Discovering Connotations as Labels for Weakly Supervised Image-Sentence Data

Aditya Mogadala
Karlsruhe Institute of Technology
Karlsruhe, Germany
aditya.mogadala@kit.edu

Bhargav Kanuparthi*
BITS
Hyderabad, India
f20140527@hyderabad.bits-pilani.ac.in

Achim Rettinger
Karlsruhe Institute of Technology
Karlsruhe, Germany
rettinger@kit.edu

York Sure-Vetter
Karlsruhe Institute of Technology
Karlsruhe, Germany
york.sure-vetter@kit.edu

ABSTRACT

Growth of multimodal content on the web and social media has generated abundant weakly aligned image-sentence pairs. However, it is hard to interpret them directly due to intrinsic *"intension"*. In this paper, we aim to annotate such image-sentence pairs with connotations as labels to capture the intrinsic *"intension"*. We achieve it with a connotation multimodal embedding model (CMEM) using a novel loss function. It's unique characteristics over previous models include: (i) the exploitation of multimodal data as opposed to only visual information, (ii) robustness to outlier labels in a multi-label scenario and (iii) works effectively with large-scale weakly supervised data. With extensive quantitative evaluation, we exhibit the effectiveness of CMEM for detection of multiple labels over other state-of-the-art approaches. Also, we show that in addition to annotation of image-sentence pairs with connotation labels, byproduct of our model inherently supports cross-modal retrieval i.e. image query - sentence retrieval.

CCS CONCEPTS

• **Information systems → Web searching and information discovery**; • **Computing methodologies → Neural networks**; *Image representations*; *Learning settings*;

KEYWORDS

Image-Sentence Connotation Labels, Weakly Supervised Deep Learning, Multi-label Prediction

ACM Reference Format:
Aditya Mogadala, Bhargav Kanuparthi, Achim Rettinger, and York Sure-Vetter. 2018. Discovering Connotations as Labels for Weakly Supervised Image-Sentence Data. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3184558.3186352

*Work done while doing an internship at Institute AIFB.

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3186352

1 INTRODUCTION

Vast amount of visual data is created daily and major chunk of it is found on the web and social media. Many approaches are built to leverage such data (e.g. Flickr[1]) for building datasets [8, 19, 21] by employing human efforts to filter the noisy images and annotate them with object categories. However, human involvement includes cost and also acquire other problems such as incompleteness and bias [22]. Hence, an alternative approach would be is to learn visual features and object detectors directly without using any manual labeling.

Hitherto, some approaches have explored the idea of automatically leveraging different types of web data from sources constituting only images [7] and images accompanied with text [36] to build visual models [31]. Although, it is asserted that the data is automatically extracted (e.g. search engines) and trained. Models are generally subjected to bias added by sources from which they are acquired. For example, image search engines (e.g. Google) usually concentrate on acquiring high-precision over recall and hence rank those images higher where a single object is centered with a clean background. In this case, images obtained may contain false positives but images themselves are not very complex to interpret i.e. images represent objects which can be easily localized.

However, other forms of web data (e.g. social media) usually contain complex images which can be refereed with labels denoted by different connotations. Linguistically, a connotation is refereed to an idea that a word may hold which is in addition to its main or literal meaning (i.e. denotation). Pertaining images, it denotes that an image can also be described with connotations (e.g. abstract meaning) in addition to their usual denotations (e.g. visual objects depicting WordNet categories). Also from the perspective of logic and semantics, connotation refers to *intension*[2] [5]. Figure 1 shows sample image-tweet pairs where an image-tweet pair augmented with connotations along with their denotations is better interpretable when compared against rest. It is evident that adding connotations to complex images usually found on social media platforms is beneficial. Howbeit, most part of current research usually concentrate only building visual models that handle denotations and only learn from images.

[1] https://www.flickr.com/
[2] Not to be confused with "intention".
[3] https://clarifai.com/

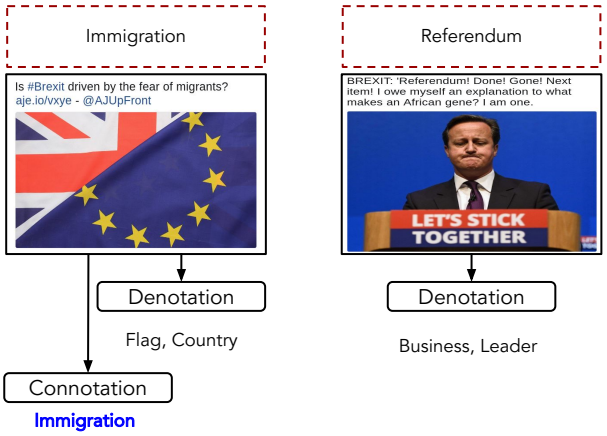

Figure 1: Augmenting connotations acquired from an image-tweet pair along with denotations provide better comprehension of "intension". Boxes in red denote ground-truth "intension" observed in the image-tweet pair. Denotations are captured only from an image with a commercial image recognition API[3].

