УДК 519.6

V. Andreieva, N. Shvai

DOI: https://doi.org/10.18523/2617-7080320203-10

GENERALIZATION OF CROSS-ENTROPY LOSS FUNCTION FOR IMAGE CLASSIFICATION

Classification task is one of the most common tasks in machine learning. This supervised learning problem consists in assigning each input to one of a finite number of discrete categories. Classification task appears naturally in numerous applications, such as medical image processing, speech recognition, maintenance systems, accident detection, autonomous driving etc.

In the last decade methods of deep learning have proven to be extremely efficient in multiple machine learning problems, including classification. Whereas the neural network architecture might depend a lot on data type and restrictions posed by the nature of the problem (for example, real-time applications), the process of its training (i.e. finding model's parameters) is almost always presented as loss function optimization problem.

Cross-entropy is a loss function often used for multiclass classification problems, as it allows to achieve high accuracy results.

Here we propose to use a generalized version of this loss based on Renyi divergence and entropy. We remark that in case of binary labels proposed generalization is reduced to cross-entropy, thus we work in the context of soft labels. Specifically, we consider a problem of image classification being solved by application of convolution neural networks with mixup regularizer. The latter expands the training set by taking convex combination of pairs of data samples and corresponding labels. Consequently, labels are no longer binary (corresponding to single class), but have a form of vector of probabilities. In such settings cross-entropy and proposed generalization with Renyi divergence and entropy are distinct, and their comparison makes sense.

To measure effectiveness of the proposed loss function we consider image classification problem on benchmark CIFAR-10 dataset. This dataset consists of 60000 images belonging to 10 classes, where images are color and have the size of 32×32. Training set consists of 50000 images, and the test set contains 10000 images.

For the convolution neural network, we follow [1] where the same classification task was studied with respect to different loss functions and consider the same neural network architecture in order to obtain comparable results.

Experiments demonstrate superiority of the proposed method over cross-entropy for loss function parameter value $\alpha < 1$. For parameter value $\alpha > 1$ proposed method shows worse results than cross-entropy loss function. Finally, parameter value $\alpha = 1$ corresponds to cross-entropy.

Keywords: loss function, image classification, Renyi entropy, Renyi divergence.

Introduction

In recent years, deep learning methods have been showing steady success in various areas of applications, such as computer vision, natural language processing, autonomous driving etc. This success can be attributed to high performance of deep learning models in comparison to other methods, notably for unstructured data, such as images or text. Naturally, a big part of research community effort is aimed at the model performance improvement.

Introduction of new loss functions can often lead to model qualities amelioration. For example, focal loss [2] allows to effectively address the object-background inbalance in object detector training. Center loss [3], when used with the softmax loss to jointly supervise the learning of CNNs, can highly enhance the discriminative power of the deeply learned features for robust face recognition. Large-margin softmax loss [3] explicitly encourages intra-class compactness and inter-class separability between learned features in classification problem.

In our work we propose the use of Renyi entropy and divergence as a loss function for classification problem. While with one-hot encoding for target class Renyi entropy and divergence are simply equivalent to cross-entropy, for soft labels it's not the case anymore. More specifically, we consider data with non-binary labels obtained with mixup data augmentation method, proposed recently in the context of computer vision tasks.

To study the efficiency of the proposed loss function we conduct experiment on CIFAR-10 dataset [4]. We follow the experiment outline given in [1], where a number of different loss functions were compared in the context of image classification problem. We discover that Renyi cross-entropy and Renyi divergence with parameter $\alpha <$ < 1 lead to higher validation accuracy then cross-entropy.

The paper is organized as follows. We start with an overview of the related work. The following section describes the proposed method. Next, experiments and their results are presented. Finally, we conclude the paper the paper with summarizing statements.

Related work

Renyi divergence and entropy. An extensive work [5] is dedicated to the overview and extension of Renyi divergence properties. Those include convexity, continuity, limits of σ-algebras, a generalization of the Pythagorean inequality to general orders. A number of results on relationships between Renyi divergence, Kullback-Leibler divergence and Chernoff information in hypothesis testing have been obtained.

