Semantic Relationship Expansion for Image-Text Matching

Ms. Shaik Heena #1, Mr. Venigalla Vamsi Krishna #2, Ms. Padarthi Aswani #3,
Ms. Nalluri Haritha #4, Mr. Eemani Venkata Sai Ram #5,

#1 Assistant Professor, Dept Of CSE, Qis College of Engineering and Technology, Ongole, Prakasam (Dt)
#2 Student, Dept Of CSE, Qis College of Engineering and Technology, Ongole, Prakasam (Dt)
#3 Student, Dept Of CSE, Qis College of Engineering and Technology, Ongole, Prakasam (Dt)
#4 Student, Dept Of CSE, Qis College of Engineering and Technology, Ongole, Prakasam (Dt)
#5 Student, Dept Of CSE, Qis College of Engineering and Technology, Ongole, Prakasam (Dt)

Abstract: Image-text matching has been a hot research topic bridging the vision and language areas. It remains challenging because the current representation of image usually lacks global semantic concepts as in its corresponding text caption. To address this issue, we propose a simple and interpretable reasoning model to generate visual representation that captures key objects and semantic concepts of a scene. Specifically, we first build up connections between image regions and perform reasoning with Graph Convolutional Networks to generate features with semantic relationships. Then, we propose to use the gate and memory mechanism to perform global semantic reasoning on these relationship-enhanced features, select the discriminative information and gradually generate the representation for the whole scene. Experiments validate that our method achieves a new state-of-the-art for the image-text matching on MS-COCO and Flickr30K datasets. It outperforms the current best method by 6.8% relatively for image retrieval and 4.8% relatively for caption retrieval on MS-COCO (Recall@1 using 1K test set). On Flickr30K, our model improves image retrieval by 12.6% relatively and caption retrieval by 5.8% relatively (Recall@1).

Index Terms—Fundamental visual concept, neighboring concept distributing, heterogeneous media

I. INTRODUCTION

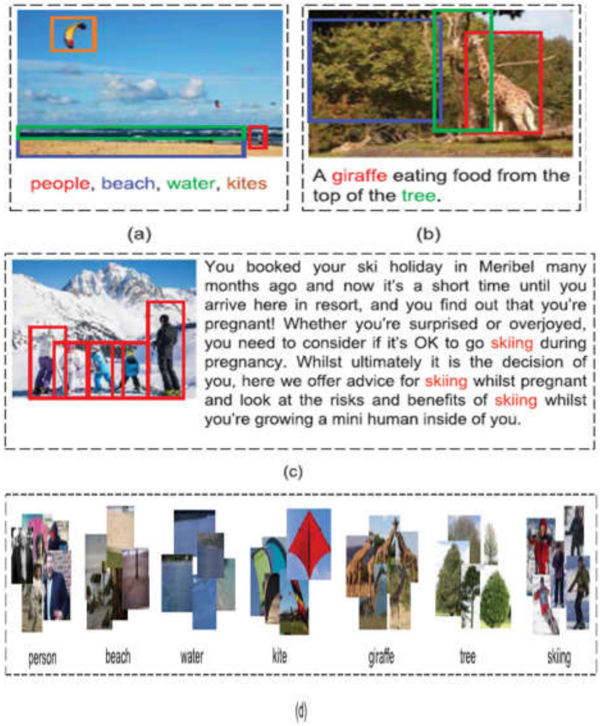
THE worldwide web is full of images, videos, audio, and text, which are not only growing rapidly in terms of quantity but are also becoming increasingly rich in terms of content. Heterogeneous web data usually coexist in multimedia documents and use similar semantics to describe the same subject from different perspectives. The various modalities of documents may be complementary in terms of expressing the semantics of content.

For example, an image can vividly inspire imagination but incompletely describe a concept. In contrast, while text can accurately describe the details of a concept, it is not intuitive enough. Currently, many real-world Internet applications involve multi-modal data, such as cross media retrieval [1], [2], [3], [4], image tagging [5], multimedia searching [6] and multimedia caption generation [7]. Common to these applications, the relations between different modalities need to be considered and learned at the document level granularity under the supervision of labeled data.

In this work, we focus on a problem that is different from traditional cross-media learning problems. Suppose we have a set of multimedia documents, each including an image and a textual description in the form of keywords, sentences or paragraphs, as shown in Fig. 1. Each image generally consists of a few visual patches, each of which can be visually represented simply and has a single semantics; the correlated textual description consists of meaningful keywords, and each keyword can be considered as a concept label of the visual patches. In this work, we aim to learn the concept label of each visual patch under the supervision at the granularity of images and textual documents. As shown in Fig. 1(d), a set of visual patches with the same semantics and their corresponding concept label make up a fundamental visual concept, and the concepts compose the complex web data. It can be envisioned that, for a mass of correlated heterogeneous documents, computers can automatically learn the fundamental concepts that compose the data describing our world without any supervision at the granularity of the fundamental concepts.

The fundamental visual concepts achieved can be used in many applications including multimedia search, recommendation and annotation, without the expensive cost of labeling. The problem has several characteristics: 1) it does not need a pre trained concept detector or classifier for each concept; 2) it allows concepts to be continuously learned from increasingly complex data (e.g., from “image + keywords” to “image + paragraph”). Based on the above analysis, we consider that the task in this work has some differences from the following related problems:

- *Multi-instance multi label learning* (MIML) [8], [9], [10]: MIML is a learning paradigm where each example is simultaneously represented by a bag of instances and associated with a set of class labels. Most MIML approaches aim to predict the labels of new bags instead of instances.
- *Image annotation*: In general, the task of the image annotation is to predict multiple textual labels that describe the content or visual appearance of an unseen image [11].



A giraffe eating food from the top of the tree.

You booked your ski holiday in Meribel many months ago and now it's a short time until you arrive here in resort, and you find out that you're pregnant! Whether you're surprised or overjoyed, you need to consider if it's OK to go skiing during pregnancy. Whilst ultimately it is the decision of you, here we offer advice for skiing whilst pregnant and look at the risks and benefits of skiing whilst you're growing a mini human inside of you.

