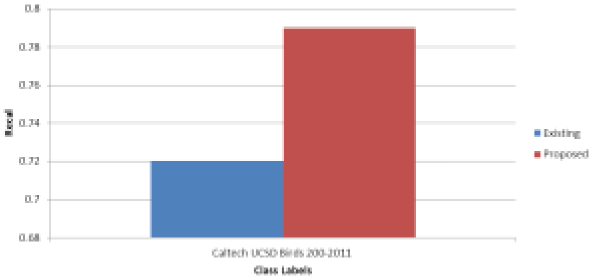


IV. MATHEMATICAL MODEL

S= {D, FE, PL, DL, HM, SG}:
Where: D: Set of Dataset
FE: Feature Extraction
PL: Prediction of Class label
DL: Descriptive Loss
HM: HMM
SG: Sentence Generation
Input:
I1: Set of Image as an input: {q1, q2,qn}
I2: Dataset will also be an input.
Functions:
F1: Image Preprocessing
F2:Feature Extraction.
F3:Prediction of class label.
F4:HMM.
F5: Sentence generation
Output:
O1: Visual Explanation

V.RESULT ANALYSIS

Below table shows sample explanations produced by first outputting a declaration of the predicted class label in terms of recall For the remainder of our qualitative results, we omit the class declaration for easier comparison.



	Existing	Proposed
Caltech UCSD Birds 200-2011	0.72	0.79

A)Dataset used for Experimental Analysis

(CUB) dataset which contains 200 classes of North American bird species and 11,788 images in total.

B) Accuracy Analysis:

Below table shows sample explanations produced by first outputting a declaration of the predicted class label in terms of recall For the remainder of our qualitative results, we omit the class declaration for easier comparison.

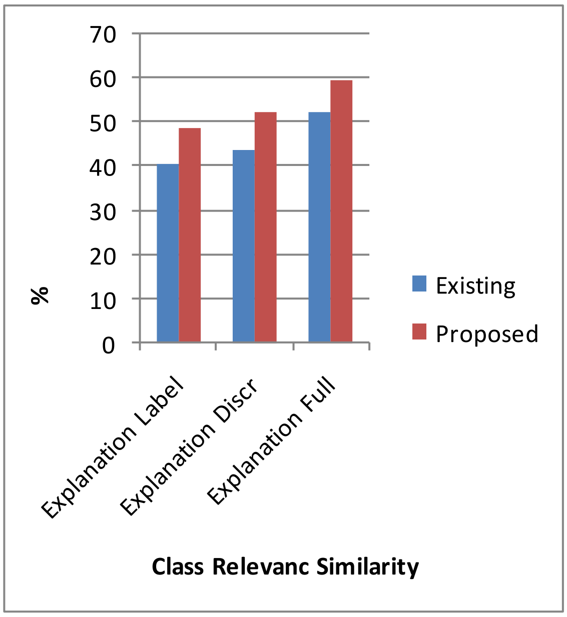
Recall:
In information retrieval, recall is the fraction of the relevant sentence or class labels that are successfully generated.

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

For example, for a sentence or description generation, recall is the number of correct sentence or parts recognized divided by the number of results that should have been returned.

In binary classification, recall is called sensitivity. It can be viewed as the probability that a relevant visual description or sentences is retrieved by the query.

C.)Class relevance similarity:



	Existing	Proposed
Explanation Label	40.86	48.6
Explanation Discr	43.61	52.6
Explanation Full	52.25	59.68

The full model proves to be superior to both baselines in image relevance and class relevance, as well as showing the importance of conditioning on the labels and using the discriminative loss to produce better results overall, demonstrated by its surpassing of both explanation ablations. Also the Explanation Label fairs only marginally better than the definition model, where as the Explanation-Discriminative achieves convincingly higher values in contrast. With respect to class relevance, the

definition model trumps the description model as expected ,and any addition working with class information improves the model, as seen in the consistently better values from both ablation models. Adding the discriminative loss however doesn’t discern between classes as well as when adding the label to the baseline models, as can be seen in column 4row 4 being worse than row 3. Also surprising is that the raw definition baseline comes second best to the grand model ,showing that adding the label and discriminative loss works better in tandem than each alone.

VI. CONCLUSION

To generates natural language descriptions of image regions based on weak labels in form of a dataset of images and sentences, and with very few hardcoded assumptions, we introduced this model. Explanation is an important capability for formation of intelligent systems. Especially as the field of computer vision continues to employ and improve deep models which are not easily interpretable, Visual explanation is a rich research direction.

ACKNOWLEDGMENT

I would sincerely like to thank our Head of Department Prof. (Dr.) Amol Potgantwar, Computer Engineering, and Professor Amit H.Palve , Department of Computer Engineering, SITRC, Nashik for his guidance, encouragement and the interest shown in this project by timely suggestions in this work. His scholarly feedback and expert suggestions had greatly enhanced the effectiveness of this work.