An interesting application to computer vision problems was studied in [6], where Bhattacharyya distance (*i.e.* exponent of Renyi divergence with $\alpha = \frac{1}{2}$) was used for vehicle tracking and counting. Spevifically, Bhattacharyya distance was applied to matching vehicles detected in different frames.

In a recent work of [7] a new loss function for generative adversarial networks (GANs) was proposed. In particular, the loss function for generator network was based on Renyi cross-entropy. The authors have shown the advantages of proposed method in comparison to baseline in terms of generated images quality and training stability.

A shifted version of Renyi entropy was proposed in [8]. In such formulation the definition becomes aligned with Hölder mean: r-th Hölder mean is the inverse of r-th Renyi entropy exponent. The properties of shifted Renyi entropy and its relations to different weighted power means are studied.

Paper [9] provides solution to the problem of Renyi divergence minimization. Properties of the corresponding functional are analyzed, and specific distribution cases are studies in detail. Additionally, a comprehensive overview of Renyi divergence applications was made.

Loss functions. A survey of loss functions used in machine learning is provided by [10]. In total, 31 loss functions are analyzed with respect to their purpose task and application scenario. This work presents both classical machine learning and deep learning standing points.

In paper [1] comparison of loss functions with application to image classification task is done. While cross-entropy remains the most common choice for this type of problem, authors demonstrate efficiency of loss functions that are usually reserved for regression tasks, in particular L_1 and L_2 losses. They continue with the analysis of loss function influence on the model training and final model characteristics, such as robustness to input and target noise.

Methodology

Renyi divergence and cross-entropy. Alfred Renyi defined divergence, or rather a spectrum of divergence measures that generalize the Kullback–Leibler divergence in [11]. The Renyi divergence of order α or the a distribution P from a distribution Q is defined as follows [11]:

$$D_\alpha(P||Q) = \frac{1}{\alpha - 1} \log \sum_{i=1}^{n} p_i^\alpha q_i^{1-\alpha} \tag{1}$$

where α is a positive parameter (not equal to 1), p_i is the value of the probability distribution P for $i = 1, 2, ..., n$, q_i is the value of the probability distribution Q for $i = 1, 2, ..., n$.

The Renyi entropy is defined [11] as

$$H_\alpha(P) = \frac{1}{1 - \alpha} \log \sum_{i=1}^{n} p_i^\alpha \tag{2}$$

Cross-entropy (or Kullback–Leibler cross-entropy) is a combination of the Shannon entropy of the distribution P and the Kullback–Leibler divergence between P and Q. Having that definition of cross-entropy, the Renyi cross-entropy can be determined by analogy. The Renyi entropy is reduced to the Shannon entropy when the value of the parameter α goes to 1, just as the Renyi divergence are reduced to the Kullback–Leibler divergence when the value of the parameter α goes to 1.

That is, the Renyi cross-entropy can be defined as follows:

$$H_\alpha(P, Q) = H_\alpha(P) + D_\alpha(P, Q) \tag{3}$$

where $H_\alpha(P)$ is the Renyi entropy, $D_\alpha(P, Q)$ is the Renyi divergence.

Substituting in the above formula 3 the Renyi entropy and Renyi divergence 1, we have the following definition:

$$H_\alpha(P, Q) = \frac{1}{1 - \alpha} log \frac{\sum_{i=1}^{n} p_i^\alpha q_i^{1-\alpha}}{\sum_{i=1}^{n} p_i^\alpha} \tag{4}$$

where α is a positive parameter (not equal to 1), p_i is the value of the probability distribution P

for $i = 1, 2, ..., n$, q_i is the value of the probability distribution Q for $i = 1, 2, ..., n$.

We remark that originally the Renyi cross-entropy proposed by Alfred Renyi [11] has the following definition:

$$H_{\alpha}^{or}(P,Q) = \frac{1}{1-\alpha} log \sum_{i=1}^{n} p_i q_i^{\alpha-1}. \qquad (5)$$

In the experiments, both definitions were tested, with the emphasize on the first one.