Fig. 1. Illustration of the task of fundamental visual concept learning from correlated images and text the correlated data include multiple increasingly complex cases. e.g., (a) Image + keywords. (b) Image + sentence. (c) Image + paragraph. The task aims to learn the concepts from the correlated media, without the classifiers or detectors trained at the granularity of concepts.

In addition, a few studies focused on predicting the labels of the regions in images [12].

- *Fined-grained image classification* (FGIC): FGIC usually involves classifying the subclasses of objects belonging to the same class.

In each class, objects of different subclasses are both semantically and visually similar to each other [13]. This paper formulates the task of learning fundamental visual concepts from correlated images and text in the form of keywords and sentences and proposes an approach named *neighboring concept distributing* (*NCD*) to address this task. In this work, visual patches and text descriptions are represented based on deep networks and one-hot vectors, respectively. The NCD approach models all data using a concept graph, which considers the visual patches as nodes and generates inter image edges between visual patches belonging to different images and intra-image edges between visual patches in the same image.

The concept label is distributed from the images to the visual patches and propagated across the latter based on measurements over the concept graph, including fitness, distinctiveness, smoothness and sparseness. Based on the proposed approach, we can learn the fundamental visual concepts that compose multimedia documents from correlated images and text. In summary, our contributions are three-fold: (1) we introduce and formulate the problem of fundamental visual concept learning from correlated images and text, which is different from current learning problems such as MIML, image annotation and FGIC; (2) we present a neighboring concept distributing approach to this problem, which models the data as a concept graph, distributes concept labels from images to visual patches and propagates them across the patches over the concept graph; and (3) we analyze the learn ability of the proposed approach and find that, under some conditions, all concepts can be correctly learned with the probability $1 - \delta$ when the amount of data M is larger than $O(\ln 1/\delta)$, i.e., with an arbitrarily high probability as the amount of data increases.

The rest of this paper is organized as follows. Section II presents a brief overview of related work. Section III briefly introduces the representation of correlated images and text. Section IV describes our approach to fundamental visual concept learning. Section V discusses the learn ability analysis of the proposed approach. Section VI provides the experimental results, and Section VII concludes the paper.

II. RELATED WORK

Image-Text Matching Our work is related to existing methods proposed for image-text matching, where the key issue is measuring the visual-semantic similarity between a text and an image. Learning a common space where text and image feature vectors are comparable is a typical solution for this task. Frome et al. [13] propose a feature embedding framework that uses Skip-Gram [21] and CNN to extract feature representations for cross-modal. Then a rank ing loss is adopted to encourage the distance between the mismatched image-text pair is larger than that between the matched pair. Kiros et al. [17] use a similar framework and adopt LSTM [12] instead of Skip-Gram for the learning of text representations. Vendrov et al. [20] design a new objective function that encourages the order structure of visual semantic can be preserved hierarchy.

Attention Mechanism. Our work is also inspired by bottom-up attention mechanism and recent image-text matching methods based on it. Bottom-up attention [16] refers to salient region detection at stuff/object level can be analogized to the spontaneous bottom-up attention that is consistent with human vision system [16]. Similar observation has motivated other existing work. In [15], R-CNN [7] is adopted to detect and encode image regions at object level. Image-text similarity is then obtained by aggregating all word-region pair's similarity scores. Huang et al. [14] train a multi-label CNN to classify each image region into multi-labels of objects and semantic relations; so that the improved image representation can capture semantic concepts within the local region. However, to the best of our knowledge, no study has attempted to incorporate global spatial or semantic reasoning when learning visual representations for image-text matching.

Relational Reasoning Methods Symbolic approaches [22] are the earliest form of reasoning in artificial intelligence. In these methods, relations between symbols are represented by the form of logic and mathematics, reasoning happens by abduction and deduction [11] etc. However, in order to make these systems can be used practically; symbols need to be grounded in advance. More recent methods, such as path ranking algorithm [18], perform reasoning on structured knowledge bases by taking use of statistical learning to extract effective patterns. As an active research area, graph-based methods [21] have been very popular in recent years and shown to be an efficient way of relation reasoning. Graph Convolution Networks (GCN) [18] is proposed for semi-supervised classification. Yao et al. [19] train a visual relationship detection model on Visual Genome dataset [21] and use a GCN-based encoder to encode the detected relationship information into an image captioning framework.

III. MULTIMEDIA REPORT REPRESENTATION

We awareness on the internet multimedia document Di that contains a visual photograph Ii and a corresponding textual description Ti , which are assumed to have the same semantics in describing the net multimedia document. For the textual modality, if the textual description Ti is within the form of sentences, we take away forestall phrases and choose key phrases to represent it

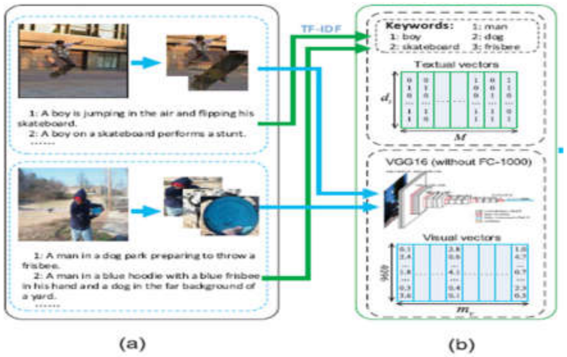
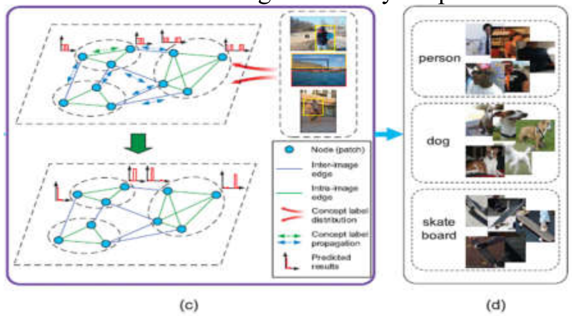

TF-IDF
1: A boy is jumping in the air and flipping his skateboard.
2: A boy on a skateboard performs a stunt.
......
1: A man in a dog park preparing to throw a frisbee.
2: A man in a blue hoodie with a blue frisbee in his hand and a dog in the far background of a yard.
......
Keywords:
1: boy
2: skateboard
1: man
2: dog
3: frisbee
Textual vectors
d_t
M
VGG16 (without FC-1000)
Visual vectors
4096
m_v
(a)
(b)

Fig. 2. The framework for fundamental visual concept learning. (a) The input data are correlated images and text. In this work, we consider only text in the form of keywords and sentences. (b) Sentences are transform to multiple meaningful keywords, which are represented by one-hot vectors, and visual patches are represented with the feature vectors generated by deep networks.



