



Improving loss function for deep convolutional neural network applied in automatic image annotation

Ali Salar¹ · Ali Ahmadi¹

Accepted: 12 April 2023 / Published online: 12 May 2023

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract

Automatic image annotation (AIA) is a mechanism for describing the visual content of an image with a list of semantic labels. Typically, there is a massive imbalance between positive and negative tags in a picture—in other words, an image includes much fewer positive labels than negative ones. This imbalance can negatively affect the optimization process and diminish the emphasis on gradients from positive labels during training. Although traditional annotation models mainly focus on model structure design, we propose a novel unsymmetrical loss function for a deep convolutional neural network (CNN) that performs differently on positives and negatives, which leads to a reduction in the loss contribution from negative labels and also highlights the contribution of positive ones. During the annotation process, we specify a threshold for each label separately based on the Matthews correlation coefficient (MCC). Extensive experiments on high-vocabulary datasets like Corel 5k, IAPR TC-12, and Esp Game reveal that despite ignoring the semantic relationships between labels, our suggested approach achieves remarkable results compared to the state-of-the-art automatic image annotation models.

Keywords Image annotation · Deep learning · Unsymmetrical loss function · Threshold estimation

1 Introduction

Nowadays, as social networks gain popularity, a massive amount of image data is available on the internet, making it necessary to analyze and annotate them. Traditional image annotation methods that manually label image contents are no longer applicable as they have two main weaknesses [1]. First, manual annotation of this enormous amount of image data is impractical, and second, human annotators may have completely different interpretations of a single image. Consequently, automatically extracting a list of relative semantic labels becomes necessary, which means every automatic image annotation (AIA) method has to generate a list of labels describing a given image content.

Due to the wide variety of deep learning methods that have been used in this field over the past few years, AIA techniques can be classified into two primary categories: deep

learning-based methods and non-deep methods. According to [1], non-deep methods are classified into four categories: Generative models, Nearest neighbor models, Discriminative models, and Tag completions.

AIA techniques based on generative models are aimed to maximize the generative likelihood of image features and labels. These models have made significant contributions to the development of AIA. Nevertheless, estimating the generative likelihood between image features and annotations is insufficient to guarantee optimal label prediction. Furthermore, the complex relationship between labels and image features may not be captured accurately by these models.

AIA methods based on nearest neighbor models suppose that images with similar visual appearances tend to have identical labels. When a test image is presented, a group of resembling images are retrieved by these models from the training dataset; then, tags of the test image are derived using the vocabularies of these training images. However, the size of training datasets and retrieval performance may affect the performance of nearest neighbor model-based AIA methods.

Discriminative models treat image annotation as a classification problem with multiple labels [2, 3] and learn an independent binary classifier for each tag. Next, these binary classifiers are utilized to predict labels for test

✉ Ali Ahmadi
ahmadi@kntu.ac.ir

Ali Salar
parham.slrst@email.kntu.ac.ir

¹ Faculty of Computer Engineering, K. N. Toosi University of Technology, Tehran, Iran

Fig. 1 As shown in the illustration above, positive labels in images are much less than negative ones, resulting in a high imbalance between them



images. Nonetheless, these models have some drawbacks. For instance, it is often overlooked how image labels relate to visual features. Moreover, as discriminative models often rely on label correlations, the quality of training data is critical to their performance.

AIA techniques based on tag completion are distinctive in comparison to other annotation methods. These methods can be analyzed from two aspects. The first is adding missing labels for given images automatically, and the second is, removing noisy labels for stated images, which is called tag refinement [4]. However, optimizing the tag completion method can be computationally complex and time-consuming, and cannot ensure global optimization.

Recent developments in deep learning methods have helped AIA tasks to be solved using deep learning-based feature representations. It is possible to summarize these methods in two categories. The first category focuses on modeling semantic label relationships for an image. Encoder-decoder models such as CNN-RNN [5] and more recent models like graph convolutional networks (GCN) [6] are in this category. Inspired by current advances in image captioning [7], the CNN-RNN model encodes image content using a convolutional neural network (CNN) and feeds it into a recurrent neural network (RNN) to generate a label sequence. Nevertheless, one fundamental limitation of this approach is that the original training labels are orderless, while the RNN requires an ordered sequential list of labels as the input. But capturing label correlations using a GCN combined with a CNN has solved the previous limitation and has led to promising results in recent papers [8, 9]. In summary, GCN uses a predefined adjacency matrix as an explicit relationship between labels to improve the correlation of similar embedded words. Even though these approaches are practical, their architectures are complex and require external information such as natural language processing.

The second category asks whether such complicated solutions are essential for achieving high performance in AIA. Since CNNs have had remarkable achievements in single-label multi-class classification problems [10–12], there is a growing interest in using them to generate robust visual features for AIA. Specifically, it has been shown that carefully designed loss functions can significantly improve annotation accuracy while keeping standard architectures. For example,

some techniques, including the CNN + WARP (weighted approximate ranking pairwise) model [13] and CNN + LSEP (log-sum-exp pairwise) model [14], use ranking loss functions instead of traditional binary cross-entropy (BCE) loss [15, 16] to train deep CNNs in multi-label classification problems. However, a fundamental issue that has received little attention is the imbalance between positive and negative labels, meaning that the contribution of positive labels is much lower than the negative ones in images (see Fig. 1).

Traditionally, AIA techniques used a fixed threshold (e.g., select labels with p greater than a single threshold θ) or top- k values (e.g., pick top k results out of a ranked list) for label assignment, resulting in either over-labeling or under-labeling in images. Some recent methods, such as [15] and [17], have attempted to handle this issue differently, which obtained promising results.

In this research, we propose a novel unsymmetrical loss function for deep learning techniques to deal with the imbalance between positive and negative tags in high-vocabulary annotation datasets (datasets including a large number of words). Our loss function is based on two properties: first, we design a piecewise loss function to highlight the loss contribution from hard positives (low probability, less than 0.25) and semi-hard positives (medium probability, e.g., between 0.25 and 0.5) as well as easy ones (higher probability than 0.5) by a simple change on the positive part of the focal loss [18] and BCE loss. Second, owing to the abundance of negative tags, our approach down-weights their overall contribution from the loss function by increasing the exponential decay factor of the negative component of the focal loss. Finally, each label is given a different threshold according to the Matthews correlation coefficient (MCC) in order to improve the amount of the F1-score. Our main contributions are summarized as follows:

- Proposing a new loss function for deep convolutional networks that employs probability shifting, logit shifting, and variable exponential parameters to boost the contribution of positives to the loss function while decreasing the contribution of negatives.
- Proposing a threshold estimation method that uses a different threshold for each label rather than a single fixed threshold for all.

- Evaluation of the proposed method with commonly used annotation datasets.