Mixup. It should be noted that Renyi divergence and cross-entropy are reduced to Shannon cross-entropy if the distribution P is concentrated at one point, *i.e.* has the distribution vector of the form $(0, \dots, 0, 1, 0, \dots 0)$. However, this is the common setting of classification task, where the true class of an instance is usually one-hot encoded.

To escape this scenario, we use mixup technique [12], where linear combinations of images and corresponding classes are used in training. When considering "mixed" images during the training of a neural network, the corresponding "true" probability distribution P is be concentrated at one point anymore (for example, for 10 classes problem it may look like $(0, 0, 0.7, 0.3, 0, 0, 0, 0, 0, 0)$). Due to this, the Renyi divergence and cross-entropy are not reduced to cross-entropy.

As a general comment, the fact that Renyi cross-entropy and Renyi divergence are reduced to cross-entropy at a certain value of the parameter α allows to assume that the use of Renyi cross-entropy and Renyi divergence as a loss function for the image classification problem will give no worse results than cross-entropy (which is a most common loss function choice for this type of problem problems) and may improve the results by investigating loss functions with different values of the parameter α.

Convolutional Neural Network. Convolutional Neural Networks (CNN) is a special type of neural network for processing grid-type data, such as images (2D grid), videos (3D grid) or time series (1D grid). The basic architectural ideas of a CNN [13] consist of the local receptive fields via the convolution operation and the spatial sub-sampling via the pooling operation. The convolution operation can be formally written as:

$$f_{x,y,k}^{C,l} = \mathbf{w}_k^{l\,T} f_{x,y}^{Op,l-1} + b_k^l \qquad (6)$$

where $\mathbf{w}_k^l$ and b_k^l are the weights and bias of the k^{th} feature map, $f^{Op,l-1}$ and $f_{x,y,k}^{C,l}$ are the input and output feature maps, l denotes the layer and (x, y) is the spatial image coordinate. The superscript C denotes convolution and Op represents various operations, *e.g.*, input (when $l = 1$), convolution, pooling, activation, etc.

Pooling applies local operations, *e.g.*, computing the maximum within a local neighborhood has the following form:

$$f_{x,y,k}^{P_{max},l} = max_{(m,n)\in\mathcal{N}_{x,y}}(f_{m,n,k}^{Op,l-1}) \qquad (7)$$

where $\mathcal{N}_{x,y}$ denotes the local spatial neighborhood and P_{max} denotes the max pooling. Often a spatial resolution reduction is applied after the max-pooling operation.

Besides the two above-mentioned operations, there are several strategies applied within the CNN models, such as non-linear activation (*e.g.*, the Rectified Linear Unit (ReLU) [14]), dropout [15] and batch normalization [16]. A Fully Connected (FC) layer, can be added at the end of the concatenated layers. It takes all nodes (neurons) from the feature maps of the previous layer as input and connects it to every nodes (neurons) of the output feature map.

On the last dense layer of the CNN model (referred to as the prediction layer), it is common to apply the Softmax activation function defined as follows:

$$\text{Softmax} = \frac{\exp(z_j)}{\sum_{k=1}^{K} \exp(z_k)} \qquad (8)$$

where K denotes the number of training samples.

Experiments

Dataset. The experiments were performed for the CIFAR-10 dataset [17], a popular dataset that is widely used in machine learning tasks. It consists of 60000 32x32 color images, the images are divided into 10 different classes - "airplane", "automobile", "bird", "cat", "deer", "dog", "frog", "horse", "ship", "truck". Each class contains 6,000 images, and the entire dataset is divided into 50,000 training images and 10,000 test images. Image classes are completely independent. There are no overlap between "automobile" and "truck". "Automobile" includes sedans, SUVs, etc. "Truck" includes only large trucks. Neither the "automobile" class nor the "truck" class includes pickups.

CNN architecture. For the CNN architecture we follow [1], where different loss functions were examined in application to the same image classification problem (CIFAR-10). This gives us an opportunity to compare the results of their work with our experiments.