(c) In the concept graph, each node denotes a visual patch, and all nodes enclosed in a dashed ellipse belong to an image. The green and blue double-headed arrows denote concept label propagation across the visual patches within an image and across different images, respectively. The histograms shown next to the nodes denote the prediction of the concept labels for corresponding patches. (d) A fundamental visual concept is comprised of a concept label (or keyword) with a set of corresponding representative visual patches.

based on TF-IDF. We do no longer carry out any technique if T*i* is in the shape of keywords. Subsequently, the textual description can be represented by using n_i keywords $T_i = \{w^1_i, w^2_i\ldots w^{ni}_I\}$ without attention of the contextual information. For a corpora of internet multimedia documents $\{D_i\}^M_{i=1}$, we have a vocabulary that carries d_t distinct key phrases, i.e., $\{w_1,w_2,\ldots,w_{dt}\}$, in which each keyword may be taken into consideration as a concept label.

An image Ii can generally be segmented into mi visible patches and represented by way of $Ii = \{p^1_i, p^2_i,\ldots, pm^i_I\}$ without attention of the contextual data. We extract the visible feature vector $p^j_I \in R^{dv}$ for each patch p^j_i primarily based on the VGG internet-16 deep network [17] pre-trained at the ImageNet dataset. We use the visible patches resized to 224×224 because the input and constitute them with the 4096-dimensional feature vectors received from the second one absolutely related layer of VGG internet-sixteen. Within the experiments, we undertake the segmentation results given inside the public datasets, and choose the equal variety of keywords from the sentences as the patches for a photograph, i.e., $n_i = m_i$.

IV. FUNDAMENTAL VISIBLE IDEA GAINING KNOWLEDGE

On this phase, the problem system is first given for essential visual concept studying and then followed with the aid of the proposed NCD method to this trouble.

A. TROUBLE COMPONENTS

We first give the outline of the essential visible idea (FVC): FVC refers to a set of photo patches with the identical semantics that may be categorized by means of an idea label, in which each patch normally has a simple visual appearance.

Given a corpora of photograph-text paired internet files D = {Di}M $_{i=1}$, wherein simplest the correspondence at the granularity of photos and textual descriptions is to be had, the key problem is to estimate the concept label of every visual patch p^j_i (j = 1, 2, · · · ,mi) and can be formulated as follows:

$$\hat{w}^j_i = C_{fvc}(p^j_i)$$

wherein C^f_{vc} denotes a change from visible patches to idea labels. Furthermore, primarily based at the end result of problem (1), we can acquire all corresponding visible patches for a given idea label w^j_i . A learned FVC ck with concept label wk can be defined within the form of

$$c_k = .w_k\,,\ \{(\,p_i\,,\,conf_i\,)|C_{f\,vc}(\,p_i\,) = w_k\,\})$$

where con fi denotes the confidence that the visual patch pi is mapped to the idea label wk . All learned standards can be recorded as C = {ck} of the cardinality |C| = dt .

B. CONCEPT GRAPH CREATION

Distinct from not unusual graph-based totally studying paintings, we need to distribute the semantic facts from pix (a subset of nodes) to every patch (node). The traditional kNN graph used formerly isn't very powerful right here. In these paintings, we assemble a kNN-like graph for our idea graph. The concept graph consists of two kinds of edges: inter-picture edges and intra-picture edges, which confer with the rims between nodes belonging to 2 exceptional pix and to the identical photograph, respectively. First, we introduce a way of producing inter-photograph edges. To better distribute the semantic information throughout unique photographs, we generate inter-picture edges by means of measuring the distinguish ability between two nodes as follows:

$$Dist(p^k_i, p^l_j) = \phi_{inter}(p^k_i, p^l_j) + \lambda_I \phi_I(I_i, I_j; p^k_i, p^l_j) \\ + \lambda_T \phi_T(T_i, T_j),$$

A huge distinguish ability with recognize to p^l_j and p^k_i states that there is an excessive similarity among both patches and a big divergence among associated photograph-textual content report pairs. For a given patch, we pick out okay patches with the biggest distinguish ability from the other pics because the inter-photo neighboring nodes Ninter (·). Over the concept graph, the semantic facts will be dispensed down to patches from photos greater efficaciously and correctly, as discussed in the following phase. For patches belonging to the same picture, the similarity between them is described as follows:

$$\phi_{intra}(p^k_i, p^l_i) = \frac{\exp(\mathbf{p}^{k\,T}_i \mathbf{p}^l_i)}{\sum_z \exp(\mathbf{p}^{k\,T}_i \mathbf{p}^z_i)}$$

wherein φintra(·, ·) is described by means of the softmax characteristic that maps all values into (0, 1) and tends to push them to 0 due to its convexity. Typically, an picture includes only a few visible patches, and we recall all of them because the intra-photograph neighboring nodes Nintra(·) for every different. The load of every part is defined as follows:

$$E^{(1)}_{ij} = \begin{cases} \phi_{inter}(v_i, v_j) & if\ \ v_j \in \mathcal{N}_{inter}(v_i) \\ 0 & otherwise \end{cases}$$

$$E^{(2)}_{ij} = \begin{cases} \phi_{intra}(v_i, v_j) & if\ \ v_j \in \mathcal{N}_{intra}(v_i) \\ 0 & otherwise \end{cases}$$