There are two main criteria used to evaluate AIA methods: F1-score and N^+ . F1-score is the harmonic mean of precision (PR) and recall (RC). N^+ shows the number of labels with non-zero recalls. We use both to compare our approach with other AIA methods in the literature.

Following is an outline of the rest of this paper: In Sect. 2, we talk about the related works. Details and analysis of the proposed approach are presented in Sect. 3. The experimental results and evaluations are given in Sect. 4. Finally, we conclude our method in Sect. 5.

2 Related works

In this section, we first clarify how automatic image annotation (AIA) differs from image classification, then briefly explain the four non-deep annotation methods in the literature, and finally describe deep learning-based methods in greater detail since our proposed loss function is related to these techniques.

2.1 Comparison of image annotation and classification

Despite the resemblances between image classification and AIA, the two fields are fundamentally separate. Image classification consists of assigning a label or class to an entire image. Nonetheless, AIA produces a list of semantic tags to describe the visual content of each image. Consequently, AIA does not distinguish between the foregrounds and backgrounds of an image.

There are some differences between datasets of image annotation and image classification, which makes image annotation a challenging task. One of these challenges is the ambiguity in the contents of images. In classification tasks, each image is defined by a single label, which typically corresponds to the most prominent object in the image. However, it is impossible to annotate all concepts in an image due to the various perspectives of human annotators. Therefore, annotation datasets often include fewer labels than the original content of the image. Moreover, datasets of image annotation are highly unbalanced in terms of the number of images per label (see Fig. 2). As a result, labels with low frequencies are very tough to learn. Another challenge that has received less attention is the high imbalance between positive and negative labels in images of annotation datasets. The superiority of negative labels in number causes positive labels to contribute less to the learning process.

2.2 Non-deep methods

The cross-media relevance model (CMRM) [19] annotates an image based on a probability formula derived from a joint distribution of semantic tags and visual features of the entire image. Following the CMRM, Wang et al. [20] use a kernel-based estimation technique and multiple Bernoulli distributions to calculate the probability distribution of visual features and model semantic labels. In [21], the joint equal contribution (JEC) model is proposed, which determines the nearest neighbors of a picture using low-level image features such as shape, color, and texture information and some primary distance measures. The more recent 2PKNN model [22] combines the advantages of both image-to-image and image-to-label resemblance to solve the problem of image annotation. With its generative and discriminative methods, the SVM-DMBRM [23] is a powerful model for AIA. Weak labeling and unbalanced class issues are addressed by SVM and DMBRM, respectively. A graph structure is used in [24], in which nodes represent semantic labels and edges show co-occurrence links.

As mentioned earlier, one of the challenges in image annotation is that manual labels are often incomplete and unreliable. In order to solve this problem, a new model called multi-label learning with missing labels (MLML) is presented by [25, 26].

2.3 Deep learning-based methods

Recent advances in deep learning methods have led to breakthroughs in computer vision and image classification areas. They have also made considerable progress in the field of multi-label image annotation. The CNN-RNN model [5], CNN-GCN model [8], and CCA-KNN model [27] (which is based on the canonical correlation analysis (CCA) framework) use CNN and word embedding vectors for visual and textual feature extraction, respectively. Using word embedding vectors allows the semantic relationship between labels to play a vital role in image annotation. In order to model label correlations, Xue et al. [28] suggested a brand-new channel correlation network that is entirely based on CNNs. Visual features are convoluted by a new attention module to match the label and channel-wise feature map. Then, to properly examine the label correlation, they apply squeeze and excitation (SE) and convolution processes sequentially to get rid of unnecessary information. Niu et al. [15] extracted textual features from noisy tags through a multilayer perception subnetwork to enhance visual features extracted by a multi-scale CNN subnetwork. Eventually, these integrated features were used to annotate images in a fully connected layer. The diverse and distinct image annotation (D^2IA) model [29] creates a subset of related and unrelated labels using sequential sampling from a determinantal point process

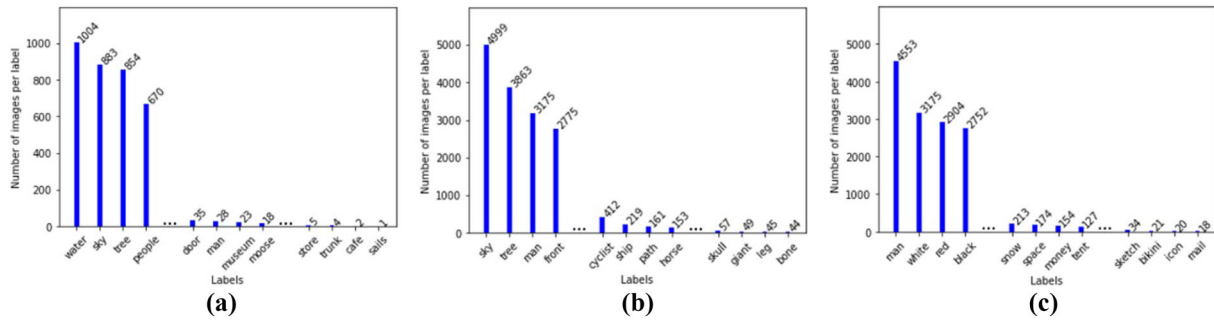


Fig. 2 Number of images per label in image annotation datasets (training data). **a** Corel 5k, **b** IAPR TC-12, and **c** ESP Game. The horizontal axis shows labels, and the vertical axis indicates the number of images per label

(DPP) model and employs a generative adversarial network (GAN) for the training process. In order to train annotation models using side information (e.g., semantic relations between labels) extracted by deep networks, more data is required. Ke et al. [16] designed an end-to-end AIA model consisting of a deep CNN (E2E-DCNN) and multi-label data augmentation that utilizes Wasserstein GAN for data augmentation. Although the multi-label data augmentation presented in [16] provides the amount of data required by the model and increases the number of tags with low frequency, it preserves the ratio between them, meaning that the number of all tags increases almost equally. As a result, the problem of imbalance between the number of images per label remains.

As an alternative to complex models, some methods modified the loss function used in CNNs to extract more appropriate visual features. For instance, Gong et al. [13] applied pairwise ranking to train deep CNNs for image annotation problems. The loss function is based on a multi-label form of the weighted approximate ranking pairwise (WARP) loss function. Li et al. [14] proposed an innovative loss function for pairwise ranking on the basis of a log-sum-exp pairwise (LSEP) function that is smooth everywhere and makes the optimization process more straightforward. To address the high imbalance between positive and negative labels in pictures, Ridnik et al. [30] suggested an asymmetric loss function, which has separate functionality for them.

