International Journal of Computer Applications (0975 - 8887)
Volume 186 - No.9, February 2024

Semi-supervised Learning for Image Quality Assessment Problem

Tuan Linh Dang ✉
Hanoi University of Science and Technology
No. 1, Dai Co Viet Road
Hanoi 100000, Vietnam

Thuy Ha Hoang
Hanoi University of Science and Technology
No. 1, Dai Co Viet Road
Hanoi 100000, Vietnam

Minh Hoang Cu
Hanoi University of Science and Technology
No. 1, Dai Co Viet Road
Hanoi 100000, Vietnam

Duc Quang Nguyen
Hanoi University of Science and Technology
No. 1, Dai Co Viet Road
Hanoi 100000, Vietnam

Huu Phuc Hoang
Hanoi University of Science and Technology
No. 1, Dai Co Viet Road
Hanoi 100000, Vietnam

ABSTRACT

We live in the 21st century, a period of digital data explosion. Images are one example. Millions of photos are created yearly, so how can we evaluate their quality? In this article, we will introduce SSL algorithms to solve the problem of image quality assessment. We combined the KONIQ-10K and KADIS-700K datasets to create a new dataset and fix the image quality issues in the old datasets. We conducted comprehensive testing on the Vision Transformer in combination with 5 SSL algorithms, and the results we obtained were exceptional. Compared to ViT, ViT combined with the CRMatch algorithm gave outstanding results, with MAE reduced from 0.53 to 0.40.

General Terms

Computer Science, Computer Vision, Machine Learning

Keywords

Semi-supervised learning, image quality assessment

1. INTRODUCTION

We are living in a digital era, a period of data explosion. Every hour, thousands or millions of photos are posted on social networks. Besides, millions and billions of photo interactions exist on social networking platforms. It can be seen that social networks are a profitable environment for business communication. A quality photo and an impressive album can attract customers and increase sales. Therefore, evaluating image quality is essential and profoundly affects how users are approached. So, how can we evaluate image quality most accurately in this data explosion period? This has been a long-standing problem in the information technology industry: image quality assessment (IQA). IQA has sparked much research in supervised learning. However, with the growing data landscape, we cannot label it manually, and supervised learning becomes very difficult. Therefore, this article will present a new approach to the IQA problem: semi-supervised learning.

Semi-supervised learning combines supervised and unsupervised approaches. This machine method uses limited annotated samples and many unlabeled data for training. The objective of semi-supervised learning is akin to that of supervised learning. It aims to predict the outcome based on input.

The main contributions of our manuscript can be summarized as follows:

—Proposing semi-supervised learning to improve Image Quality Assessment.

—Conducting experiments on various datasets to identify the most effective SSL algorithm.

The structure of this paper is organized as follows. The literature review is shown in Section 2. Section 3 outlines the architecture we are proposing. The findings from the experiments are presented in Section 4. In conclusion, Section 5 wraps up this manuscript.

2. RELATED WORKS

2.1 IQA datasets

—LIVE IQA

LIVE IQA [1] is a dataset containing 29 high-resolution color

images from various sources, including the internet and photographic CD-ROMs. The dataset includes images of faces, people, animals, nature scenes, artificial objects, and images without specific objects. It includes 779 synthetic images created by applying diferent types of single distortions such as JPEG and JPEG2000 compressions, white noise, Gaussian blur, and bit errors in JPEG2000 bit stream. The dataset is manually annotated, but its size and content representation are limited. Its small number of images may limit its ability to provide a comprehensive assessment of image quality across diverse scenarios, and it may not catch the various distortions commonly found in real-world images.

—**TID 2008**
TID-2008 [2] dataset, similar to the LIVE IQA dataset, uses single distortions to create synthetic images from 25 reference images. These distortions include Gaussian blur, mean shift, and contrast change. The dataset consists of 1700 synthetic images. However, it primarily applies single distortions to reference images, neglecting complex mixtures of distortions found in real-world images. This may limit its ability to cover the diverse range of distortions in practical scenarios, potentially limiting its suitability for training and evaluating large-scale deep learning models.

—**TID 2013**
TID-2013 [3]is an image quality assessment dataset with 25 reference images and 3000 distorted images, similar to TID-2008. Despite its larger number, it may not fully cover the diverse array of distortions in practical scenarios.

—**KONIQ-10K**
KONIQ-10K [4], one of the most enormous IQA datasets by our understanding, consists of 10,073 samples with the scores of quality. This is the first database in the real world that focuses on ecological validity. It considers the authenticity of distortions, the variety of content, and quality-related indicators. Using crowdsourcing, this dataset gathered 1.2 million dependable quality ratings from 1,459 users, leading to the development of more universal IQA models.