C. NEIGHBORING CONCEPT DISTRIBUTING OVER IDEA GRAPHS

As defined above, T = (Ti) ∈ $R^{dt\times M}$ and C ∈ $R^{dt\times mv}$ represent the actual concept label matrix for pics and the expected concept label matrix for all patches, respectively. Three important phrases need to be considered in neighboring idea distributing to attain our intention. The anticipated label of each patch is predicted to be close to its true label. However, the latter is unknown in our assignment. Subsequently, we first introduce a term called fitness to measure the difference among the summation of the anticipated outcomes over all patches belonging to an image and the actual concept label matrix of this picture. An excellent answer for the prediction requires excessive health. We outline the fitness as follows:

$$\sum_{i=1}^{M} \|\mathbf{T}_i - \sum_{k=1}^{m_v} O_{ik}\mathbf{c}_k\|^2 = \|\mathbf{T} - \mathbf{C}\mathbf{O}^T\|^2_F,$$

in which O = (O_{ik}) ∈ $R^{M\times m}{}_v$, in which O_{ik} = 1 if the node vk belongs to the photo Ii and O_{ik} = zero in any other case. As shown in Eq. (8), an awesome expected result at patch-degree granularity ends in a small distinction among the anticipated and the true concept label matrices at photograph-stage granularity. 2nd, we study that, for natural photographs, the patches inside the same photo have a tendency to have different semantics due to the rich and diverse content material contained in the pix. Subsequently, the predicted idea label of a patch is predicted to be different from those of the alternative patches within the equal picture. We introduce a term referred to as strong point to degree the difference. Over all snap shots, the uniqueness term can be formulated by

$$\sum_{i \sim j, i \neq j} \left(\mathbf{c}_i^T \mathbf{c}_j - E_{ij}^{intra}\right)^2 = \|(\mathbf{C}^T\mathbf{C} - \mathbf{E}) \circ \mathbf{Q}\|_F^2,$$

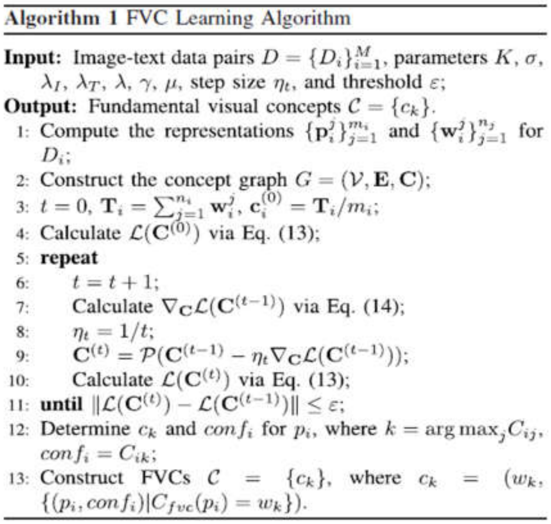
in which i ~ j method that the nodes vi and v j correspond to the equal picture, the operator denotes the Hadamard product of two matrices, and Q ∈ $R^{mv \times m}{}_{v}$, in which the access $Q^{i}{}_{j}$ = 1 if the nodes vi and v_j (i ≠ j) correspond to the identical picture and Q_{ij} = 0 otherwise. Eq. (nine) reinforces the distinctiveness of semantics because of the convexity of the softmax characteristic in Eq. A small price of the uniqueness term states that every patch can have a anticipated idea label this is very unique from those of the others inside the same image if their visible representations are distinct from every different, which facilitates uncouple the textual semantics in the idea dispensing from photos to patches.

D. OPTIMIZATION

To remedy the optimization problem (12), we introduce the projected sub gradient method to clear up it. First, we want to compute the sub gradient of L(C):

$$\nabla_{\mathbf{C}}\mathcal{L}(\mathbf{C}) = 2(\mathbf{C}\mathbf{O}^T - \mathbf{T})\mathbf{O} + 4\lambda((\mathbf{C}^T\mathbf{C} - \mathbf{E}) \circ \mathbf{Q})(\mathbf{C} \circ \mathbf{Q}) \\ + 2\gamma\,\mathbf{C}\mathbf{L} + \mu\,\Delta, \quad s.t.\ \mathbf{C} \geq 0,$$

wherein Δ ∈ $R^{dt \times m}{}_{v}$ is defined as Δ_{ij} = sgn(C_{ij}), sgn(z) outputs 1 when z > zero, −1 when z < 0, and a random quantity uniformly distributed among −1 and +1 whilst z = zero.



Algorithm 1 FVC Learning Algorithm

```
Input: Image-text data pairs $D = \{D_i\}_{i=1}^{M}$, parameters $K$, $\sigma$,
    $\lambda_I$, $\lambda_T$, $\lambda$, $\gamma$, $\mu$, step size $\eta_t$, and threshold $\varepsilon$;
Output: Fundamental visual concepts $\mathcal{C} = \{c_k\}$.
Compute the representations $\{\mathbf{p}_i^j\}_{j=1}^{m_i}$ and $\{\mathbf{w}_i^j\}_{j=1}^{n_j}$ for $D_i$;
Construct the concept graph $G = (\mathcal{V}, \mathbf{E}, \mathbf{C})$;
$t = 0$, $\mathbf{T}_i = \sum_{j=1}^{n_i} \mathbf{w}_i^j$, $\mathbf{c}_i^{(0)} = \mathbf{T}_i/m_i$;
Calculate $\mathcal{L}(\mathbf{C}^{(0)})$ via Eq. (13);
repeat
    $t = t + 1$;
    Calculate $\nabla_{\mathbf{C}}\mathcal{L}(\mathbf{C}^{(t-1)})$ via Eq. (14);
    $\eta_t = 1/t$;
    $\mathbf{C}^{(t)} = \mathcal{P}(\mathbf{C}^{(t-1)} - \eta_t \nabla_{\mathbf{C}}\mathcal{L}(\mathbf{C}^{(t-1)}))$;
    Calculate $\mathcal{L}(\mathbf{C}^{(t)})$ via Eq. (13);
until $\|\mathcal{L}(\mathbf{C}^{(t)}) - \mathcal{L}(\mathbf{C}^{(t-1)})\| \leq \varepsilon$;
Determine $c_k$ and $conf_i$ for $p_i$, where $k = \arg\max_j C_{ij}$,
    $conf_i = C_{ik}$;
Construct FVCs $\mathcal{C} = \{c_k\}$, where $c_k = (w_k,$
    $\{(p_i, conf_i) | C_{fvc}(p_i) = w_k\})$.
```

here, we use the following definition:

P(x) = max(zero, x) (17)

inside the experiments, we choose the step length ηt = 1/t. The algorithm is summarized in algorithm 1.