Namely, the CNN used in experiments consists of three convolutional layers, each of a size of 5x5 and 64 filters, with ReLU activation function, batch-normalization and pooling operations between them. After the first layer max pooling was used, and average pooling after the next two, all with kernel 3x3, stride 2. The convolutional block is followed by a single fully connected layer with 128 neurons, ReLU activation, and the last softmax layer with 10 neurons. Both fully connected and softmax layer are preceded by dropout layers with dropout probability 0.125.

Training settings. The training lasted for 100 epochs. As an optimizer stochastic gradient descent (SGD) was used. Learning rate was decreasing from 0.01 to 0.001 with cosine learning rate scheduler [18]. For data augmentation, we apply width and height shift of maximum 5 pixels, horizontal flip, and random channel shift of range 10%. Finally, as mentioned before, we apply mixup [12], which results in convex combinations of images their labels. This serves both as regularization technique, and allows us to use Renyi cross-entropy and divergence as a loss function due to the presence of soft labels.

Loss functions and parameters. As loss functions for the experiments were used Renyi divergence (eq. 1) and Renyi cross-entropy (eq. 4) with different parameters α. The following α values were considered: 0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 1, 1.5, 2, 3, 5, 10. We remind that Renyi cross-entropy and Renyi divergence coincides with Shannon cross-entropy when the parameter $\alpha = 1$.

Results and discussions

Figure 1 shows the graphs of learning evaluation of neural network models with different alpha parameters based on training data - accuracy and losses. Figure 2 shows the same indicators but on the test data. Only several values of α are presented in order to make graphs more readable.

Evaluation of the obtained models. Overall, the loss and accuracy graphs look good, because there are no signs of overfitting or underfitting (Figure 1, Figure 2). There is a gap between the training and test graph lines. Test graphs have higher accuracy and lower losses compared to the training graphs. This can be explained by use of regularizers, such as dropout and mixup.

The set of alpha parameters can be roughly generalized by dividing into two parts: when alpha is less than one and when alpha is greater than one.

The results of the experiment show that the accuracy at smaller values of the alpha is slightly higher than the accuracy at higher alpha values, this can also be seen in Figure 1 and Figure 2. Moreover, the best result can be obtained with the value $\alpha = 0.3$, when the the accuracy reached almost 86% on the test dataset. However, it should be noted that there is no strong gap between the values of $\alpha < 1$, the accuracy of the model ranges from 1-2%. The model that used Renyi divergence with the value $\alpha = 10$ proved to be the worst. The trend of accuracy with increasing alpha value is declining. When the $\alpha = 10$, the accuracy reached about 81%, that is 5% less than the accuracy where $\alpha = 0.3$.

Figure 1. A line plot of model accuracy and loss on the training data with different values of α for Renyi divergence.

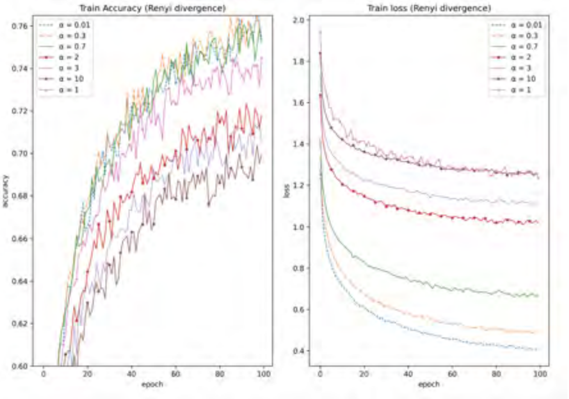
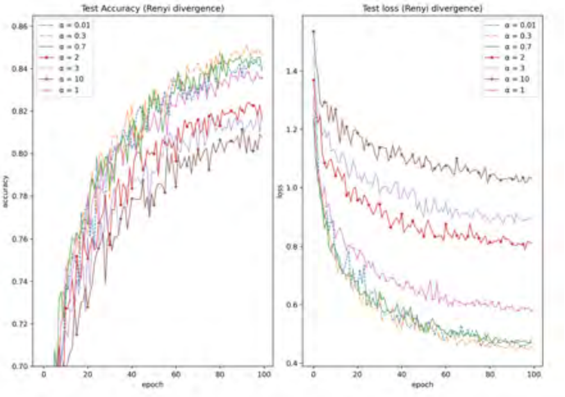
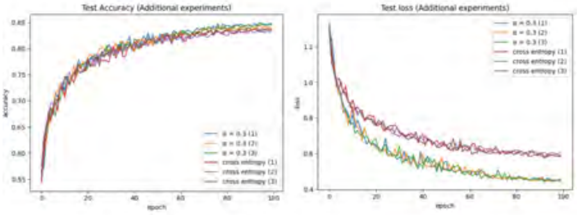