3 The proposed methodology

Our proposed loss function is described in detail in this section. Section 3.1 reviews binary cross-entropy (BCE) and focal loss. Positive and negative parts of our loss function are introduced in Sects. 3.2 and 3.3, respectively. We present our asymmetric loss function in Sect. 3.4, which handles the positive–negative imbalance in annotation datasets. Finally, the threshold estimation algorithm is discussed in Sect. 3.5.

3.1 Preliminary

Deep learning-based approaches often treat image annotation as a multi-label classification task. Classification with multiple labels is typically converted into a series of binary classification problems. Given M labels, the network predicts the logit o_i of the i -th label independently, then the probability of each label is obtained by activating the logit with the sigmoid function as $p_i = \sigma(o_i)$. Assuming y_i is the ground-truth for the i -th label, the binary classification loss is computed as follows:

$$L = - \sum_{i=1}^M (y_i L_i^+ + (1 - y_i) L_i^-) \quad (1)$$

The positive and negative losses for i -th label are represented by L_i^+ and L_i^- , respectively. For brevity, the subscript i has been omitted from the following equations.

One of the most used loss functions in multi-label image annotation is the BCE loss function, which is calculated by:

$$\begin{cases} L_{BCE}^+ = \log(p) \\ L_{BCE}^- = \log(1 - p) \end{cases} \quad (2)$$

Another loss function that has been used primarily for object detection tasks and can handle the problem of class imbalance is focal loss [18]:

$$\begin{cases} L_{focal}^+ = \alpha_+ (1 - p)^\gamma \log(p) \\ L_{focal}^- = \alpha_- p^\gamma \log(1 - p) \end{cases}, \quad (3)$$

where γ is a focus parameter, and raising its amount increases the attention paid to the hard positives and hard negatives (higher probability than 0.75). The weight parameters α_+ and α_- are utilized to tackle the class imbalance issue.

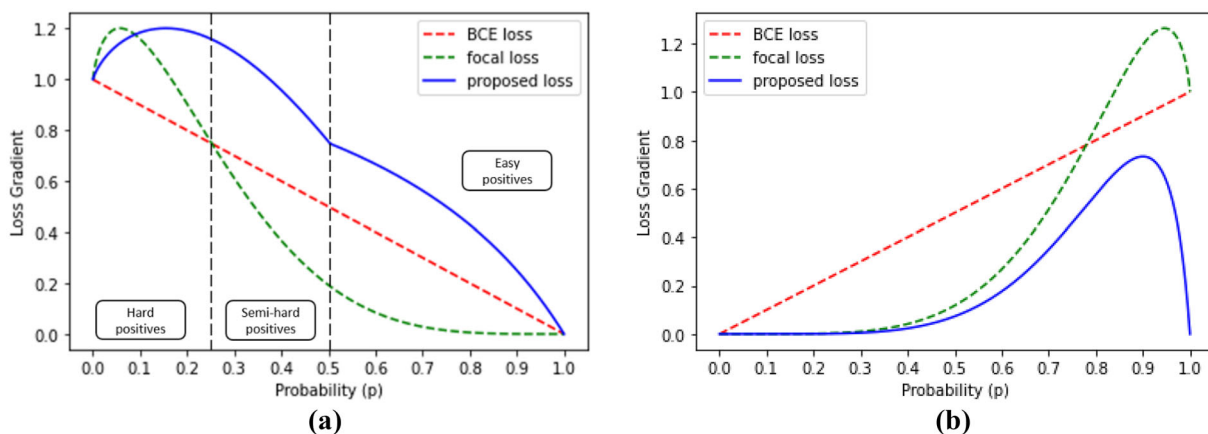


Fig. 3 Gradient Analysis of the proposed loss in comparison with binary cross-entropy (BCE) and focal loss. **a** shows that by choosing a proper margin (m_1), the proposed loss endows greater weight on hard and semi-hard positives than the focal loss with the same γ_+ parameter and

also highlights the weight of easy positives. **b** indicates that probability shifting eliminates negatives with very low probabilities, assigns lower weight to semi-hard and hard negatives over the focal loss, and discards mislabeled negatives and declines the gradient values for missing labels

3.2 Positive part of the loss function

What we are seeking is emphasizing gradients from positive labels during training. The fundamental explanation for this is the low contribution of positives to the entire labels in the high-vocabulary annotation dataset. Unlike BCE loss (Eq. 2), focal loss (Eq. 3) puts more attention on the hard positives. But, it neglects a large proportion of semi-hard positives and eliminates the gradients from easy ones. Moreover, since the formulation of focal loss is symmetrical, increasing the value of γ to decrease the contribution of easy negatives to the loss function aggravates the mentioned problem. As a result, learning features from positive labels might be under-emphasized in the network. We propose a piecewise loss function to highlight all positive labels rather than just focusing on the hard positives.

In the first instance, we subtract a margin from the logits, which reduces their amounts and gives them larger gradients. The formula is defined as:

$$p_{m_1} = \sigma(o - m_1) \tag{4}$$

where m_1 is a margin parameter. (The concept of margin is widely used in the loss function design, but [31] was the first work that combines margin and sigmoid activation). Next, we consider two sub-functions, one for hard and semi-hard positives and the other for easy positives. The first sub-function uses p_{m_1} instead of p in the positive part of the focal loss, resulting in more emphasizing gradients from hard and especially semi-hard positive labels. Similarly, the second sub-function uses p_{m_1} in place of p , but in the positive part of the BCE loss, which not only has no problems of focal loss for easy positives but accentuates their gradients. The

positive part of our loss function is defined as:

$$L^+ = \begin{cases} (1 - p_{m_1})^{\gamma_+} \log(p_{m_1}), & p_{m_1} < th \\ \log(p_{m_1}), & p_{m_1} \geq th \end{cases} \tag{5}$$

When p_{m_1} is less than th , we select the focal loss, and in case p_{m_1} is greater than th , we choose the BCE loss. th is determined in such a way that makes the loss gradients continuous.

We compare the gradients of our proposed loss function with the gradients of BCE and focal loss, which is helpful to understand how our loss function behaves. The loss gradients from positive labels are as follows:

$$\begin{aligned} \frac{dL^+}{do} &= \frac{\partial L^+}{\partial p_{m_1}} \frac{\partial p_{m_1}}{\partial o} \\ &= \begin{cases} (1 - p_{m_1})^{\gamma_+} \left[\frac{1}{p_{m_1}} - \frac{\gamma_+ \log(p_{m_1})}{1 - p_{m_1}} \right], & p_{m_1} < th \\ 1 - p_{m_1}, & p_{m_1} \geq th \end{cases} \end{aligned} \tag{6}$$

The gradients of different losses for positive labels are shown in Fig. 3a. Our suggested loss gives larger weight to semi-hard and easy positives compared to focal loss and BCE loss, respectively.

