Proceedings of the 2nd International Conference on Computing Innovation and Applied Physics (CONF-CIAP 2023)
DOI: 10.54254/2753-8818/5/20230455

Take a close look at deep learning applications from loss view

Bowen Shi

Chengdu University of Technology, Erxianqiao East Third Road, Chengdu City, Sichuan Province, China

18831244448@163.com

Abstract. Deep learning is an intrinsic learning style and representation through which knowledge acquired helps interpret data such as words, images and sounds. The ultimate aim is to enable computers to analyse and recognize data, such as words, images and sounds, in the same way as humans. In artificial intelligence, loss function is a very effective method. In the deep study, the loss function is used to determine the relationship between the target and the prediction. This paper analyzes and summarizes the loss functions in face recognition, object detection and face recognition, and focuses on some key loss functions.

Keywords: loss function, face recognition, object detection, semantic segmentation.

1. Introduction

Deep learning has reached its peak in recent years. As a particularly important part of machine learning, in order to explore higher-order and more complex functions and further develop, simulation experiments like human brain and cognition have been carried out for research. And it has made great breakthroughs in visual structure, natural language processing, data mining and other composite applications [1]. Deep architectures include several levels of nonlinear operations, such as neural networks with many hidden layers, in multilayer graphical models with many latent variables, or in complex formulas that reuse many sub formulas [2]. However, in the deep learning of neural networks, there is a certain deviation between the actual learning and the prediction of the model, and the input data of the loss function is often based on the distance measurement. On this basis, an appropriate sampling loss function is used to measure the actual value of the feature space and the distance predicted by the model.

The loss function can better reflect the difference between the model and the real situation. The knowledge of the attrition function allows for a better analysis and understanding of the subsequent optimal methods (such as slope reduction, etc.). In all statistical models, the loss function plays a very key role, that is, defining the performance evaluation index of a model, and determining the learning parameters of the model at the minimum cost. The loss function transforms a theoretical proposition into a practical one. Building highly accurate predictors requires continuously iterating the problem by asking questions, modeling the problem with the chosen method, and testing. The only criterion by which a statistical model is scrutinized is its performance - how accurate the model's decisions are. This requires a way to measure how far a particular iteration of the model is from the actual value. This is where the loss function comes in.

DOI: 10.54254/2753-8818/5/20230455

Another superiority of the loss function lies in the specificity that a customed loss function can solve the specific problem came across in various applications, such as the dice loss in semantic segmentation. From this perspective, In this paper, deep learning in face recognition, object detection, semantic segmentation and other aspects are systematically summarized and analyzed, and corresponding to the loss function

2. Face recognition

Facial recognition technology, as has been widely applied in monitoring, security, identification and mood, more complex domains such as age, the role of play will also receive the influence of many factors, resulting in test and actual gap, shape similarity, including human form into effect and clarity and other internal factors and external factors, for the computer. The image is represented by a multi-dimensional number matrix, so the recognition task is difficult8/8/2023.

The basic process of face recognition, first for face detection, through the face detection algorithm to obtain the face position information, usually with a rectangular upper left corner coordinate value and length and width of the face region in the image position, according to the position information from the image cropped face region image. The second is the key point detection, through the face key algorithm to obtain the coordinates of each point on the face region, such as around eyes and mouth in the image coordinate values, by affine transformation to these coordinates and predefined templates point (that is defined in advance what position) of these key points should be located in the image is aligned, All face images will be normalized to the same pattern. Finally, face modeling will be used to extract face features from standardized face images for subsequent comparison or verification [3].

For face recognition technology in the existence of the loss, the most commonly used are Softmax, which the formula is

$$f(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{1}$$

and Hardmax, which the formula is

$$f(x_i) = \frac{x_i}{\sum_j x_j} \tag{2}$$

Compared with Hardmax, Softmax is easy to achieve the ultimate goal of one-hot under the same output features due to its convergence, which reduces the difficulty of training. However, Softmax does not encourage too much feature separation, instead, it encourages feature differentiation of different categories, and does not require intra class compactness and inter-class separation. It is not completely good at face recognition task, need later transformation.