Figure 2. A line plot of model accuracy and loss on the test data with different values of α for Renyi divergence.



In addition, three more experiments were performed for the $\alpha = 0.3$ and for plain cross-entropy to exclude the chance of randomness in accuracy difference. Figure shows graphs of average accuracy and loss for test data.

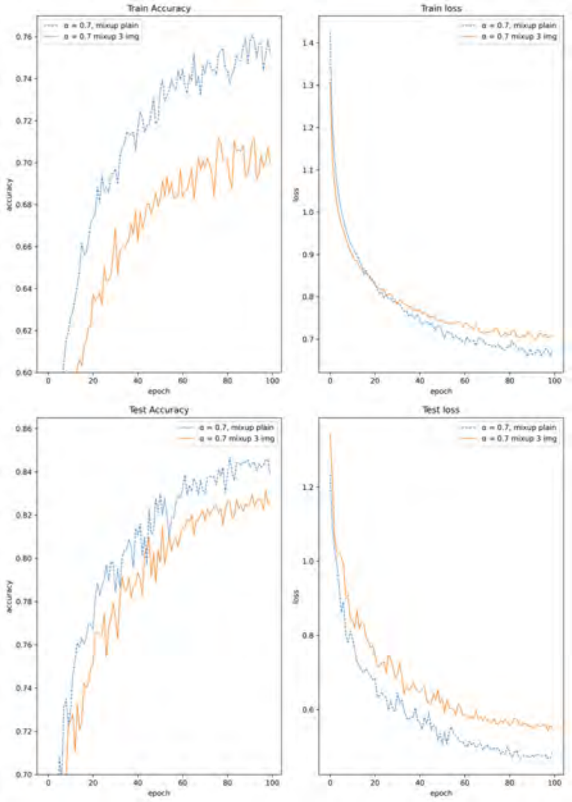
Figure 3. The results of the repeated experiments with cross-entropy and Renyi divergence with parameter $\alpha = 0.3$



As a result of additional experiments, the average accuracy of models with the $\alpha = 0.3$ was 84.1%. In contrast, the average accuracy of models with cross-entropy was 82.4%. The repeated experiments have shown consistent improvement over baseline with an average margin of + 1.7%, which likely eliminates the influence of chance on the results of the tests.

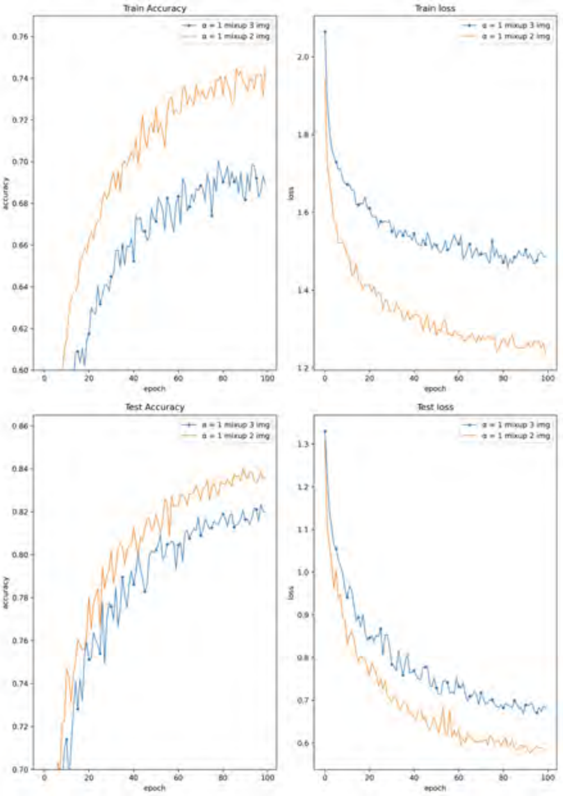
Comparison of experimental results with the baseline. The results can also be compared with the results reported in [1], where the authors considered different loss functions in similar experiment setting.