Earlier than the optimization method, we remember that the semantics of each patch randomly corresponds to 1 idea label of the image to which the patch belongs. Consequently, we initialize the iterative algorithm as follows: $c^{(0)}{}_{I}$ = Ti/mi . Based totally on the optimization set of rules, the price of the element in *ci* that corresponds to the genuine idea label will improve and the others will lower primarily based on the information distribution and propagation over the idea graph.

V. EXPERIMENTAL EFFECTS AND EVALUATION

In this phase, we examine the effectiveness of the proposed NCD method over three widely used public datasets and record the as compared outcomes.

A. EXPERIMENTAL SETUP

The experiments are conducted on 3 datasets: MSRC, VOC20121 and MSCOCO. The MSRC dataset contains 591 pictures with 23 concepts. The VOC2012 dataset incorporates 17,a hundred twenty five pictures with 20 concepts. The MSCOCO dataset contains eighty two, 783 snap shots with 80 concepts, and every image is segmented into multiple rectangle patches and described by using distinctive numbers of tags and 5 sentences. The patches similar to a concept have been separated with the aid of area masks or bounding containers within the datasets. The patch-degree tags are used within the testing technique as ground reality for evaluating the overall performance. As described in phase III, the textual functions are represented by one-warm vectors. The size of textual features are dt = 23, 20 and eighty for MSRC, VOC2012 and MSCOCO, respectively. For photographs, each visible patch is first resized to a set size and then represented as a 4096-dimensional visual feature vector (i.e., dv = 4096) via VGG net-16 deep networks.

1) The kNN *Algorithm*: For each patch within an photo, we first pick out its okay nearest neighboring patches from the other photos primarily based on the visible similarity φinter (·, ·) described in phase IV.B and then depend the range of occurrences of every key-word related to snap shots that encompass as a minimum 1 of the okay acquaintances. The key-word with the maximum occurrences is chosen because the idea label of the given patch. Primarily based at the experiments, we select k = 8 for MSRC, and okay = 12 for each VOC2012 and MSCOCO.

2) M-E Graph: The M-E Graph constructs a multi-area graph, wherein each node represents an image, and resolves an optimization hassle by means of the slicing plane method. EM-MIML

3) EM-MIML: considers patch-level label records as hidden variables and maximizes the probability of image-degree labels given the visible features with the expectation-maximization set of rules. We enforce the set of rules based at the identical visible illustration and key phrases with our work. Deep-MIML

4) Deep-MIML: goals to be expecting the idea labels of visible patches from photograph-degree textual annotations primarily based on a changed VGG-16 architecture. We take a look at the overall performance of Deep-MIML on the MSCOCO dataset based on the released code.2

5) MIMLfast: MIMLfast constructs a low-dimensional subspace shared through all labels and trains label-nique linear fashions to optimize the approximated rating loss through stochastic gradient descent. MIMLfast is a fast algorithm to deal with big datasets inside the multi-instance multi-label undertaking. We use precision, don't forget and F1-score to assess the overall performance of fundamental visual idea studying. The metrics are defined as follows

$$precision = \frac{\#\, of\ correctly\ labeled\ patches}{\#\, of\ totally\ labeled\ patches}$$

$$recall = \frac{\#\, of\ correctly\ labeled\ patches}{\#\, of\ patches\ of\ the\ same\ concept}$$

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

The metrics are described for each concept. The overall performance over an entire dataset is measured via the common on all concepts.

B. PARAMETER TUNING

For the MSRC and VOC2012 datasets, we use approximately 1/3 of the statistics because the validation set, and we use the three-fold go-validation for parameter determination. Regarding the MSCOCO dataset, we use the validation set for parameter determination. In total, we sequentially optimize the parameters to reap the (neighborhood) most reliable solution.

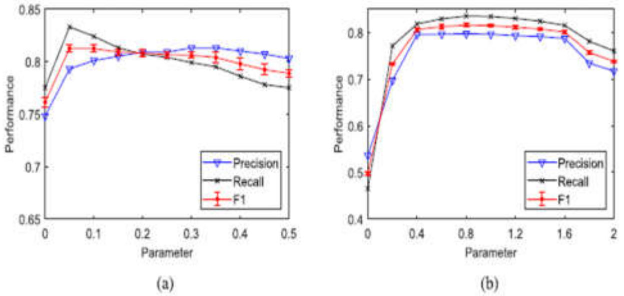


Fig. 3. The effect of parameters λ, γ , and μ in Equation (13) on the performance of the proposed NCD over the MSRC dataset. Each subfigure demonstrates the change of the learning performance with the different values of a parameter given the other two optimal parameters. (a) λ for distinctiveness term. (b) γ for smoothness term.

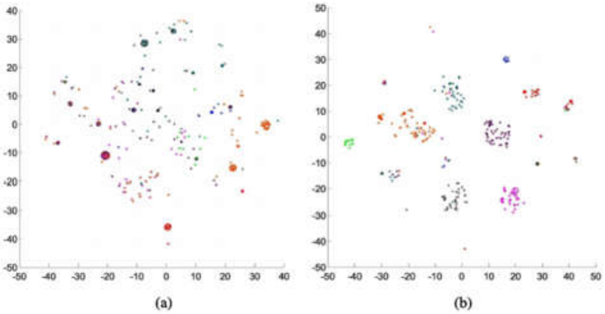


Fig. 4. Distribution of the concept labels learned with the NCD approach on the MSRC dataset in a 2-dimensional space derived by t-SNE. We show the total 939 patches of 8 concepts, i.e., 'bird', 'boat', 'building', 'face', 'sky', 'road', 'grass' and 'tree', instead of all 23 concepts of SRC due to distinguish ability of colors. In this figure, each point denotes a visual patch, and its location and color represent the predicted concept label vector and true concept label, respectively. The clustered points with the same color indicate a good performance of concept learning. In the right subfigure, the three components in a triple shown next to a cluster denote the concept label, number of patches and F1-score of the corresponding concept, respectively. (a) Before update. (b) After 12 updates.

the validation set of MSRC for the NCD technique. In every subfigure, we alternate one parameter even as solving the other parameters to the highest quality values. Fig. 4(a) illustrates the alternate of gaining knowledge of outcomes because the parameter λ will increase. We take a look at that a small cost of λ = zero.05 can attain the satisfactory overall performance. In a extra exact test, we discover that λ = 0.06 can cause a barely better result than λ = zero.05, and we ultimately select the value of zero.06. Fig. 4(b) suggests the impact of the smoothness measured by using γ on the studying performance.