3.3 Negative part of the loss function

Due to the high number of negative labels in AIA tasks, we aim to reduce their contribution to the loss. Although this attenuation can be significantly satisfied by the negative part of focal loss for easy and semi-hard negatives, this is not always adequate because of the high imbalance

in image annotation tasks. According to [30], we eliminate very easy negatives after shifting their probabilities, which down-weights the contribution of all negative labels from the loss function. The shifted probability, p_{m_2} , is given by:

$$p_{m_2} = \max(p - m_2, 0), \quad (7)$$

where m_2 is a probability margin parameter. It is worth mentioning that the concept of logit shifting in Eq. (4) differs from probability shifting because it uses a non-linear sigmoid function. The negative part of our loss function uses p_{m_2} instead of p in the negative part of the focal loss:

$$L^- = (p_{m_2})^{\gamma_-} \log(1 - p_{m_2}). \quad (8)$$

Similarly, we investigate the behavior of loss gradients for negatives compared with those for BCE and focal loss:

$$\begin{aligned} \frac{dL^-}{do} &= \frac{\partial L^-}{\partial p} \frac{\partial p}{\partial o} \\ &= (p_{m_2})^{\gamma_-} \left[\frac{-1}{1 - p_{m_2}} + \frac{\gamma_- \log(1 - p_{m_2})}{p_{m_2}} \right] p(1 - p). \end{aligned} \quad (9)$$

In Fig. 3b, we show gradients of the loss function for negative labels and compare them to other losses. Overall, the suggested loss gives less weight to negatives and completely removes easy ones. This portion of the loss function can also deal with missing labels, tags that are highly probable in an image but labeled negatively. In contrast to the negative portions of BCE and Focal Loss, which penalize the model if it predicts very hard negatives, the mentioned loss function minimizes their gradient magnitudes. Thus, the model is not penalized too much if it annotates missing labels. In other words, if the model predicts an incorrect label with a high probability (around or greater than 0.9), this label is accepted as a missing label. (Recall that recent studies [16] have shown that manual annotation may miss some tags in images, which means different people may annotate the same image differently.)

3.4 Asymmetric loss function

The following is a definition of the proposed loss function:

$$\begin{cases} L^+ = \begin{cases} (1 - p_{m_1})^{\gamma_+} \log(p_{m_1}), & p_{m_1} < th \\ \log(p_{m_1}), & p_{m_1} \geq th, \end{cases} \\ L^- = (p_{m_2})^{\gamma_-} \log(1 - p_{m_2}) \end{cases} \quad (10)$$

where p_{m_1} and p_{m_2} are defined in Eqs. (4) and (7), respectively. Since the proposed loss function is asymmetric, it does not have the drawbacks of symmetric losses, such as focal loss, so we can set $\gamma_- > \gamma_+$. To address the positive–negative imbalance in image annotation datasets, the mentioned loss function behaves differently on negative and positive labels. This results in decreasing the contribution of negatives and an increase in the contribution of positives from the loss.

3.5 Threshold estimation algorithm

As convolutional network training is completed, we use a simple but effective method based on the Matthews correlation coefficient (MCC) to estimate a threshold for each label. Matthews operation calculates the correlation coefficient for each label separately, considering the actual and predicted values of that label, resulting in a number between -1 and 1. According to [32], MCC is more informative than F1-score and accuracy when evaluating binary classification problems since its formula considers the proportion of each component in the confusion matrix (true positive (TP), true negative (TN), false positive (FP), and false negative (FN)). It only receives a good score if the classifier performs well on both the negative and positive examples. Although accuracy and F1-score are frequently employed in statistics, both can be deceiving. When the dataset is unbalanced, for instance, accuracy is no longer an appropriate metric since it provides an overoptimistic estimate of the classifier's ability for the majority class. The F1-score also has some drawbacks. For example, it varies for class switching (if the positive class is renamed negative and vice versa) and is independent of the number of samples correctly identified as negative, whereas MCC is not susceptible to these issues. The correlation's formula is:

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP) + (TN + FN)}}. \quad (11)$$

First, we define a set of thresholds in the range [0.05, 0.7] with a step of 0.05. Then we compare the model's predicted values for each label in all training images or validation images (if any) with each of the mentioned thresholds separately. In case these values are greater than the thresholds, they are changed to 1 and otherwise to 0.

Next, we perform the Matthews operation using the obtained values on different thresholds and the grand-truth values for each label. Finally, the threshold with the best MCC will be chosen for each label. Algorithm 1 illustrates the pseudocode of the proposed method.

Algorithm 1 Pseudocode for threshold estimation method.

INPUT: Two $N \times M$ dimensional matrices $Y \in \{0,1\}^{N \times M}$ (actual value for labels) and $\hat{Y} \in \{0,1\}^{N \times M}$ (predicted value for labels). N and M represent the total number of training images and labels, respectively;

OUTPUT: Best thresholds

- 1) Define a set of thresholds: thresholds = [0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7];
- 2) Take the m -th column (m begins at 1) of \hat{Y} , which indicates the m -th label of every training image (\hat{Y}_m);
- 3) Select the i -th value of the threshold set (i begins at 1) and compare it to the items in \hat{Y}_m . It will be one if it is greater or equal to i ; otherwise, it will be zero;
- 4) Apply the Matthews operation to \hat{Y}_m and Y_m to get the result;
- 5) Continue following steps 3 and 4 for every other i ;
- 6) Identify the threshold value with the highest MCC and use it as the best threshold for the m -th label;
- 7) Steps 2–6 should be repeated in order to find the best threshold for each label.

4 Experiments

This section presents the results of experiments using the proposed loss function. In Sect. 4.1, the most commonly used evaluation metrics in image annotation are stated. A summary of benchmark datasets is provided in Sect. 4.2. The details of the implementation and methods for determining the parameters are discussed in Sect. 4.3. In Sect. 4.4 we compare our approach with existing loss functions to demonstrate its superiority. Then, we evaluate the results of the threshold estimation algorithm in Sect. 4.5. Eventually, comparisons with existing models will be demonstrated in Sect. 4.6.

4.1 Evaluation metrics

In order to determine the performance of AIA methods, several metrics are available; among them, the most commonly used are precision (PR) and recall (RC), F1-score, and N^+ .