Figure 4. Comparative graphs with three images mixup and two images mixup for the value of the Renyi divergence parameter $\alpha = 0.7$.



The authors took 100 thousands iterations of 100 images, which corresponds to 200 epochs of the CIFAR-10 dataset with 50 thousand images. For our experiments 100 epochs were taken, due to limited computational resources. However, this does not prevent from comparing the results, although if another 100 epochs had been held, the results would have been somewhat better. The authors reported that the best results were achieved with L2 and higher-order hinge losses, accuracy reached about 80%, slightly lower with Cauchy-Schwartz divergence, and even lower with log loss (cross-entropy). In particular, for the latter the accuracy on the test dataset reached 78%, which is 6.1% lower than the results obtained during our experiments. One of plausible reasons of accuracy difference on the test set for cross-entropy with respect to the results reported in baseline source is the usage of mixup regularization method.

Figure 5. Comparative graphs with three images mixup and two images mixup for the cross-entropy.



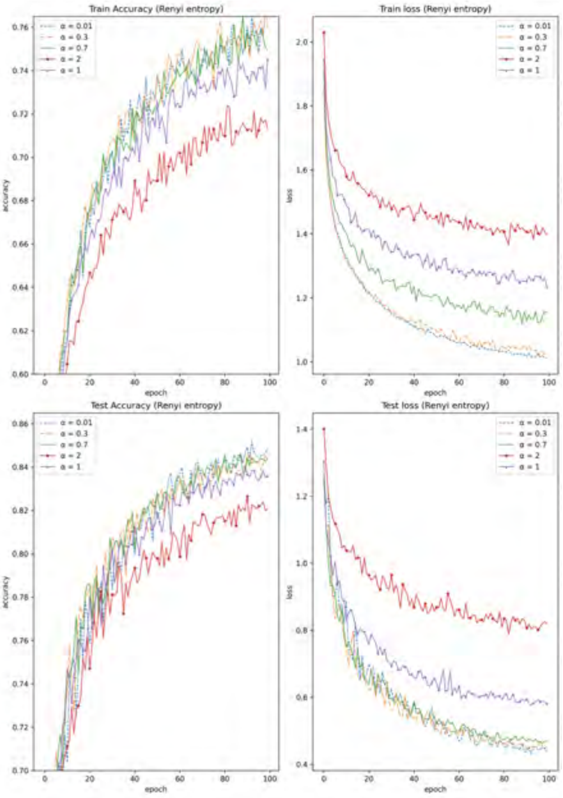
Experiments with mixup using three images. Using Renyi divergence as loss function with a mixup as a prerequisite showed quite confident improvements of the model accuracy. For the mixup, we considered convex combinations of two images being “mixed”. In terms of divergence, we moved from the “true” distribution concentrated in one point to the distribution concentrated in two points. It is logical to ask a question whether using more images for the mixup is going to change the model performance. For this purpose, we conducted an experiment with three images in mixup.

Two models were trained with Renyi divergence with parameter $\alpha = 0.7$, one with mixup of two images (plain mixup) and one with the mixup of three images. Comparative graphs can be seen in Figure 4 and Figure 5, respectively.

For mixup with the three images, the results were slightly worse than simple mixup. This may be since the new sample turned out to be “noisier” for the model, and therefore it was harder to learn and had greater losses.

For cross-entropy, the result of a mixup with three images can be interpreted in the same way as for the Renyi divergence. The results were slightly worse than the results with plain mixup.

Figure 6. Training and test accuracy and loss values corresponding to the model training with Renyi cross-entropy for different values of α.