C. Overall Performance And Analysis

Due to randomness within the gradient of $|| \bullet ||_1$, i.e., the matrix Δ in Eq, we perform 5 unbiased experiments for our technique and record the average overall performance. In each optimization system, we set the quantity of iterations to twenty to acquire the convergence in preference to setting the convergence threshold ε.

1) The replace in concept mastering: We first illustrate how the idea gaining knowledge of result modifications with the aid of iteration inside the concept area the use of the proposed NCD technique. In Fig. five, we pick out 939 visible patches from eight principles, i.e., 'chicken', 'boat', 'building', 'face', 'sky', 'avenue', 'grass' and 'tree', rather than all 23 standards from the MSRC dataset due to the distinguish ability of colors. In Fig. 5 (a), the concept label of a patch is initialized consistent with the key phrases of the image to which the patch belongs, e.g., [0.25, 0.25, 0.25, 0.25, 0, · · ·, 0] T, as brought in the method of NCD in phase III.D.

TABLE I
LEARNING PERFORMANCE (PERCENT) OF TOTAL 23 CONCEPTS IN MSRC WITH THE NCD APPROACH

Concept	aero plane	build ing	moun tain	bicy cle	flow er	chair	grass	horse	sheep	water	bird	boat	body	book	face	road	sign	tree	car	cat	cow	dog	sky
Precision	73.2	80.0	14.3	97.0	89.2	100.0	86.3	50.0	100.0	80.0	100.0	77.5	35.1	87.5	41.5	92.0	96.9	83.2	88.0	100.0	97.7	96.8	70.1
Recall	100.0	78.4	8.0	100.0	94.3	100.0	97.6	66.7	100.0	42.7	100.0	93.9	29.4	100.0	37.3	68.9	100.0	70.2	100.0	100.0	95.6	100.0	98.1
F1	84.5	79.2	10.3	98.5	91.7	100.0	91.6	57.1	100.0	55.7	100.0	84.9	32.0	93.3	39.3	78.8	98.4	76.1	93.6	100.0	96.6	98.4	81.8

most factors with the equal hues are assembled into multiple companies, this means that that the equal concept labels tend to be allotted and propagated to visible patches that virtually have the equal semantics (denoted by the identical colour). After convergence, we discover that most visual patches with the identical semantics are placed near together, because of this that the very near concept labels were allotted to them.

2) Gaining knowledge of outcomes of every concept: We test the learning overall performance of our NCD method on three datasets and document the precision, recollect and F1-score of each concept. Desk I suggests the mastering effects over all 23 standards of the MSRC dataset. As proven inside the table, most concepts may be found out nicely, e.g., 'fowl', 'cat', and 'chair'. From the dataset, we discover that these standards are distinctly simple and represented sincerely by means of the visual patches. We also have a look at that a few principles, consisting of 'frame', 'face' and 'mountain', achieve low overall performance (under forty %). From the picture examples, we note that patches related to these standards have a big variation of appearance, which ends up in the shortage of edges between the corresponding nodes and influences the concept label distributing amongst those nodes. It needs to be noted that there is no label to coerce so that the patches with the equal semantics are close in the function space.

TABLE II
PERFORMANCE (PERCENT) COMPARISON OF DIFFERENT APPROACHES ON THE MSRC, VOC2012 AND MSCOCO DATASETS

Algorithms	MSRC			VOC2012			MSCOCO		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
KNN	74.5	63.2	68.4	79.3	67.5	72.9	69.3	40.9	51.4
M-E Graph ([32], 2010)	72.0	-	67.0	-	-	-	-	-	-
EM-MIML([10], 2015)	72.6	73.0	72.8	78.6	74.9	76.7	59.8	60.6	60.2
Deep-MIML([12], 2017)	-	-	-	-	-	-	70.2	57.0	62.9
MIMLfast([46], 2018)	66.0 (Accuracy)			-	-	-	-	-	-
NCD (without fitness)	38.3	18.2	24.7	64.4	18.2	28.4	58.2	32.7	41.9
NCD (without distinctiveness)	73.8	76.5	75.1	86.6	74.7	80.1	61.5	56.3	58.8
NCD (without smoothness)	52.4	45.5	48.7	82.0	63.1	71.3	63.8	30.4	41.2
NCD (without sparseness)	77.2	81.1	79.1	88.5	76.5	82.1	64.3	59.1	61.6
NCD	79.8	81.9	80.8	92.4	78.8	85.1	68.2	62.4	65.2