Let N_i^p represent the number of images annotated with the i -th label, N_i^c represent how many images have been correctly annotated with the i -th label, and N_i^g denote the number of images annotated with the label i , using the ground-truth

data. Then the average PR and average RC are formulated as:

$$PR = \frac{1}{M} \sum_{i=1}^M \frac{N_i^c}{N_i^p}, \tag{12}$$

$$RC = \frac{1}{M} \sum_{i=1}^M \frac{N_i^c}{N_i^g}. \tag{13}$$

It is problematic to assess AIA models only by comparing PR and RC as these metrics conflict. Using the F1-score (the average harmonic of PR and RC) is proven to be more accurate for evaluating models. Another important metric that shows the number of labels with non-zero recalls (i.e., labels that were correctly annotated at least once by the model) is N^+ . F1-score and N^+ are calculated as follows:

$$F1 - score = \frac{2 \times PR \times RC}{PR + RC}, \tag{14}$$

$$N^+ = \sum_{i=1}^M Sgn(RC_i). \tag{15}$$

$$Sgn(x) = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \end{cases}$$

4.2 Datasets

There are three well-known datasets that are mostly used in AIA tasks. The first is Corel 5k [33], which has 5000 images (a training set of 4500 images and a test set of 500 images) with 260 labels. IAPR TC-12 [34] is the second dataset, which consists of 19,627 images (17,665 for training, 1962 for testing) that represent various scenes from everyday life, such as landscape images, action pictures, cities, buildings, sports, plants, and animals. A total of 291 labels are presented in this dataset. The more challenging dataset is ESP Game [35], which contains logos, drawings, scenery, and personal photos. In total, there are 20,770 images (18,689 for training and 2081 for testing) labeled with 268 keywords. The wide diversity of objects and the extensive number of words included in these datasets set them apart from the other data and made them more challenging. Table 1 provides details about these datasets.

In order to clarify the imbalance between positive and negative labels, we use an average positive–negative ratio, which is expressed as follows:

$$\text{positive} - \text{negative ratio} = \frac{LpI}{NoV - LpI}. \tag{16}$$

Table 1 An overview of three benchmark annotation datasets

Dataset	Number of images	Number of training images	Number of testing images	Number of vocabularies (NoV)	Labels per image (Lpl)	Image per label	Positive–negative ratio
Corel 5k	5,000	4,500	500	260	3.4	58.6	0.013
IAPR TC-12	19,627	17,665	1962	291	5.7	347.7	0.019
ESP Game	20,770	18,689	2081	268	4.7	326.7	0.017

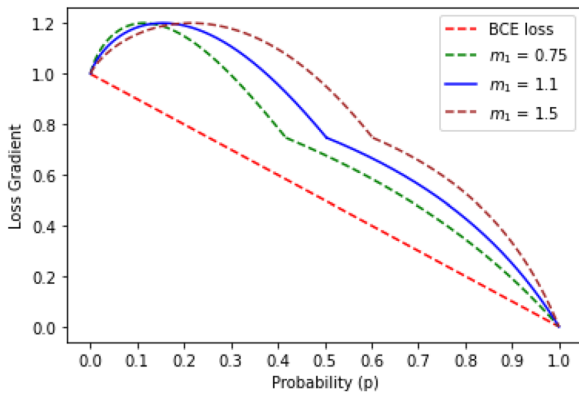


Fig. 4 Gradient analysis of the positive part of the proposed loss with different margins and $\gamma_+ = 3$ in comparison with BCE. Choosing an appropriate margin ($m_1 \in (0.5, 2)$) enables our loss to accentuate hard and semi-hard positives along with easy ones

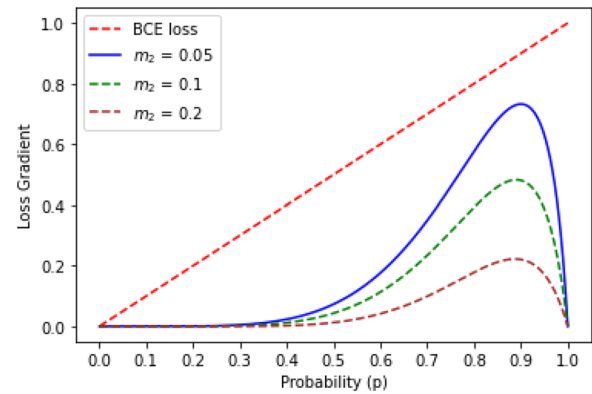


Fig. 5 Gradient analysis of the negative part of the proposed loss with different probability margins and $\gamma_- = 4$ in comparison with BCE. As the margin increases, semi-hard and hard negatives are given lower weights

4.3 Implementation

In the experiments, we use the convolutional architecture TResNet-M [36] as the backbone network for feature extraction. It is a new version of the ResNet [11] that enhances accuracy by utilizing several design tricks, including a Space-ToDepth stem, Anti-Alias downsampling, In-Place Activated BatchNorm, blocks selection, and squeeze-and-excitation layers. Input images in all datasets are uniformly resized to 448×448 . To optimize the network, we use Adam with a cycle learning rate schedule [37] with a maximum learning rate of 0.0001.¹

We set the $\gamma_+ = 3$ and the $\gamma_- = 4$, regarding that our goal is to pay more attention to hard and semi-hard positive cases and reduce the loss contribution from negative labels. According to Fig. 4, choosing a right margin (m_1) is crucial since it directly affects the weights assigned to different types of positive labels. On the one hand, a margin with low values allocates lower weight to semi-hard positives, but on the other hand, by setting high values for the margin, easy positives receive more weight than they should. As a result of experiments, selecting a proper probability margin (m_2)

can also be essential. A higher probability margin greatly down-weights the loss contribution from semi-hard and hard negatives, so annotating incorrect labels will not have much impact on learning and will lead to over-labeling. The negative gradients of our loss with different probability margins can be seen in Fig. 5.

4.4 Comparisons with existing loss functions

Since the suggested loss function is composed of two parts, and the negative part is quite similar to the recent state-of-the-art method ASL, we conduct an experiment based on ASL to confirm the piecewise positive part effectiveness.

First, we compare BCE loss with focal loss to explain that hard mining methods ($\gamma_+ > 0$), which down-weight the gradient magnitudes for semi-hard positives and eliminate them for easy positives, lead to a reduction in N^+ . Next, we apply AF + PS (asymmetric focusing and probability shifting), suggested in [30], to the negative part of the loss function to demonstrate that down-weighting incorrect labels (reducing the magnitude of their gradients) during the training process can significantly improve the final results. To make a fair comparison, we apply the positive component of both

¹ You can find our implementation at: <https://github.com/parham1998/Improving-Loss-Function-for-Deep-CNN-based-AIA>.