Experiments with Renyi cross-entropy. Experiments were also performed for Renyi cross-entropy with the following values of the alpha: 0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 1 (cross-entropy), 2. For experiments, was used the definition of Renyi cross-entropy made as extension of cross-entropy definition (eq. (4)). Original definition of Renyi cross-entropy (eq. (5)) was also tested, however, no improvement was achieved.

Figure 6 shows graphs of accuracy and loss during training on a training dataset and a testing dataset.

The experiment showed that for the model with Renyi entropy the results are almost indistinguishable from the result with Renyi divergence. Similarly, smaller values of the alpha parameter lead to a more accurate model of the neural network, larger alphas have yield less accuracy.

The accuracy of models trained with $\alpha < 1$ is almost identical, but it is greater than the accuracy of the model trained with cross-entropy. Compared with the literature [1], we also observe superiority of obtained results with values of $\alpha < 1$ over other loss functions.

Conclusions

In this work we have proposed to use Renyi divergence and Renyi cross-entropy as a loss function for classification task with soft labels that generalizes cross-entropy. For experimental part, an image classification task (CIFAR-10) was considered, with mixup regularization as a source of soft labels.

Experiments have demonstrated that the proposed method of solving the image classification problem gives improvement with respect to the baseline and competitive loss functions [1]. The results showed that models trained with alpha values that are less than one have greater accuracy $w.r.t$ cross-entropy and lower loss, whereas larger values of alpha lead to less accurate model. The optimal value of parameter alpha remains an open question.

An additional experiment was also conducted to investigate how the performance of the model would change if mixup is done not with two but with three images. The result showed that mixing up three images negatively affects the training of the model with respect to the option with two images, which is aligned with the observations of the original mixup paper [12] for cross-entropy.

The choice of the loss function in the image classification problem has a crucial influence on the model accuracy and obtained experimental results attest to this statement. There are more suitable loss functions than the conventional ones, which, like the proposed method, may improve the learning of neural network models, thus this topic remains relevant and open for further research.

References

1. K. Janocha and W. M. Czarnecki, *On loss functions for deep neural networks in classification* (2017). Retrieved from arXiv preprint arXiv:1702.05659.
2. Tsung-Yi Lin, Priya Goyal, Ross Girshick et al. Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision (2017), pp. 2980–2988.
3. Yandong Wen, Kaipeng Zhang, Zhifeng Li and Yu Qiao, A discriminative feature learning approach for deep face recognition, in: European conference on computer vision / Springer (2016), pp. 499–515.
4. A. Krizhevsky, G. Hinton, A. Alex Krizhevsky et al., Learning multiple layers of features from tiny images (2009).
5. T. Van Erven and P. Harremos, "Rényi divergence and Kullback-Leibler divergence", IEEE Transactions on Information Theory. **60** (7), 3797–3820 (2014).
6. J. B. Baskoro, A. Wibisono and W. Jatmiko, *Bhattacharyya distance-based tracking: A vehicle counting application*, in: International Conference on Advanced Computer Science and Information Systems (ICACSIS) / IEEE (2017), pp. 439–444.
7. H. Bhatia, W. Paul, F. Alajaji et al., *Rényi generative adversarial networks* (2020). Retrieved from arXiv preprint arXiv:2006.02479.
8. F. J. Valverde-Albacete and Carmen Peláez-Moreno, "The case for shifting the Rényi entropy", Entropy. **21** (1), 46 (2019).
9. J.-F. Bercher, "On some entropy functionals derived from Rényi information divergence", Information Sciences. **178** (12), 2489–2506 (2008).
10. Qi Wang, Yue Ma, Kun Zhao, Yingjie Tian, A comprehensive survey of loss functions in machine learning, in: *Annals of Data Science* (2020), pp. 1–26.
11. A. Rényi et al. On measures of entropy and information, in: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics. The Regents of the University of California* (1961).
12. Hongyi Zhang, Moustapha Cisse, Yann N Dauphin and David Lopez-Paz, *mixup: Beyond empirical risk minimization* (2017). Retrieved from arXiv preprint arXiv:1710.09412.
13. Yann Lecun, Leon Bottou, Y. Bengio and Patrick Haffner, "Gradient-based learning applied to document recognition", Proceedings of the IEEE. **86** (11), 2278–2324 (1998).
14. Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification", IEEE International Conference on Computer Vision (ICCV 2015). **1502** (2015).
15. Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky et al., "Dropout: A simple way to prevent neural networks from overfitting", *Journal of Machine Learning Research.* **15**, 1929–1958 (2014).
16. Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift", Proceedings of the 32Nd International Conference on International Conference on Machine Learning. **37**, 448–456 (2015). Retrieved from http://dl.acm.org/citation.cfm?id=3045118.3045167.
17. Alex Krizhevsky, Vinod Nair and Geoffrey Hinton, *Cifar-10 (canadian institute for advanced research).* Retrieved from http://www.cs.toronto.edu/kriz/cifar.html.
18. Ilya Loshchilov, Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts (2016). Retrieved from arXiv preprint arXiv:1608.03983.