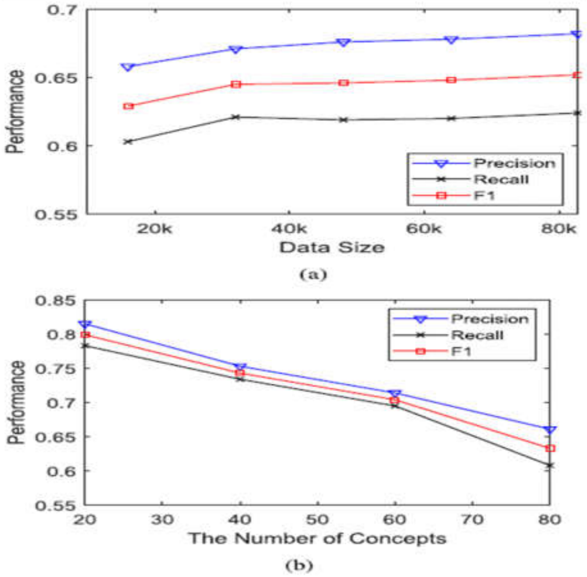
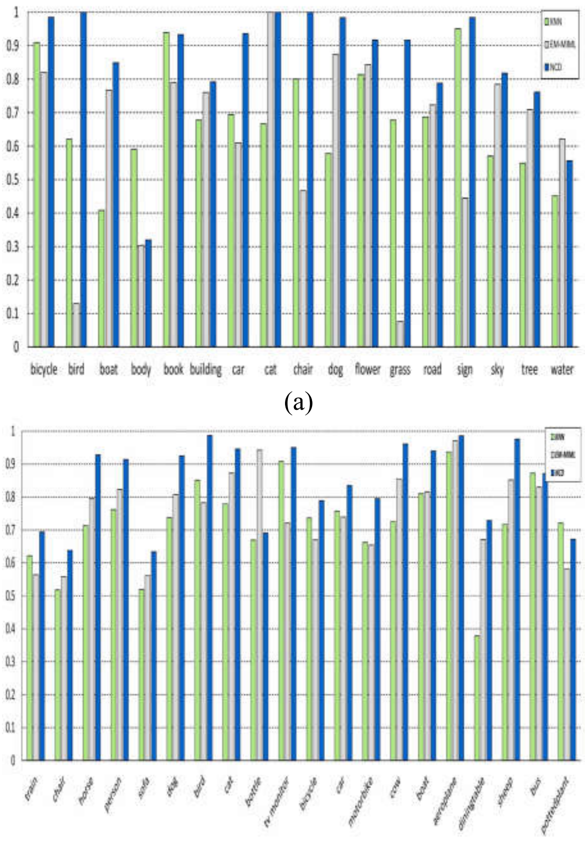


Figure 5: Learning performance with (a) different data sizes and (b) different numbers of concepts. In subfigure (a), we report the learning performance for 20%, 40%, 60%, 80% and 100% of the data in MSCOCO for all 80 concepts. In subfigure (b), we report the learning performance for 20, 40, 60 and 80 concepts on the corresponding 20,000 examples chosen from MSCOCO.

that on MSCOCO. As analyzed in the learn ability analysis in section V, we recollect that a crucial reason for the differences in mastering overall performance is the one of a kind numbers of each ideas and pictures.

3) Consequences of the number of records Examples and ideas: (a) illustrates the common performance of the NCD technique with the trade of the records length. We randomly pick out 20%, 40%, 60% and 80% of the facts from MSCOCO for all 80 ideas and perform five independent experiments for every percentage. From this discern, we find that the mastering overall performance normally improves with the growth of the percent. We additionally observe that the take into account has a minor lower while the proportion adjustments from 40% to 60%. We do not forget the principle purpose is that the enrichment of facts introduces a few picture-textual content information pairs that pose problems in the concept learning. In Fig. 6(b), we analyze the mastering overall performance when different numbers of concepts are taken into consideration with a fixed facts size.

4) Ablation study: to show the contribution of each thing in Eq. (thirteen), we test the gaining knowledge of overall performance of NCD with specific configurations. The variations include: 1) NCD (without health), which eliminates the fitness term, 2) NCD (without distinctiveness), which is acquired by using removing area of expertise, three) NCD (without smoothness), which is obtained by using eliminating the smoothness time period, and four) NCD (without sparseness), which ignores the L1-norm regularization that constrains the sparseness of the gaining knowledge of results. Desk IV indicates the evaluation effects of the variants. From the table, we observe that the versions NCD (without health) and NCD (without smoothness) typically perform worse than the others. The phenomenon is constant with the evaluation within the method segment.



(a)

(b)

Figure 6: The comparison results in terms of F1-score over the three datasets. (a) The comparison results in terms of F1-score for 17 concepts randomly chosen from the MSRC dataset. (b) The comparison results in terms of F1-score for total 20 concepts in the VOC2012 dataset.

NCD approach. Thinking about that important gadgets in maximum natural photographs have a tendency to have exceptional semantics, we introduce the individuality term, which helps push away the concept labels of patches belonging to the identical photo?

5) Assessment of learning effects: desk IV reports the mastering effects, inclusive of precision, bear in mind and F1-score, for the NCD method and the 5 associated techniques. From this desk, we take a look at that the NCD approach outperforms the in comparison strategies on the averaged idea studying outcomes over the three datasets. In comparison with the related strategies, our method improves the F1-score by using 8.zero% on MSRC, via eight. Four% on VOC2012 and through 2.3% on MSCOCO. To our surprise, the kNN technique, basically hired as a simple voting strategy over graphs in these paintings, achieves an appropriate result. It similarly demonstrates that the graph is a affordable version for FVC studying thru which semantics may be correctly propagated. MIMLfast is an effective and effective approach for MIML duties. Without the label of each instance (the situation applied to our technique and the different as compared methods), MIMLfast can gain an accuracy of zero. Sixty six±zero.03 (mean±std) on the MSRC dataset with very rapid computing pace. At the same dataset, our technique surely achieves the accuracy of 0.809 and outperforms MIMLfast. Fig. illustrates the contrast of our proposed approach with the associated work in phrases of F1-score for the three datasets. From the determine, we examine that the NCD method achieves better overall performance for maximum of the concepts than the opposite techniques.

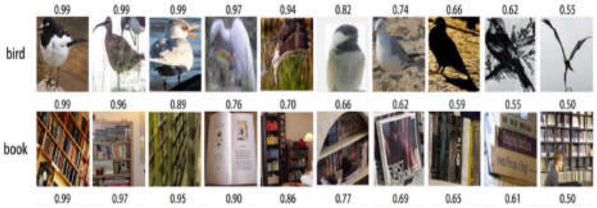


Fig. 7. Example illustration of 10 concepts. The value over each patch denotes the confidence that it belongs to the corresponding concept category. For each concept category, we show 10 representative patches in descending order of confidence.