Table 2 Comparison between our approach and prior loss functions on the Corel 5k dataset

Methods	γ_+	γ_-	m_1	m_2	th	Corel 5k			
						PR	RC	F1	N +
BCE loss	0	0	0	0	–	0.453	0.385	0.416	147
Focal loss	3	3	0	0	–	0.452	0.366	0.405	139
BCE ⁺ and AF + PS [30]	0	4	0	0.05	–	0.480	0.522	0.500	179
Focal ⁺ and AF + PS	3	4	0	0.05	–	0.460	0.463	0.461	164
Proposed loss	3	4	1.1	0.05	0.25	0.466	0.554	0.506	189

BCE^+ (the positive part of binary cross-entropy loss), $Focal^+$ (the positive part of the focal loss), AF (asymmetric focusing), PS (probability shifting). γ_+ and γ_- are the focus parameters for the positive and negative parts of the loss function, respectively. m_1 and m_2 denote the logit and probability margins, respectively, whereas th stands for the threshold. The best performances are formatted in bold

BCE loss and focal loss to the positive part of the loss function. It can be deduced that setting the positive component of the focal loss as the positive part of the loss function will cause the same problem stated in the previous paragraph. In contrast, the network focuses more on semi-hard and easy positives when the positive half of the BCE loss is used as the positive part of the loss function and generates more labels. In circumstances with a large number of easy positive tags, strategies such as focal loss may be helpful.

Finally, we indicate that using a piecewise positive part, which has both focal loss features for hard positives and BCE loss features for easy and semi-hard positives, can considerably enhance results. Furthermore, we subtracted a margin from the logits of correct labels to emphasize gradients from these labels and increased their loss function contribution.

It is important to note that the suggested asymmetric loss begins the training process with a higher recall value (a higher number of incorrect labels) due to the reduction and increment in gradient magnitudes for negative and positive labels, respectively. Consequently, the model is not penalized excessively if it mistakenly tags incorrect labels, but it is penalized significantly in the case of not predicting correct labels. However, over the course of the training process, the recall values gradually become fixed and the precision value keeps rising (the model learns to remove incorrect labels). This is exactly the reverse of what occurs in BCE and focal loss. Table 2 summarizes the results of the experiments.

4.5 Evaluation of threshold estimation method

In deep learning-based AIA, labels are often allocated based on a fixed threshold (e.g., 0.5). Since the sigmoid activation function is typically applied to multi-label classification by deep networks, it is not far-fetched to use a threshold value of

Table 3 Evaluation of threshold estimation algorithm

Datasets	Methods	F1	N +
Corel 5k	Proposed loss	0.506	189
	Proposed loss + MCC	0.520	191
IAPR TC-12	Proposed loss	0.531	285
	Proposed loss + MCC	0.537	277
ESP Game	Proposed loss	0.452	261
	Proposed loss + MCC	0.461	255

The best performances are formatted in bold

0.5. Over-labeling or under-labeling are the major drawbacks of this method. In contrast, we introduced a novel threshold estimation method based on the Matthews correlation coefficient (MCC), which assigns different thresholds to each label.

On the one hand, the properties of the proposed loss function make the model capable of detecting missing labels, which exist in the image content but are not annotated by human annotators. Our threshold estimation solution, on the other hand, is highly dependent on the actual (ground truth) value of each label, which means human annotations play a significant role in its calculation. This contradiction causes some of the missing labels predicted by the model to be removed after calculating the new thresholds. In consequence, precision increases while recall decreases since a portion of correct labels may also have been omitted. In other words, it is a trade-off between predicting labels more conservatively (with a lower error rate) and more freely (with a higher error rate).

The results in Table 3 indicate that the algorithm increases F1-score in all three datasets and decreases N⁺ in IAPR TC-12 and ESP Game.

4.6 Comparisons with existing models

This section compares our approach with several classical and state-of-the-art models proposed in recent years. As shown in Table 4, traditional annotation models with poor performances have fallen into three categories: Generative models (e.g., MBRM [38]), Nearest neighbor models (e.g., 2PKNN [22]), and Discriminative models (e.g., MLDL [39]). The advanced deep-learning models proposed over the last few years, like CCA-KNN [27], RIA [5], VSE + 2PKNN-ML [40], PRM Deep [41], SEM [42], E2E-DCNN [16], SAIA [17], and SSL-AWF [43], are also expressed after them. (In each column, the best result is formatted in italic, and the second-best result is formatted in bold).

According to Table 4, our solution has higher recall values compared to other methods, demonstrating its superiority in detecting missing labels (correct labels that are not included in the ground-truth). Nevertheless, it adversely affects the F1-score, one of the main criteria in AIA, and precision. The threshold estimation method has been introduced to decrease missing labels predicted by the network and increase the F1-score. It is worth noting that we do not require architecture modifications, and our solution does not increase training times. This is different from previous solutions, which involved modifying the architecture (RNNs [5], GCNs [8]) and incorporating external information like label embedding.

Table 5 illustrates pictures annotated by the proposed technique on three benchmark annotation datasets. (There are pictures of Corel 5k, IAPR TC-12, and ESP Game in the first, second, and third rows, respectively.) The manual annotation is shown on the left column, and the automatic annotation is shown on the right one. Although the blue labels have not been manually annotated, they can convey the content of images well.

5 Conclusions










This paper comes up with a novel loss function for deep CNN-based image annotation. The introduced loss function can be classified as an unsymmetrical loss function, which performs differently on positive and negative labels. Since the number of negative labels is always much higher than positive ones in images, we reduce their contribution from the loss and increase the contribution of positive labels during network training. Analyzing our loss derivatives provided a deeper understanding of the loss properties. Additionally, a different threshold is used for each tag rather than a fixed threshold, which boosts the F1-score for all datasets. The results of a comprehensive analysis of three benchmark datasets with four evaluation metrics reveal that, in spite of its simplicity,

Table 4 Comparison between our approach and other AIA methods

Models	Methods				Corel 5k				IAPR TC-12				ESP Game			
	PR	RC	F1	N +	PR	RC	F1	N +	PR	RC	F1	N +	PR	RC	F1	N +
Generative models	0.24	0.25	0.24	122	0.24	0.23	0.23	223	0.24	0.23	0.23	223	0.18	0.19	0.18	209
Nearest neighbor	0.44	0.46	0.45	191	0.54	0.37	0.44	278	0.54	0.37	0.44	278	0.53	0.27	0.36	259
Discriminative	0.45	0.49	0.47	198	0.56	0.40	0.47	282	0.56	0.40	0.47	282	0.56	0.31	0.40	259
Deep learning	0.42	0.52	0.46	201	0.45	0.38	0.41	278	0.45	0.38	0.41	278	0.46	0.36	0.41	260
	0.32	0.35	0.32	139	0.35	0.34	0.33	267	0.35	0.34	0.33	267	0.32	0.32	0.31	249
	0.41	0.52	0.46	205	0.58	0.43	0.49	281	0.58	0.43	0.49	281	0.50	0.36	0.42	262
	0.453	0.5173	0.483	201	0.492	0.402	0.443	281	0.492	0.402	0.443	281	-	-	-	-
	0.370	0.520	0.430	-	0.410	0.390	0.400	-	0.410	0.390	0.400	-	0.380	0.420	0.400	-
	0.41	0.55	0.47	192	0.48	0.43	0.45	285	0.48	0.43	0.45	285	0.48	0.39	0.43	261
	0.554	0.565	0.56	212	0.655	0.438	0.525	291	0.655	0.438	0.525	291	0.637	0.387	0.482	268
	0.51	0.48	0.495	203	0.53	0.43	0.475	283	0.53	0.43	0.475	283	0.49	0.39	0.434	261
	0.466	0.554	0.506	189	0.503	0.562	0.537	285	0.503	0.562	0.537	285	0.423	0.484	0.452	261
	0.484	0.563	0.520	191	0.562	0.575	0.537	277	0.562	0.575	0.537	277	0.508	0.421	0.461	255