—**KADID-10K & KADIS-700K**
The KADID-10K dataset contains 81 original images, each altered by 25 distortions at five different levels [5]. The KADIS-700K has a collection of 140,000 images. Each image has five degraded versions that were randomly selected.

2.2 Models

DBCNN [13] The DBCNN architecture is a deep bilinear model that can handle synthetic and authentic distortions in blind image quality assessment (BIQA). The model has two streams of convolutional neural networks (CNNs). Each stream addresses a different distortion scenario. The CNNs are pre-trained on different tasks: one for distortion type and level classification and the other for image classification. The characteristics from both CNNs are combined using bilinear pooling to create a single representation for predicting the overall quality. A variant of stochastic gradient descent fine-tunes the entire network on target databases.
The DBCNN results demonstrate superior results on different IQA databases. It also shows how it can be employed in the Waterloo Exploration Database on a large scale.

2.2.0.1 ***HyperIQA***. [12] A novel method for blindly assessing image quality in the wild, which means it can handle images with various contents and distortions without any prior knowledge. HyperIQA comprises three phases: content comprehension, perception rule acquisition, and quality prediction. Semantic features are extracted from the input image during the content comprehension stage, utilizing a pre-trained ResNet-50 model. The perception rule learning stage uses a self-adaptive hyper-network to generate the weights of a quality prediction network based on the semantic features. The quality prediction network then outputs a quality score for the input image. HyperIQA is designed to adapt to different image contents and distortions automatically, and thus achieve better performance on authentic image databases than existing methods. HyperIQA also achieves competitive results on synthetic image databases, although it is not explicitly trained for them. The architecture of HyperIQA is shown in the following figure.
HyperIQA outperforms the other methods in challenging authentic image databases such as KonIQ-10k and LIVE Challenge and ranks among the top approach on synthetic image databases such as CSIQ and TID2013. This demonstrates the effectiveness and robustness of HyperIQA for blindly assessing image quality in the wild.

2.2.0.2 ***Re-IQA***. [15] A machine learning method for assessing image quality without using reference images or types of distortions. Re-IQA consists of two stages: feature learning and quality prediction. Two encoders are trained using expert methods during the feature learning stage. The encoders learn high-level content and low-level quality features from unlabeled images. The high-level content features are extracted from a pre-trained ResNet-50 model, and the low-level quality features are learned by a contrastive learning framework called MoCo v2. The quality prediction stage uses a linear regression model to map the features to the ground truth quality scores. Re-IQA can generate quality features that are complementary to content features and thus achieve high performance on various IQA databases, both synthetic and authentic.
Re-IQA performs best on synthetic databases like LIVE, CSIQ, and TID2013, where the images have controlled distortions and similar content. Re-IQA also performs well on authentic databases, such as KonIQ-10k and LIVE Challenge, where the images have various contents and distortions.

2.2.0.3 ***ViT***. [14] The basic methodology of the Transformer model is separating a sequence of raw inputs and using a self-attention mechanism to calculate the contribution of each part to the output. They are well-designed to apply in NLP tasks when they split each word in a sentence. But it will be different when we talk about image processing. We just can't calculate attention weights for every single pixel in the image. It's impractical in the case of datasets with high resolution. Instead, ViT divides an image into smaller patches with the size P (16-32). Once the image is partitioned into patches, the ViT model linearly embeds each patch into a fixed-dimensional vector space. These embeddings serve as the input tokens for the Transformer encoder. The encoder has multiple layers. Each layer has a self-attention mechanism and a neural network for processing. Self-attention helps the model understand the importance of different patches in an image to make predictions and capture long-range connections within the image.

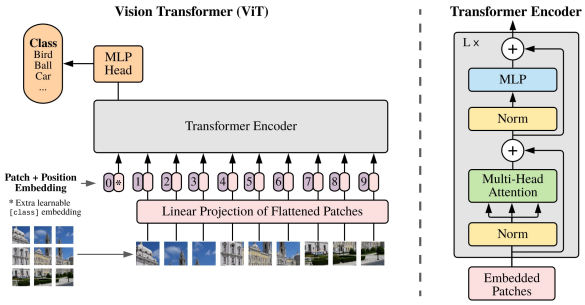

Fig. 1. ViT architecture from [14]

2.3 Semi-supervised learning

2.3.0.1 ***Fixmatch***. The FixMatch [6] approach generates synthetic labels, known as pseudo-labels by using a weakly augmented unlabeled image where augmentation like flip-and-shift is applied.These pseudo-labels are then used as targets when training the model with heavily augmented versions of the same images. For unlabeled data, we augment 2 twice: weak augment and strong augment. The model receives a weakly augmented image and generates predictions. After we use a predetermined threshold, the resulting prediction is transformed into a one-hot pseudo-label. With strong augmented image, the model calculated to predictions. And the supervised loss is calculated between predictions of strong image and pseudo-label. The supervised loss is designed to facilitate accurate prediction of labeled data, where the model aims to approximate the true labels closely. In Fixmatch, it is to optimize both supervised loss and unsupervised loss.