The contrastive loss is mainly used to reduce the dimension. By reducing the distance between two similar samples or increasing the distance between different samples in the training set [4], the loss function is

$$L = \frac{1}{2N}\sum_{n=1}^{N} y \cdot \|a_n - b_n\|^2 + (1-y) \cdot max\left(\text{margin} - \|a_n - b_n\|^2, 0\right) \tag{3}$$

Where d = | | the an - bn | | 2, represents the Euclidean distance of two sample characteristics, y label samples for two matches, y = 1 on behalf of the two samples or similar match, y = 0 represents mismatch, margin for setting the threshold. The value of y (0 or 1) is used to distinguish the sample types (different or similar). Obviously, if y=1, this equation represents the initial similar sample. In the case of large Euclidean spacing, this model is not suitable for the existing models. If y=0, it means that the sample is not like. If the Euclidean distance is larger, the loss is less, and the expected purpose is achieved. However, due to the limit of fixed margin, it has some distortion For contrastive loss, choosing hard example usually leads to faster convergence.

Triplet Loss, As the name suggests, it has a structural atmosphere: a standard image, a positive sample (same as the target), and a negative sample (different from the target). In the field of face recognition, Triplet Loss is often used to extract the embedding of the face, and it is also an improvement

DOI: 10.54254/2753-8818/5/20230455

of Contrastive Loss. For samples with the same label, their embedding space will be narrowed as far as possible, while for samples with different labels, their embedding distance will be as far as possible [5]. Its loss function is denoted by

$$\sum_i^N \left[\left\|f(x_i^a) - f(x_i^p)\right\|_2^2 - \left\|f(x_i^a) - f(x_i^n)\right\|_2^2 + \alpha\right]_+ \quad (4)$$

Triplet Loss has excellent recognition capability and is especially suitable for image classification. Triplet Loss is able to better represent two different input vectors when there is a great deal of similarity between them, so that the classification can be done well.

Compared with other classification Loss functions, Triplet Loss can obtain better details. Specifically, Triplet Loss can be set at specific thresholds, depending on the training needs of the model. The network architecture of Triplet Loss usually sets a critical limit during training. The designer can adjust the spacing between normal and negative states by adjusting the tolerance. Triplet Loss is a good method, but its disadvantages are as follows: the selection of Triplet will cause the uneven distribution of data, which makes the model unstable in the learning process, and the parameters will be adjusted slowly in the learning process. Triplet Loss is more likely to be overfitted than classification loss. Therefore, in most cases, we will use this method for model preview, or combined with softmax function (margin loss function) [5].

In addition, Center Loss, L-softmax that strengthens classification conditions and Center Loss are all common Loss types in the application field of face recognition technology.

3. Object detection

Object detection refers to technology in computer vision and image processing, which means detecting visual images and videos in specific semantic objects. At the same time, developments like this, continuous neural networks and related learning systems, will strengthen neural network algorithms and have a fundamental impact on object detection methods that are considered learning systems.

Focal Loss is one of the most common target detection loss functions. This method mainly solves the problem of serious imbalance between positive and negative samples in a kind of sample detection algorithm, and reduces a large number of simple negative samples, which is a kind of difficult sample mining. Focal Loss is an improved cross-entropy loss function. Its loss function is

$$FL(p_t) = -(1 - p_t)^{\gamma} \log(p_t) \quad (5)$$

Among them,

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{otherwise} \end{cases} \quad (6)$$

γ is constant, when it is 0, FL is consistent with the ordinary cross-entropy loss function. The parameter γ is used to distinguish easily distinguishable samples from hard distinguishable samples. It can reduce the loss weight of simple samples (large number), make the loss function more focused on difficult samples (number), and prevent simple samples from dominating the whole loss function. Although Focal loss solves the problem of imbalance between positive and negative samples (foreground and background), reduces the weight of simple samples, and makes the loss function pay more attention to difficult samples, the model has been interfered and the model overfitting deteriorates, and the parameters need to be manually tested, otherwise the fitting effect will be directly affected [6].