The best performances are formatted in bold, and the second-best results are formatted in italic

Table 5 Annotation examples of the proposed approach

Images	Manual annotation	Automatic annotation
	Light, shore	Sky, water, boats, sunset
	Water, river, deer, white-tailed	Water, tree, rocks, deer, white-tailed, horns, elk
	Birds, booby	Grass, birds, booby, nest
	Bike, hand, hill, man, mountain, road, woman	Bike, <i>forest</i> , helmet, hill, man, mountain, road, sky, woman
	Cloud, desert, dune, man, sky, woman	Cloud, desert, dune, man, sand, sky, woman
	Boy, brick, front, room, shirt, wall	Boy, classroom, kid, man, pullover, shirt, table, wall
	Man, photo, red, tie	Black, man, people, red, tie, suit
	Cartoon, glasses, smile	Brown, cartoon, glasses, man, orange, yellow
	Circle, eat, food, plate, red	Blue, circle, food, plate, white

Labels that were correctly classified while not presented in the manual annotation are formatted in bold, and labels that were wrongly classified are formatted in italics

our loss performs better than old-fashioned loss functions and can compete with state-of-the-art methods.

Author contributions AS: Conceptualization, Methodology, Writing-original-draft, Formal analysis, Visualization. AA: Methodology, Validation, Resources, Writing-review & editing, Supervision.

Funding The authors did not receive support from any organization for the submitted work.

Data availability The data that support the findings of this study are openly available in the public data repository at: Corel-5 k: <https://www.kaggle.com/datasets/parhamsalar/corel5k> reference number [33]. IAPR TC-12: <https://www.kaggle.com/datasets/parhamsalar/iaprtc12> reference number [34]. Esp Game: <https://www.kaggle.com/datasets/parhamsalar/espgame> reference number [35].

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

- Cheng, Q., Zhang, Q., Fu, P., Tu, C., Li, S.: A survey and analysis on automatic image annotation. *Pattern Recognit* **79**, 242–259 (2018). <https://doi.org/10.1016/j.patcog.2018.02.017>
- Tsoumakas, G., Katakis, I.: Multi-label classification. *Int. J. Data Warehous. Min.* **3**(3), 1–13 (2007). <https://doi.org/10.4018/jdwm.2007070101>
- Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. *Mach. Learn.* **85**(3), 333–359 (2011). <https://doi.org/10.1007/s10994-011-5256-5>
- Zhu, G., Yan, S., Ma, Y.: Image tag refinement towards low-rank, content-tag prior and error sparsity. In: *MM'10—Proceedings of the ACM Multimedia 2010 International Conference*. 2010, pp. 461–470. <https://doi.org/10.1145/1873951.1874028>.
- Jin, J., Nakayama, H.: Annotation order matters: recurrent image annotator for arbitrary length image tagging. In: *Proceedings—International Conference on Pattern Recognition*. pp. 2452–2457, (2016). <https://doi.org/10.1109/ICPR.2016.7900004>.
- Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. *arXiv Prepr. arXiv1609.02907*. (2016)
- Liu, X., Xu, Q., Wang, N.: A survey on deep neural network-based image captioning. *Vis. Comput.* **35**(3), 445–470 (2019). <https://doi.org/10.1007/s00371-018-1566-y>
- Chen, Z.M., Wei, X.S., Wang, P., Guo, Y.: Multi-label image recognition with graph convolutional networks. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019, pp. 5172–5181. (2019). <https://doi.org/10.1109/CVPR.2019.00532>
- Lotfi, F., Jamzad, M., Beigy, H.: Automatic image annotation using tag relations and graph convolutional networks. In: *Proceedings of the 5th international conference on pattern recognition and image analysis, IPRIA 2021*, pp. 1–6. (2021). <https://doi.org/10.1109/IPRIA53572.2021.9483536>
- Szegedy, C. et al.: Going deeper with convolutions. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07–12, pp. 1–9. (2015). <https://doi.org/10.1109/CVPR.2015.7298594>