Hence in this paper, we aim to add such diversity in labels by harnessing large-scale data. Usually, standard web-scale image datasets (e.g. YFCC100M [30]) have shorter textual context and provide only denotations. However, connotations can be acquired only from the larger textual context. Therefore, our first goal is to acquire such image-textual data which provide such a context. Specifically, 1) we leverage Twitter[4] to collect weakly supervised image-tweet pairs data that provides such a context. Since manual annotation of images-tweet pairs at the scale of Twitter is tedious. We leverage semantic annotation and disambiguation approaches [3] for generating connotations. 2) Second, an architecture based on embedding models [13, 33] is leveraged to capture correlation between image-tweets and connotations by learning their common space representation. Further, for any given input image-tweet, connotations are ranked according to the dot product between connotation and image-tweet embeddings. 3) Lastly, byproduct of our model is used to perform cross-modal retrieval to compare its effectiveness with other similar approaches.

We believe that this work will provide a new direction for exploiting social media data to achieve varied visual tasks without human labeling effort. In rest of the paper, Section 2 presents related work and the Section 3 describes our approach to learn features from image-tweet pairs and then learn a connotation model to rank the connotations. Further, experimental setup Section 4 present the preliminaries about dataset and evaluation measure, While experimental results are presented in Section 5 followed by the conclusion and future work.

2 RELATED WORK

Our related work can be drawn from many closely aligned areas.

[4]https://twitter.com/

2.1 Labeling Images with Webly Supervised Learning

There has been a long standing interest in mining visual data from the web [6]. Many approaches [34] have focused there efforts on either cleaning the web data by leveraging pre-trained models built from datasets created with human supervision (e.g. ImageNet [8]) or aimed to automatically discover hidden patterns to train models directly from it [39]. Our work also focuses on the later objective and has an intent to tackle noise involved in such scenarios for building effectual models. Already, some approaches [35] have handled similar challenges by filtering the noise when learning visual models. However, we differ from them by directly not learning visual representation models (e.g. CNNs [16, 28]) as we understand that learning from CNN with noisy labeled data is still an open problem. But, we leverage multimodal data to address the challenge. Also, aforementioned approaches only operate with single label per image, while we predict multiple labels per image.

2.2 Cross-Modal Retrieval with Images and Text

One of the closely aligned research fields is cross-modal retrieval with images and text. Over the past few years many approaches are proposed for cross-modal retrieval concerning images and textual forms observed in variable lengths such as phrases, sentences and paragraphs. Most of these early proposed approaches belong to subspace learning methods which learn a common space for cross-modal data, in which the similarity between the modalities is measured using varied distance metrics. Several of such subspace learning methods exists such as Canonical Correlation analysis (CCA) [23] etc.

However, subspace learning methods are generally susceptible to scaling challenges. To overcome such issues, probabilistic graphical model (PGM) based approaches are proposed such as correspondence Latent Dirichlet Allocation (Corr-LDA) [2] and their variations. Howbeit, these approaches also faced drawback as exact inference in general is intractable and has to depend on the approximate inference methods, such as variational inference, expectation propagation, or Gibbs sampling.

Deep neural network based methods overcame challenges observed in subspace learning and PGM models by designing robust techniques that can scale to large data and also avoid intractable inference issues. Approaches such as deep restricted boltzmann machine (Deep RBM) [29], deep canonical correlation analysis (DCCA) [1], correspondence autoencoder (Corr-AE) [12] and deep visual-semantic embeddings [13] used multimodal inputs to learn representations of common spaces.

Our approach falls in-line with the family of deep learning methods and is proximal to the visual-semantic embedding approaches. However, our model goal is bigger than performing common space learning, we aim to predict of multiple labels by leveraging common space of each multimodal pair.

2.3 Hashtag Prediction

Prediction of connotations which captures intension in social media data is also closely aligned with the hashtag prediction [33] or recommendation [27]. Hashtags are regularly observed to capture

authors sentiment or comprehension on a particular topic. However, they are usually illustrated with n-grams or abbreviations and sometimes difficult to interpret when compared with semantically enriched connotation labels.

Nevertheless, initially several approaches have leveraged deep neural networks to build their models only with social media text (e.g. Tweets) for prediction or recommendation. However, these approaches pursued different paths to achieve their goal. Weston et al., [33] composed semantic embeddings from hashtags, while Dhingra et al. [10] utilized character-based embeddings and Gong et al., [15] used attention-based CNN. Only recently, using hashtags for image tagging was explored. Denton et al. [9] proposed a 3-way multiplicative gating approach, where the image model is conditioned on the user metadata on Facebook dataset. While, Park et al. [24] Context Sequence Memory Network (CSMN) model mainly built for personalized image captioning to predict hashtags on Instagram dataset. However, none of the aforementioned approaches leveraged multimodal social media data for utilizing larger contexts. Also, none of the hashtags were semantically enriched for better interpretation.

3 APPROACH

Let $\mathcal{S} = \{(I_j, T_j), Y_j\}_{j=1}^{N}$ be our dataset with (I_j, T_j) the j-th image-tweet pair and $Y_j \subseteq \mathcal{Y}$ the automatically extracted corresponding connotations set, where $\mathcal{Y} \triangleq \{1, 2, , K\}$ is set of all possible connotations. Each image-tweet can have different number of connotations $(I_j, T_j) = |Y_j|$.