2.3.0.2 ***AdaMatch***. AdaMatch [7] extends FixMatch by by addressing the discrepancy in data distributions between the labeled and unlabeled domains present in the batch-norm statistics, adjusting the pseudo-label confidence threshold on-the-fly, and using a modified version of distribution alignment.

2.3.0.3 ***CoMatch***. CoMatch is a co-training framework that involves two representations: a class probability from the classification head and a low-dimensional embedding from the projection head [8]. These representations work together and improve in a co-training framework. The classification head improves the robustness of the pseudo-labels used during training using memory-smoothed pseudo-labels refined with information from neighboring samples. The projection head uses contrastive learning on a pseudo-label graph, encouraging samples with similar pseudo-labels to have embeddings close together. CoMatch is the first approach to introduce contrastive learning into SSL, and it is also used on graph-based feature representations.

2.3.0.4 ***FreeMatch***. Regarding unlabeled data, the FreeMatch algorithm proposed in the study by [9] has been developed to dynamically adjust classification thresholds based on the learning progress of individual classes. This algorithm utilizes the self-adaptive thresholding (SAT) method, employing the exponential moving average (EMA) of unlabeled data confidence scores to determine both global (dataset-specific) and local thresholds (class-specific).

To enhance its efficacy in minimally supervised environments, a recommendation is made to implement a class fairness objective, encouraging the model to generate balanced predictions across all categories. The primary goal of the FreeMatch training approach is to optimize mutual information between the model's input and output, resulting in the generation of diverse and confident predictions on unlabeled data.
However, it is important to note that the generation of pseudo-labels in this approach remains entirely unsupervised, leading to the complete disregard of labeled information. Moreover, the reliability of pseudo-labels diminishes, particularly when the amount of labeled data is limited. .

2.3.0.5 ***SimMatch***. SimMatch [11] is method that aims to address the challenge of unreliable pseudo-labels in scenarios with insufficient labeled data. Pseudo-labels can lead to the "overconfidence" issue, where the model might learn from the inaccurate pseudo-labels and perform poorly. To achieve this, SimMatch works by aligning similarities at both the semantic and instance levels across different data augmentations at the same time. Specifically, the algorithm ensures that strongly augmented data maintains consistent semantic similarity (meaning label predictions align) with its weakly augmented counterpart. Moreover, it also promotes feature matching by aligning instance characteristics (similarities between individual instances) between strong and weak augmentations. Unlike previous approaches that solely rely on predictions from weakly augmented data for pseudo-labels, SimMatch introduces a unique interaction between semantic and instance pseudo-labels. It achieves this through a memory buffer that stores all labeled examples. By employing aggregating and unfolding techniques, these two types of similarities can be transformed into each other, leading to mutually enhanced accuracy and reliability of the generated matching targets.

2.3.0.6 ***CrMatch***. [10] CR-Match, a novel approach, integrates FeatDistLoss with other robust techniques, establishing a new benchmark across various settings in standard semi-supervised learning (SSL) evaluations. This includes prominent benchmarks such as CIFAR-10, CIFAR-100, SVHN, STL-10, and Mini-Imagenet. The method effectively enforces regularization on the feature representation distances derived from differently augmented images belonging to the same class.

3. PROPOSED ARCHITECTURE

3.1 Overview of the proposed architecture

An overview of our system is shown in Figure 2 with two main modules including the Supervised modules and Semi-Supervised modules. The Supervised module approach is effective to train model, particularly when a large number of labeled datasets are available. But in this problem, unlabeled data is too much. So the supervised modules is used for solve problem.