11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2016, pp. 770–778. (2016). <https://doi.org/10.1109/CVPR.2016.90>
12. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-ResNet and the impact of residual connections on learning. In: 31st AAAI Conference on Artificial Intelligence, AAAI 2017, 2017, vol. 31, no. 1, pp. 4278–4284. <https://doi.org/10.1609/aaai.v31i1.11231>
13. Gong, Y., Jia, Y., Leung, T., Toshev, A., Ioffe, S.: Deep convolutional ranking for multilabel image annotation. arXiv Prepr. arXiv1312.4894. (2013)
14. Li, Y., Song, Y., Luo, J.: Improving pairwise ranking for multi-label image classification. In: Proceedings—30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, vol. 2017, pp. 1837–1845. (2017). <https://doi.org/10.1109/CVPR.2017.199>
15. Niu, Y., Lu, Z., Wen, J.R., Xiang, T., Chang, S.F.: Multi-modal multi-scale deep learning for large-scale image annotation. IEEE Trans. Image Process. **28**(4), 1720–1731 (2019). <https://doi.org/10.1109/TIP.2018.2881928>
16. Ke, X., Zou, J., Niu, Y.: End-to-end automatic image annotation based on deep CNN and multi-label data augmentation. IEEE Trans. Multimed. **21**(8), 2093–2106 (2019). <https://doi.org/10.1109/TMM.2019.2895511>
17. Khatchatoorian, A.G., Jamzad, M.: Architecture to improve the accuracy of automatic image annotation systems. IET Comput. Vis. **14**(5), 214–223 (2020). <https://doi.org/10.1049/iet-cvi.2019.0500>
18. Lin, T. Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, vol. 2017, pp. 2999–3007. (2017). <https://doi.org/10.1109/ICCV.2017.324>
19. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pp. 119–126. (2003). <https://doi.org/10.1145/860435.860459>
20. Wang, M., Zhou, X.D., Zhang, J.Q., Xu, H.T., Le Shi, B.: Image auto-annotation via an extended generative language model. Ruan Jian Xue Bao/Journal Softw. **19**(9), 2449–2460 (2008). <https://doi.org/10.3724/SP.J.1001.2008.02449>
21. Makadia, A., Pavlovic, V., Kumar, S.: A New Baseline for Image Annotation. In: European Conference on Computer Vision, pp. 316–329. Springer, (2008). https://doi.org/10.1007/978-3-540-88690-7_24
22. Verma, Y., Jawahar, C.V.: Image annotation using metric learning in semantic neighbourhoods. In: lecture notes in computer science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 7574 LNCS, no. PART 3, pp. 836–849. (2012). https://doi.org/10.1007/978-3-642-33712-3_60
23. Murthy, V.N., Can, E.F., Manmatha, R.: A hybrid model for automatic image annotation. In: ICMR 2014—Proceedings of the ACM international conference on multimedia retrieval 2014, pp. 369–376. (2014). <https://doi.org/10.1145/2578726.2578774>
24. Feng, L., Bhanu, B.: Semantic concept co-occurrence patterns for image annotation and retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **38**(4), 785–799 (2016). <https://doi.org/10.1109/TPAMI.2015.2469281>
25. Wu, B., Lyu, S., Ghanem, B.: ML-MG: Multi-label learning with missing labels using a mixed graph. In: 2015 IEEE International Conference on Computer Vision (ICCV), vol. 2015 Inter, pp. 4157–4165. (2015). <https://doi.org/10.1109/ICCV.2015.473>
26. Wu, B., Liu, Z., Wang, S., Hu, B.G., Ji, Q.: Multi-label learning with missing labels. In: Proceedings—International Conference on Pattern Recognition, pp. 1964–1968. (2014). <https://doi.org/10.1109/ICPR.2014.343>
27. Murthy, V.N., Maji, S., Manmatha, R.: Automatic image annotation using deep learning representations. In: ICMR 2015—Proceedings of the 2015 ACM International Conference on Multimedia Retrieval, pp. 603–606. (2015). <https://doi.org/10.1145/2671188.2749391>
28. Xue, L., Jiang, D., Wang, R., Yang, J., Hu, M.: Learning semantic dependencies with channel correlation for multi-label classification. Vis. Comput. **36**(7), 1325–1335 (2020). <https://doi.org/10.1007/s00371-019-01731-5>
29. Wu, B., Chen, W., Sun, P., Liu, Ghanem, B., Lyu, S.: Tagging like humans: diverse and distinct image annotation. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp. 7967–7975. (2018). <https://doi.org/10.1109/CVPR.2018.00831>
30. Ridnik, T., et al.: Asymmetric loss for multi-label classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 82–91. (2021). <https://doi.org/10.1109/ICCV48922.2021.00015>
31. Zhang, Y. et al.: Simple and robust loss design for multi-label learning with missing labels. arXiv Prepr. arXiv2112.07368. (2021). Available: <http://arxiv.org/abs/2112.07368>
32. Chicco, D., Jurman, G.: The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics **21**(1), 6 (2020). <https://doi.org/10.1186/s12864-019-6413-7>
33. Duygulu, P., Barnard, K., de Freitas, J.F.G., Forsyth, D.A.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 2353, pp. 97–112. (2002). https://doi.org/10.1007/3-540-47979-1_7
34. Grubinger, M.: Analysis and evaluation of visual information systems performance. (2007)
35. Von Ahn, L., Dabbish, L.: Labeling images with a computer game. In: Conference on Human Factors in Computing Systems—Proceedings, pp. 319–326. (2004). <https://doi.org/10.1145/985692.985733>
36. Ridnik, T., Lawen, H., Noy, A., Ben, E., Sharir, B.G., Friedman, I.: TResNet: High performance GPU-dedicated architecture. In: Proceedings—2021 IEEE Winter Conference on Applications of Computer Vision, WACV 2021, pp. 1399–1408. (2021). <https://doi.org/10.1109/WACV48630.2021.00144>
37. Smith, L.N., Topin, N.: Super-convergence: very fast training of neural networks using large learning rates. In: Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, vol. 11006, p. 36. (2019). <https://doi.org/10.1117/12.2520589>
38. Feng, S.L., Manmatha, R., Lavrenko, V.: Multiple Bernoulli relevance models for image and video annotation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 1002–1009. (2004). <https://doi.org/10.1109/cvpr.2004.1315274>
39. Jing, X.Y., Wu, F., Li, Z., Hu, R., Zhang, D.: Multi-label dictionary learning for image annotation. IEEE Trans. Image Process. **25**(6), 2712–2725 (2016). <https://doi.org/10.1109/TIP.2016.2549459>
40. Zhang, W., Hu, H., Hu, H.: Training visual-semantic embedding network for boosting automatic image annotation. Neural Process. Lett. **48**(3), 1503–1519 (2018). <https://doi.org/10.1007/s11063-017-9753-9>
41. Khatchatoorian, A.G., Jamzad, M.: An image annotation rectifying method based on deep features. In: ACM International Conference Proceeding Series, pp. 88–92. (2018). <https://doi.org/10.1145/3193025.3193035>

42. Ma, Y., Liu, Y., Xie, Q., Li, L.: CNN-feature based automatic image annotation method. *Multimed. Tools Appl.* **78**(3), 3767–3780 (2019). <https://doi.org/10.1007/s11042-018-6038-x>
43. Li, Z., Lin, L., Zhang, C., Ma, H., Zhao, W., Shi, Z.: A Semi-supervised learning approach based on adaptive weighted fusion for automatic image annotation. *ACM Trans. Multimed. Comput. Commun. Appl.* **17**(1), 1–23 (2021). <https://doi.org/10.1145/3426974>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



Ali Salar attended the K. N. Toosi University of Technology and earned his Bachelor of Science degree in Computer Engineering in 2020. At the moment, he is pursuing his master's degree in Artificial Intelligence and Robotics at K. N. Toosi University. He is currently working as a researcher at the Intelligent Processing of information and software testing Lab. Deep learning, computer vision, and image processing are all areas of research that interest him. He

is especially attentive in working on multi-label image annotation and image captioning using deep models.



Ali Ahmadi received his M.Sc. and Ph.D. in Artificial Intelligence and Soft Computing from Osaka Prefecture University, Japan in 2001 and 2004, respectively. He worked as a senior researcher in Nano-device Research Center at Hiroshima University, Japan during 2004–2007. He has been with K.N. Toosi University of Technology from 2007 where he has worked as the head of Faculty of Computer Engineering during 2012–2016 and since then as associate professor in AI department. He is currently an adjunct faculty in York University, Canada. His research interests include: semantic image and text mining, image captioning, information fusion, HCI models, virtual and mix reality.