Our goal is now to learn a ranking model R(I,T,Y) that computes the confidence scores to all connotations to rank relevant connotations for a given image-tweet pair. We further decompose R(I,T,Y) = f(g($\Phi(I)$,$\Psi(T)$),E_y) where $E_{\mathcal{Y}} \in \mathbb{R}^{d \times K}$ denote connotation label embeddings matrix and g($\Phi(I)$,$\Psi(T)$): $\mathbb{R}^I \times \mathbb{R}^T \rightarrow \mathbb{R}^d$ computation model to add tweet bias to image representations. Further, f(g($\Phi(I)$,$\Psi(T)$),Y): $(\mathbb{R}^d, \mathbb{R}^d) \rightarrow \mathbb{R}^K$ computes dot product between g($\Phi(I)$,$\Psi(T)$) and connotation embedding matrix $E_{\mathcal{Y}}$ for finding confidence scores of relevant connotations Y. We now adapt a Convolutional Neural Network (CNN) [16] for an image representation ($\Phi(I)$), character-level long short-term memory (charLSTM) [17] for the tweet representation ($\Psi(T)$) and a novel loss function for learning R(I,T,Y) model.

In the following, we provide details of individual components of the ranking model R(I,T,Y).

3.1 Image-Tweet Bilinear Model

Aim of the image-tweet bilinear model is to compute g($\Phi(I)$,$\Psi(T)$). Initially, we present architectures used for extracting feature representations from both image (I) and tweet (T) i.e. $\Phi(I)$ and $\Psi(T)$ respectively followed by the bilinear model.

3.1.1 Tweet Representation. Tweets (T) are sequences upto 140 characters with inherent semantic and syntactic meaning. Encoding a tweet into a embedding vector ($\mathbb{R}^T$) can encapsulate the compositional structure of the entire tweet. Thus, we propose to leverage charLSTM i.e. $\Psi(T, \Theta)$ to build embedding for each tweet, where Θ represent parameters of charLSTM. Initially, characters in a tweet are read sequentially to be further fed as input to an charLSTM encoder for encoding a tweet it into a $\mathbb{R}^T$ vector.

3.1.2 Image Representation. For representing images (I) into fixed vector ($\mathbb{R}^I$). We use pre-trained CNN on ImageNet classes as a feature extractor i.e. $\Phi(I)$ to obtain the image embeddings from the raw image. The image vectors of dimensionality $\mathbb{R}^I$ are extracted from the final fully connected layer of the network without the top Softmax layer.

3.1.3 Tweet-Biased Image Representation. Image and tweet representations belong to two different feature spaces and do not share any common representation. To associate image and tweet representations, image-tweet bilinear model gives a simple method for leveraging tweet information by adding a tweet dependent bias term to the image embedding. In particular, the tweet-biased image embedding g($\Phi(I)$,$\Psi(T)$): $\mathbb{R}^I \times \mathbb{R}^T \rightarrow \mathbb{R}^d$ is defined by Equation 1.

$$g(\Phi(I), \Psi(T)) = W_I^T \Phi(I) + W_T^T \Psi(T) \tag{1}$$

where $W_I \in \mathbb{R}^{I \times d}$ and $W_T \in \mathbb{R}^{T \times d}$ are image and tweet parameter matrices respectively.

3.2 Connotation Multimodal Embedding Model

The connotation multimodal embedding model (CMEM) denoted using the function f(g($\Phi(I)$,$\Psi(T)$),$E_{\mathcal{Y}}$;θ) $\in \mathbb{R}^K$ learns a joint embedding space for connotation embedding matrix ($E_{\mathcal{Y}}$) and tweet-biased image representations ($g(\Phi(I), \Psi(T))$) to rank connotations. Figure 2 presents the overall model.

To learn parameters of f($\cdot$) $\in \mathbb{R}^K$, an optimization problem is solved using the loss function (l) given by Equation 2.

$$\min_{\theta} \frac{1}{N} \sum_{n=1}^{N} l(f(g(\Phi(I_n), \Psi(T_n)), E_{\mathcal{Y}}; \theta), Y_n) + \lambda ||\theta||_2^2 \tag{2}$$

where θ refers to the parameters of the CMEM.

Furthermore, we design the loss function (l) in a manner to leverage large datasets and enforce f($\cdot$) to produce results whose values for true connotations are greater than those for negative connotations for any given image-tweet pair. In particular, pairwise rank loss (PRL) [4] suits such a criteria and is given by Equation 3.

$$l_{prl} = \sum_{\hat{y} \notin Y_i} \sum_{y \in Y_i} max(0, \alpha + f_{\hat{y}}(\cdot) - f_y(\cdot)) \tag{3}$$

where $\hat{y}$ represent negative connotations for any given positive connotation y, *alpha* is the hyper-parameter that denotes margin. However, l_{prl} is not smooth everywhere and thus makes it difficult to optimize.

Therefore for CMEM, we propose to explore three different losses that provides better theoretical guarantees than the l_{prl} and makes easier for optimization. In the following, we first present the two existing techniques based on pairwise rank loss (i.e. WARP [32] and LSEP [20]) and then present our proposed loss function.