Андреєва В. П., Швай Н. О.

УЗАГАЛЬНЕННЯ ПЕРЕХРЕСНОЇ ЕНТРОПІЇ ЯК ФУНКЦІЇ ВТРАТ У ЗАДАЧАХ КЛАСИФІКАЦІЇ ЗОБРАЖЕНЬ

Задача класифікації є однією з найпоширеніших задач машинного навчання. Ця задача навчання з вчителем полягає у зіставленні кожному вхідному елементу однієї з скінченної кількості дискретних категорій.

Задача класифікації виникає природним чином у численних застосуваннях, таких як обробка медичних зображень, розпізнавання мовлення, системи технічного обслуговування, виявлення аварійних ситуацій, автономне водіння тощо. За останнє десятиліття методи глибокого навчання виявились надзвичайно ефективними для багатьох задач машинного навчання, зокрема класифікації. У той час як архітектура нейронної мережі може багато в чому залежати від типу даних та обмежень, що породжуються природою задачі (наприклад, застосування моделі у реальному часі), процес її навчання (тобто пошук параметрів моделі) майже завжди представляється як оптимізація функції втрат.

У задачах класифікації з багатьма класами у ролі функції втрат часто виступає перехресна ентропія, оскільки вона дає змогу досягти високої точності.

У цій роботі ми пропонуємо використовувати узагальнену версію цієї функції втрат, а саме розходження та ентропію Реньї. Зазначимо, що у випадку бінарних міток таке узагальнення зводиться до перехресної ентропії, тому нас буде цікавити саме контекст м'яких міток. Більш конкретно, ми розглядаємо проблему класифікації зображень, що розв'язується із застосуванням згорткових нейронних мереж та mixup регуляризації. Остання полягає у розширенні тренувального набору даних шляхом опуклих комбінацій пар елементів та відповідних міток. Відповідно, отримані мітки не є бінарними (що відповідає строгій належності до одного класу), а мають вигляд вектора ймовірностей. За таких умов перехресна ентропія та дивергенція і ентропія Реньї відрізняються, і їх можна порівняти між собою.

Для вимірювання ефективності запропонованої функції втрат ми розглядаємо проблему класифікації зображень на наборі даних CIFAR-10. Цей набір складається з 60 000 зображень, що належать до 10 класів, де зображення є кольоровими та мають розмір 32×32. Навчальний набір складається з 50 000 зображень, а тестовий набір містить 10 000 зображень.

Архітектуру згорткової нейронної мережі було обрано відповідно до [1], де була розглянута та сама задача класифікації з метою порівняння функцій втрат, з метою отримання порівнянних результатів.

Експерименти демонструють перевагу запропонованого методу над перехресною ентропією для значення параметра функції втрат $\alpha < 1$. Для значення параметра $\alpha > 1$ запропонований метод показує гірші результати, ніж функція перехресної ентропії. Нарешті, значення параметра $\alpha = 1$ відповідає перехресній ентропії.

Ключові слова: функція втрат, задача класифікації зображень, ентропія Реньї, розходження Реньї.

Матеріал надійшов 25.06.2020