The IoU loss function is based on the estimated box and labeled box IoU (set ratio). The prediction box is denoted as P, and the annotation box is denoted as G. The corresponding IoU can be expressed as:

$$IoU = \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|} \quad (7)$$

Which is the ratio of the intersection and the union of two boxes. IoU Loss is defined as:

DOI: 10.54254/2753-8818/5/20230455

$$\mathcal{L}_{IoU} = 1 - \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|} \tag{8}$$

It is easy to see from the formula that IoU is used to directly find the slope of the regression. However, if the bins do not meet, then IoU=0, the bins that are close to each other and the bins that are far away will lose their slopes, so the optimal solution cannot be performed. This is the problem of its IoU loss. GIoU Loss, adding a penalty term to IoU Loss, When the distance between Bboxes is larger, the penalty term will be larger, and it the gradient problem of IoU is solvedIts loss function is

$$\mathcal{L}_{GIoU} = 1 - IoU + \frac{|C - B \cup B^{gt}|}{|C|} \tag{9}$$

This is particularly similar to the case of IoU Loss. The initial IoU is 0,1, and the GIoU is -1,1. If the two images overlap, IoU=1; If the two pictures are infinitely far apart, then IoU is equal to 0 and GIoU=-1. GIoU not only pays attention to overlap, but also pays attention to non-overlap, which can better reflect the overlap of the two images. It can be seen that the estimation of image overlap by GIoU has improved to some extent, but it still cannot reflect the similarity between images [7].

4. Semantic segmentation

Semantic segmentation is to classify a pixel with different meanings and classify it as a pixel. This is a computer vision orientation problem, which needs to collect raw data and then transform them into images with highlighting effect [8]. It plays a very important role in image recognition in autonomous driving, robot and image search engine.

The first conventional, as well as the most used Loss design for semantic segmentation is the Cross Entropy Loss. This method is used to judge the relationship between the real output and the expected output, and to explain the gap between the real output and the expected output. In other words, when the mutual entropy is low, the distribution of these two possibilities will be more similar. The crossing of each pixel is analyzed, and the prediction of each pixel classification (or possibly assigned vector) is compared with our single statistical heat vector (one -hot form). The formula for Cross Entropy is

$$H(p, q) = -\sum_{i=1}^{n} p(x_i) \log\left(q(x_i)\right) \tag{10}$$

When p=q, cross-entropy achieves the minimum value, so cross-entropy can be used to compare the coincidence between one distribution and another. Cross-entropy is close to entropy, and q is a better approximation top. In fact, the closer the output of the model is to the expected output, the smaller the cross-entropy will be, which meets the needs.

Weighted cross entropy is a parameter used to represent the importance of sampling in LOSS. When the number of samples is small, its role in loss should be strengthened, and when the number of samples is large, its role in loss should be reduced. The definition for

$$WCE = -\frac{1}{N}\sum_{i=1}^{N} \alpha y_i \log\, p_i + (1 - y_i) \log\,(1 - p_i) \tag{11}$$

Among them,

$$\alpha = \frac{N_{neg}}{N_{\text{pos}}} \tag{12}$$

Compared with the cross-entropy introduced above, the weighted cross-entropy is further realized. The weighted interaction entropy is the weight of each classification, which makes the classification of samples with lower ability more important [9, 10].

Dice Loss derives from the Dice Coefficient, a later developed statistical indicator that measures the coincidence and similarity of two samples. Its definition is as follows.

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|} \tag{13}$$

Or what it means in the dichotomous problem is

DOI: 10.54254/2753-8818/5/20230455

$$DSC = \frac{2TP}{2TP+FP+FN} \tag{14}$$

However, Dice Loss is relatively suitable for the case of extremely uneven samples. In general, the use of Dice loss will cause adverse effects on back propagation and easily make training unstable.