3.2.1 Weighted Approximate Rank Pairwise (WARP) Loss. Weston et al. [32] extended pairwise rank loss provided in Equation 3 by adding weights on violations with Weighted Approximate Rank Pairwise (WARP) loss given by Equation 4.

$$l_{warp} = \sum_{\hat{y} \notin Y_i} \sum_{y \in Y_i} w(r_i^y) max(0, \alpha + f_{\hat{y}}(\cdot) - f_y(\cdot)) \tag{4}$$

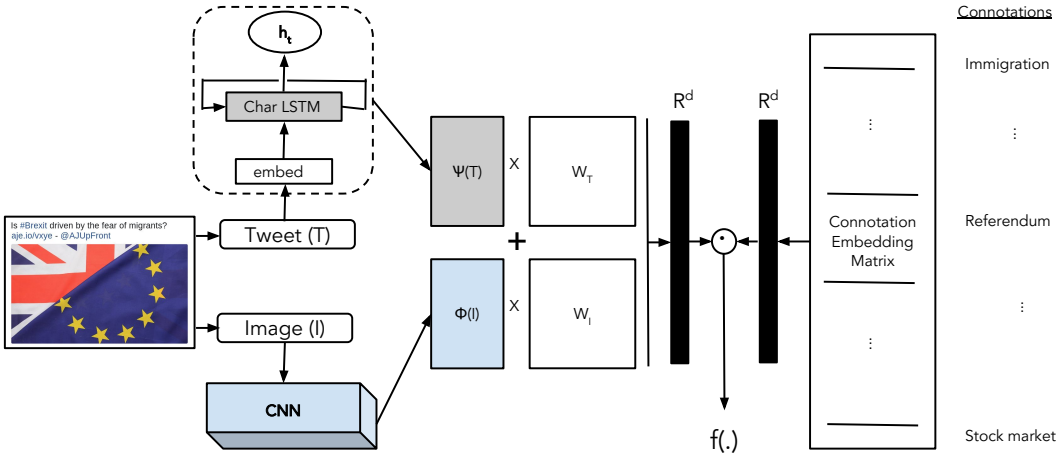

Figure 2: Connotation Multimodal embedding model with its different constituents. $\mathbb{R}^d$ refers to the final d-dimension representation of image-tweet pair in-line with the connotation embeddings dimensions. $\odot$ denote element-wise dot product.

where w(·) denote monotonically increasing function and r_i^y is the predicted rank of the positive connotation y. The intuition is that if the positive connotation is ranked lower, then the violation should be penalized higher. However, due its non-smoothness it is not differentiable everywhere and makes optimization difficult.

3.2.2 Log-Sum-Exp Pairwise (LSEP) loss. Addressing issues in l_{prl} and l_{warp} such as non-smoothness, adaptable margins etc., Li et al. [20] proposed Log-Sum-Exp pairwise (LSEP) loss by modifying exponential pairwise rank loss (l_{epl}) [38] given by the Equation 5.

$$l_{lsep} = log\left(1 + \sum_{\hat{y} \notin Y_i} \sum_{y \in Y_i} exp(f_{\hat{y}}(\cdot) - f_y(\cdot))\right) \tag{5}$$

LSEP is expected to provide flexibility to the learning problem by allowing adaptable margins per sample pair and also making it smooth everywhere. Also, LSEP do not use weight function w(·) as it is expected have implicit weight effect to penalize the lower ranked positive connotations harder. Although, LSEP have many advantages such as it can linearly scale with vocabulary size with negative sampling technique [14] and provide better numeric stability. Nevertheless, it still lack two key abilities. (1) l_{lsep} is not α-convex [25]. This means that we cannot place a bound on how long gradient descent takes to converge. Howbeit, it partially mitigates the problem with regularization (2) l_{lsep} uses variant of logistic loss, thus making it sensitive to outliers in the data and assigns large loss values to them. We aim to overcome such challenges with our proposed penalized-logistic-sum pairwise (PLSP) loss.

3.2.3 Penalized-Logistic-Sum Pairwise (PLSP) Loss. It can be comprehended from aforementioned sections that pairwise ranking approaches dependent on variants of hinge loss (e.g. l_{prl},l_{warp}), exponential loss (e.g. l_{epl}) and logistic loss (e.g. l_{lesp}). Our proposed approach is the variant of truncated logistic loss and is expected to be *alpha*-convex while being robust to outliers. Equation 6 shows the loss function l_{plsp}.

$$l_{plsp} = log\left(1 + \sum_{\hat{y} \notin Y_i} \sum_{y \in Y_i} exp\left(\frac{min(f_{\hat{y}}(\cdot) - f_y(\cdot), s)}{max(f_{\hat{y}}(\cdot) - f_y(\cdot), -s)}\right)\right) \tag{6}$$

where $s < 0$ denotes location of truncation. Important property of l_{plsp} is that the denominator value in the exponential cannot get extremely small because it is lower bounded by s. Similarly, numerator of the equation cannot get cannot get extremely big. Therefore, it makes l_{plsp} robust to noise of outliers and also smooth everywhere due to exponential.

4 EXPERIMENTAL SETUP

4.1 Dataset

In this section, we introduce a new dataset called *TwitterBrexit* collected from Twitter.

4.1.1 Dataset Procurement. In the following, we present varied stages involved in creation of the dataset.

Tweets Collection is specific to a domain i.e. Brexit in our case. This is undertaken to reduce noise in the collection and to ensure connotations set is interpretable. Otherwise, we will end up procuring randomly distributed labels and could lead to uninterpretable results. Initially, we attained seed topic words using Google trends[5] during the period of May 2015 and May 2016 for searching Twitter. Topic words such as Brexit, Immigration, Racism, Theresa, etc., are then used as queries to Twitter search API [6] for collecting tweets. This step is iterated several times until a long list of tweets are acquired.