VI. CONCLUSION

In this paper, we formulate the problem of fundamental visual concept learning from correlated images and text and propose an approach to this problem called neighboring concept distributing. The approach distributes the semantic information from the images to the patches and propagates it across different patches by considering fitness, distinctiveness, smoothness and sparseness. The learn ability analysis reveals that, under some conditions, all concepts can be learned with an arbitrarily high performance as the correlated images and text data increase. Experimental results demonstrate that the proposed NCD approach outperforms state-of-the-art methods. Based on this work, some significant issues will be addressed in the future as follows: 1) for human beings, neural network structures may change through learning. Likewise, the evolution of models in terms of structures and parameters for learning FVCs is significant. 2) It is difficult to achieve good performance of FVC learning in complicated cases, such as “image + paragraph” and noisy/false correspondence between images and text. Therefore, a strategy is needed to transfer learning results from simple/canonical data to complicated data.

VII. REFERENCES

[1] Adorni G, Di Manzo M, Giunchiglia F Natural Language Driven Image Generation, in *Proceedings of the 10th International Conference on Computational Linguistics and 22Nd Annual Meeting on Association for Computational Linguistics*

[2] Agrawal R, Gollapudi S, Kannan A, Kenthapadi K Enriching Textbooks with Images, in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*

[3] Bernardi R, Cakici R, Elliott D, Erdem A, Erdem E, Ikizler N, Keller F, Muscat A, Plank B (2016) Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures

[4] Bobick A, Intille SS, Davis JW, Baird F, Pinhanez C, Campbell LW, Ivanov Y, Schtte A, Wilson AD (1999) "The KidsRoom: A Perceptually-Based Interactive and Immersive Story Environment," *Presence,* pp. 369–393

[5] Boonpa SRS, Charoenporn T (2017) Relationship extraction from Thai children's tales for generating illustration, *2nd International Conference on Information Technology (INCIT)*

[6] Carney RN, Levin JR (2002) "Pictorial Illustrations Still Improve Students' Learning from Text," in *J.R. Educational Psychology Review*

[7] Chong W, Blei D, Li F (2009) "Simultaneous image classification and annotation," in *IEEE Conference on Computer Vision and Pattern Recognition*

[8] Coelho F, Ribeiro C (2011) "Automatic Illustration with Cross-media Retrieval in Large-scale Collections," in *Content-Based Multimedia Indexing (CBMI)*

[9] Coyne B, Sproat R (2001) Wordseye: an automatic text-to-scene conversion system," in *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*

[10] Csomai A, Mihalcea R (2007) Linking educational materials to encyclopedic knowledge, in *Proceedings of the 2007 Conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work*

[11] Delgado D, Magalhães J, Correia N (2010) Automated illustration of news stories, in *Proceedings of the 2010 IEEE Fourth International Conference on Semantic Computing*

[12] Dmitry U (2012) A Text-to-Picture System for Russian Language," in *Proceedings 6th Russian Young Scientist Conference for Information Retrieval*

[13] Dmitry U, Alexandr K (2012) An Ontology-Based Approach to Text-to-Picture Synthesis Systems, in *Proceedings of the Second International Workshop on Concept Discovery in Unstructured Data*

[14] S. Dupuy, A. Egges, V. Legendre and P. Nugues (2001) Generating a 3d simulation of a car accident from a written description in natural language: The carsim system, in *Workshop on Temporal and Spatial Information Processing*

[15] Duy B, Carlos N, Bruce EB, Qing ZT (2012) Automated illustration of patients instructions, *Journal of the American Medical Informatics Association,* pp. 1158–1167

[16] Elhoseiny M, Elgammal A (2015) Text to Multi-level MindMaps: A Novel Method for Hierarchical Visual Abstraction of Natural Text, in *Multimedia Tools and Applications*

[17] Erhan OV, Alexander T, Samy B, Dumitru E (2016) Show and tell: lessons learned from the 2015 {MSCOCO} image captioning challenge. IEEE Trans Pattern Anal Mach Intell 39(4):652–663

[18] Eunice MM (2006) *automatic conversion of natural language to 3D animation,* University of Ulster

[19] Ganguly D, Calixto I, Jones G (2015) "Overview of the Automated Story Illustration Task at FIRE 2015," in *FIRE*

[20] Goldberg A, Dyer CR, Eldawy M, Heng L (2008) "Easy As ABC?: Facilitating Pictorial Communication via Semantically Enhanced Layout," in *Proceedings of the Twelfth Conference on Computational Natural Language Learning*

[21] Goldberg AB, Rosin J, Zhu X, Dyer CR (2009) "Toward text-to-picture synthesis," in *Proc. NIPS 2009 Symposium on Assistive Machine Learning for People with Disabilities*

[22] Hanser E, Kevitt PM, Lunney T, Condell J (2009) Scenemaker: automatic visualisation of screenplays, in *Annual Conference on Artificial Intelligence*

Authors Profile

Ms. Shaik Heena working as Assistant Professor of CSE Department in QIS College of Engineering and Technology (Autonomous), Ongole, Andhra Pradesh, India.

Mr. Venigalla Vamsi Krishna pursuing B Tech in computer science engineering from Qis college of Engineering and Technology (Autonomous & NAAC 'A' Grade), Ponduru Road, Vengamukkalapalem, Ongole, Prakasam Dist, Affiliated to Jawaharlal Nehru Technological University, Kakinada in 2016-20 respectively.

Ms. Padarthi Aswani pursuing B Tech in computer science engineering from Qis college of Engineering and Technology (Autonomous & NAAC 'A' Grade), Ponduru Road, Vengamukkalapalem, Ongole, Prakasam Dist, Affiliated to Jawaharlal Nehru Technological University, Kakinada in 2016-20 respectively.

Ms. Nalluri Haritha pursuing B Tech in computer science engineering from Qis college of Engineering and Technology (Autonomous & NAAC 'A' Grade), Ponduru Road, Vengamukkalapalem, Ongole, Prakasam Dist, Affiliated to Jawaharlal Nehru Technological University, Kakinada in 2016-20 respectively.

Mr. Eemani Venkata Sai Ram pursuing B Tech in computer science engineering from Qis college of Engineering and Technology (Autonomous & NAAC 'A' Grade), Ponduru Road, Vengamukkalapalem, Ongole, Prakasam Dist, Affiliated to Jawaharlal Nehru Technological University, Kakinada in 2016-20 respectively.