Generalized Dice Loss is a combination of multiple Dice losses. When a small object has a pixel prediction error, the Dice coefficient will fluctuate greatly, resulting in a large gradient change. Dice Loss used an index to quantitatively describe the prediction of small targets. In addition, Dice Loss is a special type of partition Loss. Multiple Dice loss is commonly used for segmentation of multiple lesions. GDL' loss function is

$$GDL = 1 - 2\frac{\sum_{l=1}^{2} w_l \sum_n r_{ln} p_{ln}}{\sum_{l=1}^{2} w_l \sum_n r_{ln} + p_{ln}} \tag{15}$$

Where r_{ln} represents the true pixel category of category l at the n TH position, while p_{ln} represents The corresponding predicted odds and w_l represents the weight of each category. The formula for w_l is

$$w_l = \frac{1}{\sum_{i=1}^{n} r_{ln}^2} \tag{16}$$

In addition, the cross-entropy loss and mutation forms in taxonomy also include Categorical cross-entropy loss, Binary cross-entropy loss, etc. Combining different loss functions such as Dice+Cross entropy is a method of semantic segmentation.

In summary, cross-drop loss predicts each pixel as an independent sample, while Dice loss and IoU loss look at the final predicted output in a more holistic way.

5. Conclusion

In this paper, the corresponding loss functions of face recognition, object detection and semantic segmentation are summarized and analyzed, and the common and key loss functions are systematically introduced. It is not difficult to see that the use of artificial intelligence has the characteristics of humanlike inhuman, technical accuracy, great potential and so on, can effectively solve various problems. If the research and development of artificial intelligence can be effectively used and controlled, artificial intelligence will be our human sword Deep learning belongs to a paradigm of artificial intelligence, whose algorithmic nature is to minimize the loss function by constantly adjusting the network parameters. One goal of cognitive computing is to build applications that simulate and mimic the characteristics of the human mind. The natural way to do this is to develop brain-based computational models of cognitive function. When selecting the appropriate loss function to solve the corresponding problem, the main features that can best express the data should be selected first to construct the feature space based on distance or probability distribution metric. Then, a reasonable feature normalization method is chosen to keep the core content of the original data after feature vector transformation. Finally, the reasonable loss function is selected, and the parameters of the model are adjusted continuously according to the loss on the basis of the experiment, so as to achieve the classification as far as possible. The limitation of this paper is that it does not include all the loss functions except face recognition, object detection and semantic segmentation. It only explains the key loss functions of the above three aspects in detail, but not all of them. This is also the place that needs to be improved in the later stage. In the future, the role of loss function in the training of artificial neural network (ANN), the influence on the generalization ability of ANN model and other attributes will be further studied.

References

[1] Janocha, K., & Czarnecki, W. M. (2017). On loss functions for deep neural networks in classification. *arXiv preprint arXiv:1702.05659*.

[2] Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and trends® in Machine Learning*, *2*(1), 1-127.

DOI: 10.54254/2753-8818/5/20230455

[3] Arya, S., Pratap, N., & Bhatia, K. (2015). Future of face recognition: a review. *Procedia Computer Science*, *58*, 578-585.

[4] Masi, I., Wu, Y., Hassner, T., & Natarajan, P. (2018, October). Deep face recognition: A survey. In *2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*(pp. 471-478). IEEE.

[5] Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 815-823).

[6] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980-2988).

[7] Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 658-666).

[8] Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., & Garcia-Rodriguez, J. (2017). A review on deep learning techniques applied to semantic segmentation. *arXiv preprint arXiv:1704.06857*.

[9] Asgari Taghanaki, S., Abhishek, K., Cohen, J. P., Cohen-Adad, J., & Hamarneh, G. (2021). Deep semantic segmentation of natural and medical images: a review. *Artificial Intelligence Review*, *54*(1), 137-178.

[10] Crum, W. R., Camara, O., & Hill, D. L. (2006). Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE transactions on medical imaging*, *25*(11), 1451-1461.