Tweets pruning is performed to acquire only those tweets with corresponding images. We found that only 25% of the tweets collected are accompanied with images. Further, pruned image-tweet

[5]https://trends.google.com/trends/

[6]https://developer.twitter.com/en/docs/tweets/search/overview/basic-search

pairs is again processed to eliminate junk, tweets without words, English only tweets and duplicates.

4.1.2 Dataset Peculiarity. The new dataset introduced in this paper is peculiar and also challenging to process when compared against other similar datasets on the following aspects. First, dataset is collected from the social media platform. Hence, the language usually used will be informal comprising grammatical mistakes and large vocabularies. However, there are also additional characteristics which is helpful for highlighting information present in the image-tweet pair such as hashtags. Second, the association between the image and tweet is often loosely connected. Hence, they are weakly supervised. Third, dataset is useful for large-scale training and can be exploited to test robustness of multi-label classifiers.

4.1.3 Creation of Connotations. For acquiring connotations for images present in the collection, text aligned with images is leveraged. An usual strategy to make sense or extract intension from the social media text is by annotating them with semantic enricher's. Acquired connotations are considered as a brief summarization of the content present in the text. Hence, connotations also support better information interpretation. We leveraged semantic annotation and disambiguation tool [37] to obtain such labels. Since these labels are acquired from the text aligned with images, labels are also expected to describe images. However, they are not preferable for direct learning of image recognition models. As they are extracted automatically without human supervision and hence can induce noise into learning. In total, the dataset contained ~30k distinct connotations as labels. The mean number of labels per tweet was 2.3 with a standard deviation of 1.3. A large fraction of labels describe the content of the image, with many synonyms. Others describe abstract meaning representing possible intension in the image content (e.g. Economics, Xenophobia).

The distribution of labels in the dataset is far from uniform: the top 10 labels account for 47% of the total and ~27k of them appear less than 10 times throughout the whole dataset. It is difficult to predict infrequent labels, so we limit top 1387 labels which have appeared at least 25 times in the dataset to create a balanced version of the dataset. Figure 3 shows sample annotations, while Figure 4 presents top-50 frequent labels in the entire dataset. In total, our

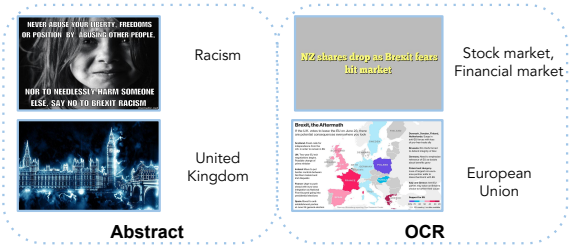

Figure 3: Example connotations as labels observed for different variants of images. For example, "OCR" variant represent images which contain text and "Abstract" denote concepts close to real-world entities.

collection comprises around 160,004 image-tweet pairs for training, 10,000 for validation and 10220 for testing.

4.1.4 Applications of the Dataset. We leverage the dataset mainly for multi-label prediction. However, we also show that it is also useful for cross-modal retrieval.

4.2 Evaluation Measures

To measure the effectiveness of discovering connotations as labels for images. We use different measures such as recall, multi-label accuracy, Hamming loss [11] and coverage.

4.2.1 Recall@k (R@k). measures of the fraction of relevant connotation labels for each test image-tweet pair that are ranked in the top k given by Equation 7.

$$\text{Accuracy} = \frac{1}{|q|}\sum_{j=1}^{|q|}\frac{|X_j \bigcap Y_j|}{|Y_j|} \tag{7}$$

where X_j refers to the predicted correct labels and Y_j the ground-truth labels for the j-th query.

4.2.2 Multi-label Accuracy@k (ML-A@k). measures the proportion of predicted correct labels that are ranked in the top k to the total number of ground-truth labels for a given image query. Overall accuracy is the average across all queries given by Equation 8.

$$\text{Accuracy} = \frac{1}{|q|}\sum_{j=1}^{|q|}\frac{|X_j \bigcap Y_j|}{|X_j \bigcup Y_j|} \tag{8}$$

Higher the value of accuracy, better the performance.

4.2.3 Hamming Loss (HL). measures how many times on an average the relevance of an instance to a class label is incorrectly predicted. Also, hamming loss consider both prediction error (i.e. prediction of incorrect label) and missing error (i.e. missing out the relevant label) normalized over total number of classes and examples given by Equation 9

$$\text{HL} = \frac{1}{|q|N}\sum_{j=1}^{|q|}\sum_{l=1}^{N}[\mathcal{F}(l \in X_j \wedge l \notin Y_j) + \mathcal{F}(l \notin X_j \wedge l \in Y_j)] \tag{9}$$

where $\mathcal{F}$ refers to the indicator function and l to the semantic labels. In practice, smaller the value of HL, better the performance.

4.2.4 Coverage. evaluates how much one needs to traverse the ranked list of labels on average to cover all the relevant labels of the sample and is provided by Equation 10.

$$\text{Cov} = \frac{1}{|q|}\sum_{j=1}^{|q|} max(rank(X_j)) - 1 \tag{10}$$

Smaller the value of coverage, better the performance.