REFERENCES

[1] Biran, O., McKeown, K.: Justification narratives for individual classifications. In:Proceedings of the AutoML workshop at ICML 2014. (2014)

[2] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems.(2012) 1097–1105

[3] Gao, Y., Beijbom, O., Zhang, N., Darrell, T.: Compact bilinear pooling. In:Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). (2016)

[4] Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell,T.: Decaf: A deep convolutional activation feature for generic visual recognition.Proceedings of the International Conference on Machine Learning (ICML) (2013)

[5] Teach, R.L., Shortliffe, E.H.: An analysis of physician attitudes regardingcomputer-based clinical consultation systems. In: Use and impact of computersin clinical medicine. Springer (1981) 68–85

[6] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scalehierarchical image database. In: Computer Vision and PatternRecognition,2009.CVPR 2009.IEEE Conference on, IEEE (2009) 248–255

[7] 7.Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural imagecaption generator. In: CVPR. (2015)

[8] Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S.,Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visualrecognition and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015)

[9] Karpathy, A., Li, F.: Deep visual-semantic alignments for generating image descriptions. In: CVPR. (2015)

[10] Xu, K., Ba, J., Kiros, R., Courville, A., Salakhutdinov, R., Zemel, R., Bengio,Y.: Show, attend and tell: Neural image caption generation with visual

[11] attention.Proceedings of the International Conference on Machine Learning (ICML) (2015)

[12] Kiros, R., Salakhutdinov, R., Zemel, R.: Multimodal

[13] neural language models. In:Proceedings of the 31st International Conference on Machine Learning (ICML-14).(2014) 595–603

[14] Hochreiter, S., Schmidhuber, J.: Long short -term memory. Neural Comput. 9(8)(November 1997) 1735– 1780

[15] Shortliffe, E.H., Buchanan, B.G.: A model of inexact reasoning in medicine. Mathematical biosciences 23(3) (1975) 351–379

[16] Lane, H.C., Core, M.G., Van Lent, M., Solomon, S., Gomboc, D.: Explainable artificial intelligence for training and tutoring. Technical report, DTIC Document(2005)

[17] 15.Core, M.G., Lane, H.C., Van Lent, M., Gomboc, D., Solomon, S., Rosenberg, M.:Building explainableartificial

[18] intelligence systems. In: Proceedings of the national conference on artificial intelligence. Volume 21., Menlo Park, CA; Cambridge, MA;London; AAAI Press; MIT Press; 1999 (2006) 1766.

[19] 16.Van Lent, M., Fisher, W., Mancuso, M.: An explainable artificial intelligence system for small-unit tactical behavior. In:

[20] PROCEEDINGS OF THE

[21] 17.Lomas, M., Chevalier, R., Cross II, E.V., Garrett, R.C., Hoare, J., Kopack, M.: Explaining robot actions. In:

[22] Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction, ACM (2012) 187–188

[23] 18.Lacave, C., D´ıez, F.J.: A review of explanation methods for Bayesian networks. The Knowledge

[24] Engineering Review 17(02) (2002) 107–127

[25] 19.Johnson, W.L.: Agents that learn to explain themselves. In: AAAI. (1994) 1257–1263.

[26] 20.Berg, T., Belhumeur, P.: How do you tell a blackbird from a crow? In: Proceedings of the IEEE International

[27] Conference on Computer Vision. (2013) 9–16.

[28] 21.Jiang, Z., Wang, Y., Davis, L., Andrews, W., Rozgic, V.: Learning discriminative features via label consistent

[29] neural network. arXiv preprint arXiv:1602.01168(2016)

[30] 22.Doersch, C., Singh, S., Gupta, A., Sivic, J., Efros, A.: What makes paris look likeparis? ACMTransactions on

[31] Graphics 31(4) (2012).

[32] 23.Kulkarni, G., Premraj, V., Dhar, S., Li, S., choi, Y., Berg, A., Berg, T.: Baby talk: understanding and

[33] generating simple image descriptions. In: CVPR. (2011)

[34] 24.Guadarrama, S., Krishnamoorthy, N., Malkarnenkar, G., Venugopalan, S., Mooney,R., Darrell, T., Saenko, K.:

[35] Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero -shot

[36] recognition. In: Proceedings of the IEEE International Conference on Computer Vision. (2013) 27122719

[37] 25.Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Doll´ar, P., Gao, J., He, X., Mitchell, M., Platt, J.C., et al.: From captions to visual concepts and back. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 1473–1482.

[38] 26. Lampert, C., Nickisch, H., Harmeling, S.: Attribute-based classification for zero shot visual object categorization. In: TPAMI. (2013)

[39] 27. Reed, S., Akata, Z., Lee, H., Schiele, B.: Learning deep representations of fine grained visual descriptions.

[40] In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)