5 EXPERIMENTS

5.1 Implementation

As discussed in aforementioned sections, important constituents of our CMEM are CNN, charLSTM, connotation embeddings and a loss function. For image representation, we explored two different CNN models, mainly VGG16 [28] and ResNet50 [16] pre-trained on the ImageNet ILSVRC dataset [26] by extracting features of dimensions 4096 and 2048 respectively from the final fully connected layer of the network without the top Softmax layer. For the charLSTM,

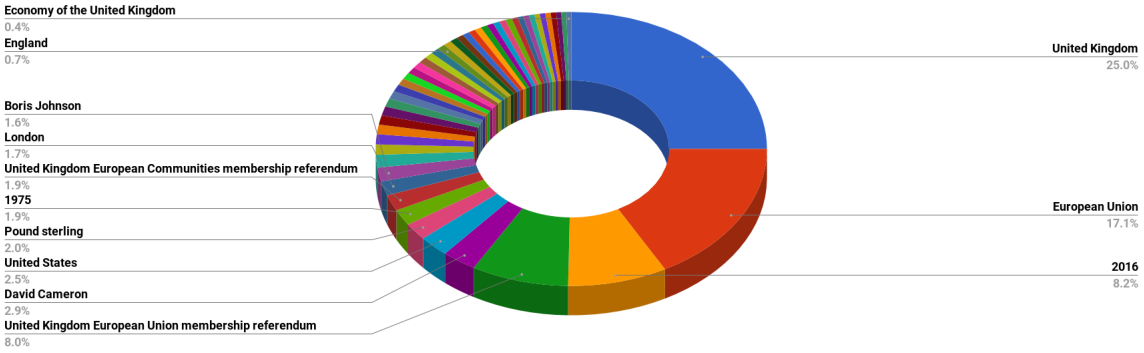

Figure 4: Top-50 frequent labels and their distribution.

we initialized character embeddings with 512 dimensions using Glorot uniform. Connotations acquired from semantic enricher are Wikipedia titles (i.e. concepts) which are also observed in DBpedia[7]. Therefore, we leveraged wiki2vec[8] to obtain 256 and 512 dimension embeddings for concepts. CMEM is now trained using Adam optimizer [18] with gradient clipping having maximum norm of 1.0 for 10 epochs. The weight decay λ in the regularization term of Equation 2 is set to 5e-5.

5.2 Results and Discussion

5.2.1 Baselines. We design our baselines based on the usage of varied loss functions with CMEM. For example, CMEM-*warp* represents our CMEM with the WARP loss.

5.2.2 Quantitative Analysis (Label Prediction). To conduct our evaluation, we leveraged the *TwitterBrexit* dataset mentioned in the Section 4.1. Different approaches are evaluated based on measures such as recall at 10 (R@10), accuracy at 10 (ML-A@10) and hamming loss (HL). Table 1 shows the results attained. We can notice that the CMEM-*prl* performed poorly when compared against all other models. However, when only advanced models like CMEM-*warp* and CMEM-*lsep* are compared, it can be observed that CMEM-*lsep* outperforms CMEM-*warp* on both recall and accuracy. Howbeit, for HL there seems to have no significant difference. This can be attributed to better optimization achieved with l_{lsep}.

Furthermore, we can perceive that CMEM with our proposed loss i.e. CMEM-*plsp* performs particularly well in terms of accuracy, recall and HL when compared against other baselines. Results also convey that our proposed loss in CMEM was particularly robust to outliers and could leverage that with significant gains in both recall and accuracy. Also, few more observations can be made about visual features and dimensions of connotation embeddings. ResNet50 performs better than VGG16, while larger dimensions for connotation embeddings perform better than their counterparts with lesser dimensions.

[7]http://wiki.dbpedia.org/
[8]https://github.com/idio/wiki2vec

Figure 5 shows average precision-recall (PR) curves that allow us to comprehend the effect of label prediction. It can be perceived that our CMEM-*plsp* outperforms others suggesting the robustness of our l_{plsp} loss with CMEM when compared to other baselines.

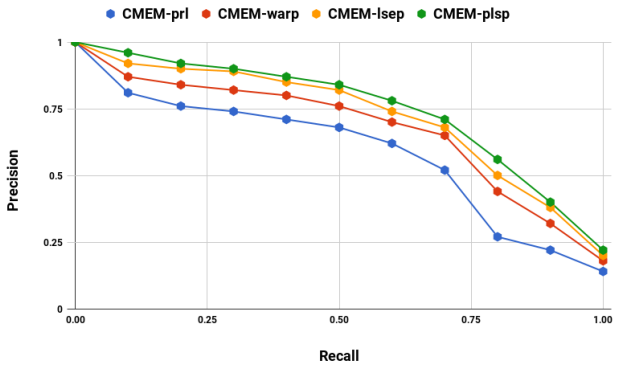

Figure 5: Average Precision-Recall Curve

5.2.3 Quantitative Analysis (Cross-modal Retrieval). A natural consequence of CMEM is learning of parameters W_I and W_T. During learning, both of them are updated jointly when optimized w.r.t connotations. Hence, they inherently share correlation between image and tweets.

In this section, we evaluate their effectiveness with image query for tweet retrieval and compare them with standard subspace learning methods such as canonical correlation analysis (CCA) and its variants (i.e. regularized CCA (RCCA)) using Mean rank. We also explored other representations for the text such as latent Dirichlet allocation (LDA) [2]. The LDA model is trained with 50 topics to represent each tweet with 50-dimensional LDA feature by the topic assignment probability distributions. Table 2 shows the comparison of different approaches for image to tweet retrieval.

5.2.4 Qualitative Analysis (Label Prediction). Figure 6 presents sample results attained with CMEM using different loss functions.

Loss Function	Connotation Embeddings	CNN Architecture	R@10	ML-A@10	HL	Cov
CMEM-prl	256	VGG16	18.11	35.42	0.416	12.61
		ResNet50	18.17	35.84	0.412	12.54
	512	VGG16	18.84	36.10	0.408	12.46
		ResNet50	18.90	36.54	0.406	12.44
CMEM-$warp$	256	VGG16	18.95	36.69	0.406	12.37
		ResNet50	19.02	36.78	0.404	12.38
	512	VGG16	19.10	37.80	0.396	12.26
		ResNet50	19.24	38.24	0.389	12.10
CMEM-$lsep$	256	VGG16	19.22	38.15	0.390	12.14
		ResNet50	19.30	38.84	0.382	12.09
	512	VGG16	19.44	39.16	0.374	11.98
		ResNet50	19.51	39.40	0.371	11.93
CMEM-$plsp$ (ours)	256	VGG16	19.54	39.84	0.369	11.88
		ResNet50	19.63	40.05	0.368	11.85
	512	VGG16	19.63	40.08	0.368	11.84
		ResNet50	**19.68**	**40.26**	**0.366**	**11.79**

Table 1: Connotation Label prediction Results on *TwitterBrexit*. R@10, ML-A@10 represent percentages (%). Bold denote best, while underline represent second best.

Image → Tweet Retrieval											
Measures	Methods	10	20	30	40	50	60	70	80	90	100
Mean Rank	CCA-LDA	5371	5570	5627	5767	5779	5753	5774	5770	5752	5766
	RCCA-100-LDA	4902	5083	5224	5303	5312	5309	5312	5306	5304	5315
	RCCA-1000-LDA	4873	5060	5203	5260	5269	5267	5272	5262	10263	5275
	CCA-charLSTM	3616	3989	4130	4347	4409	4515	4572	4627	4661	4690
	RCCA-100-charLSTM	3637	3965	4125	4328	4397	4504	4565	4613	4650	4682
	RCCA-1000-charLSTM	3708	3909	4181	4377	4451	4551	4615	4649	4683	4719
	CMEM-prl	2618	2986	3124	3341	3403	3509	3562	3617	3652	3688
	CMEM-$warp$	2627	2959	3112	3318	3381	3502	3551	3603	3641	3672
	CMEM-$lsep$	2608	**2898**	**3101**	3277	3311	**3481**	3515	3549	3583	3619
	CMEM-$plsp$	**2588**	2908	3121	**3227**	**3286**	3496	**3488**	**3529**	**3575**	**3596**

Table 2: Mean rank (lower the better) using different percentage (%) of image queries for retrieval. RCCA-* represent different regularization (100, 1000). Underline represents second best. All results are reported using ResNet50 as image features.

It can be seen that connotations extracted shows better intension from the image-tweet pair when compared against using only denotations captured from the image.

6 CONCLUSION AND FUTURE WORK

In this paper, we presented an approach to automatically extract connotations as labels for images by leveraging weakly supervised image-tweet data. We showed that the approach is scalable to many new classes and can support large scale image recognition as required in Web scenarios. In future, we aim to extend the approach to varied domains and check its generalization ability. Also, we would like to address other problems such as label inter-dependency and sparsity.

REFERENCES

[1] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *International Conference on Machine Learning*. 1247–1255.

[2] David M Blei and Michael I Jordan. 2003. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 127–134.

[3] Kalina Bontcheva and Dominic Rout. 2014. Making sense of social media streams through semantics: a survey. *Semantic Web* 5, 5 (2014), 373–403.

[4] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*. ACM, 129–136.

[5] Rudolf Carnap. 1988. *Meaning and necessity: a study in semantics and modal logic*. University of Chicago Press.

[6] Xinlei Chen and Abhinav Gupta. 2015. Webly supervised learning of convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 1431–1439.

[7] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. 2013. Neil: Extracting visual knowledge from web data. In *Proceedings of the IEEE International Conference on Computer Vision*. 1409–1416.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 248–255.

[9] Emily Denton, Jason Weston, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. 2015. User conditional hashtag prediction for images. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1731–1740.

[10] Bhuwan Dhingra, Zhong Zhou, Dylan Fitzpatrick, Michael Muehl, and William W Cohen. 2016. Tweet2vec: Character-based distributed representations for social

[9] https://clarifai.com/

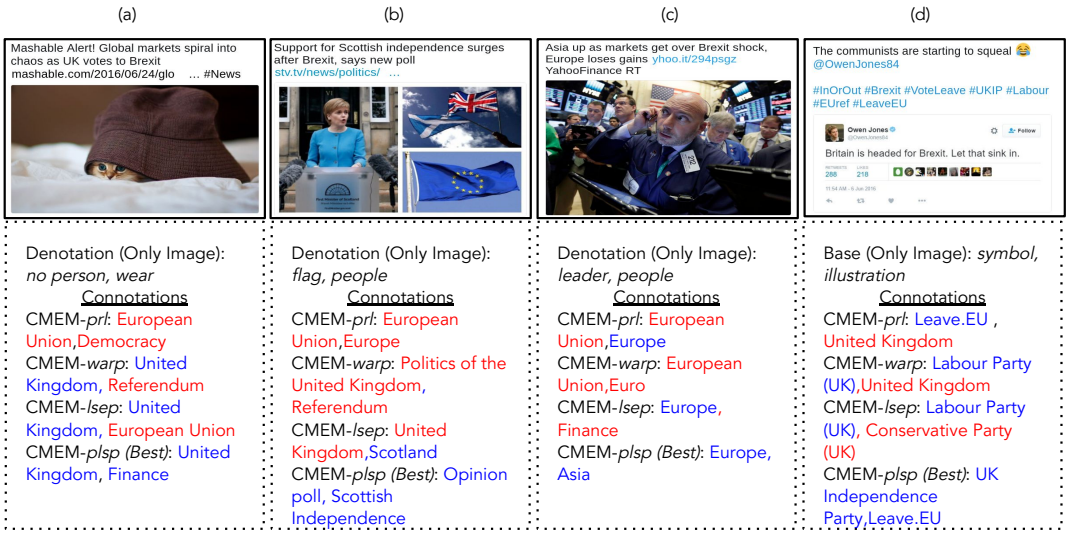

Figure 6: Sample qualitative results (red: False positives, Blue: True positives) attained with CMEM using different loss functions. Top-2 ranked connotations obtained are presented for each case. CMEM-*plsp* (Best) refers to the best model obtained from Table 1. Denotations are captured only using images with a commercial image recognition API[9].

media. *arXiv preprint arXiv:1605.03481* (2016).

[11] André Elisseeff and Jason Weston. 2002. A kernel method for multi-labelled classification. In *Advances in neural information processing systems*. 681–687.

[12] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. 2014. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 7–16.

[13] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. 2013. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*. 2121–2129.

[14] Yoav Goldberg and Omer Levy. 2014. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722* (2014).

[15] Yuyun Gong and Qi Zhang. 2016. Hashtag Recommendation Using Attention-Based Convolutional Neural Network.. In *IJCAI*. 2782–2788.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[17] Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-Aware Neural Language Models.. In *AAAI*. 2741–2749.

[18] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[19] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73.

[20] Yuncheng Li, Yale Song, and Jiebo Luo. 2017. Improving Pairwise Ranking for Multi-label Image Classification. *arXiv preprint arXiv:1704.03135* (2017).

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[22] Ishan Misra, C Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. 2016. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2930–2939.

[23] Aditya Mogadala and Achim Rettinger. 2015. Multi-modal Correlated Centroid Space for Multi-lingual Cross-Modal Retrieval. In *European Conference on Information Retrieval*. Springer, 68–79.

[24] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. 2017. Attend to You: Personalized Image Captioning with Context Sequence Memory Networks. *arXiv preprint arXiv:1704.06485* (2017).

[25] NN Pascu. 1979. Alpha-close-to-convex functions. In *Romanian-Finnish Seminar on Complex Analysis*. Springer, 331–335.

[26] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.

[27] Jieying She and Lei Chen. 2014. Tomoha: Topic model-based hashtag recommendation on twitter. In *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 371–372.

[28] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[29] Nitish Srivastava and Ruslan R Salakhutdinov. 2012. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*. 2222–2230.

[30] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: The new data in multimedia research. *Commun. ACM* 59, 2 (2016), 64–73.

[31] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. 2017. Learning From Noisy Large-Scale Datasets With Minimal Supervision. *arXiv preprint arXiv:1701.01619* (2017).

[32] Jason Weston, Samy Bengio, and Nicolas Usunier. 2010. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine learning* 81, 1 (2010), 21–35.

[33] Jason Weston, Sumit Chopra, and Keith Adams. 2014. #TagSpace: Semantic Embeddings from Hashtags. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1822–1827.

[34] Yan Xia, Xudong Cao, Fang Wen, and Jian Sun. 2014. Well begun is half done: Generating high-quality seeds for automatic image dataset construction from web. In *European Conference on Computer Vision*. Springer, 387–400.

[35] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. 2015. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2691–2699.

[36] Yazhou Yao, Jian Zhang, Fumin Shen, Xian-Sheng Hua, Jingsong Xu, and Zhenmin Tang. 2017. Exploiting Web Images for Dataset Construction: A Domain Robust Approach. *IEEE Transactions on Multimedia* (2017).

[37] Lei Zhang and Achim Rettinger. 2014. X-LiSA: cross-lingual semantic annotation. *Proceedings of the VLDB Endowment* 7, 13 (2014), 1693–1696.

[38] Min-Ling Zhang and Zhi-Hua Zhou. 2006. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE transactions on Knowledge and Data Engineering* 18, 10 (2006), 1338–1351.

[39] Yan-Tao Zheng, Ming Zhao, Yang Song, Hartwig Adam, Ulrich Buddemeier, Alessandro Bissacco, Fernando Brucher, Tat-Seng Chua, and Hartmut Neven. 2009. Tour the world: building a web-scale landmark recognition engine. In *Computer vision and pattern recognition, 2009. CVPR 2009. IEEE conference on*. IEEE, 1085–1092.