3.2 Dataset

The above datasets have many disadvantages compared to our problem. Previously released datasets have many limitations. LIVE IQA, TID-2008 and TID-2013 are all image sets consisting of only 20-30 original images, which are then 'noise' across multiple scenes to increase the size of the dataset. In this way, we can see

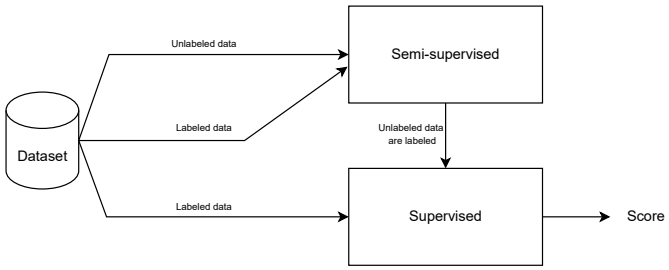

Fig. 2. Overview system

that this set of photos does not focus on the content of the photo to evaluate, but only through indicators such as high blur, etc. With the KONIQ-10K data set, the data has been evaluated. focuses more on the content of the image, but compared to the huge amount of data, about 10,000 images is too little. To solve the problem of having little data and needing more data to focus on the content of the image, we have created a new data set based on the combination of two data sets KONIQ-10K and KADIS-700K. Below is a diagram of the dataset:

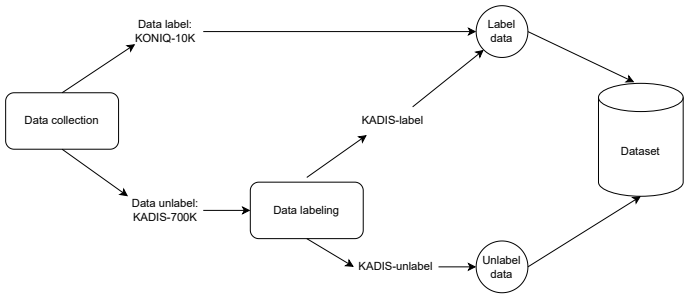

Fig. 3. Overview dataset

This KONIQ-10K dataset has 10,073 quality-scored images. Its score ranges from 0-100. We scale each image to a scale of 0-4 labels. In this dataset, KADIS-700K, we use 140,000 pristine images. Our approach labeled 159 more images with label 0 and 200 images with label 4 from 140.000 pristine images. Finally, we have 10432 labeled images compiled from KONIQ-10K, 359 newly labeled images, and 139641 unlabeled images from KADIS-700k. The new dataset contains 10432 labeled images and 139641 unlabeled images, called "**KOKA10K**".

3.3 Semi-supervised

The Semi-supervised learning will be employed for training dataset. This semi-supervised module uses backbone as the same in supervised module. Different from supervised, our approach use Semi-supervised learning algorithm to train data. Details for algorithms can be read at Section 2.3. After the training process, unlabeled data is inferenced by this semi-supervised module.

3.4 Supervised

After we have the inference of all unlabeled images, we combine it with the dataset of labeled images. At this time, we have a whole dataset enough for training with supervised model. We use ViT to extract features of images. With each feature, we use a simple full connected layer to get the logit regression ouput of it. By applying the cross-entropy loss function, the errors between the ground-truth labels and predicted ones is calculated and will be minimized by a particular optimizer.

Table 1. Train val test in KOKA10K.

Dataset	Label	Rate	Total images	
Test	0	2.19%	23	1047
	1	13.27%	139	
	2	49.18%	515	
	3	33.17%	347	
	4	2.19%	23	
Val	0	2.11%	22	1040
	1	13.17%	137	
	2	49.52%	515	
	3	33.19%	345	
	4	2.02%	21	
Train	0	2.16%	180	8345
	1	13.22%	1103	
	2	49.37%	4120	
	3	33.15%	2767	
	4	2.10%	175	

4. EXPERIMENTAL RESULTS

We used Pytorch framework to experiment. Based on our experiments, the Adam optimizer algorithm, cross-entropy loss which was our cost function for experiments with all datasets.

Table 2. Prediction results on Data10k

	Dataset 10k			
Model	**MAE**	**MAE of 3**	**MAE of 4**	**MAE of 3&4**
VIT	0.53	0.54	0.65	0.55
VIT + SimMatch	0.45	0.24	0.47	0.25
VIT + AdaMatch	0.47	0.23	0.65	0.25
VIT + CoMatch	0.46	0.22	0.3	0.23
VIT + FreeMatch	0.43	0.33	0.43	0.34
VIT + CRMatch	0.40	0.33	0.39	0.33

5. CONCLUSIONS

In this paper, we have introduced a new solution in the problem of image quality assessment: Semi-supervised learning. We also experimented this method on the general data set we generated, with results showing clear improvement with and without SSL. Using only ViT by supervised learning, the MAE is 0.53. However, when using ViT combined with semi-supervised learning, MAE decreased. Especially on the CRMatch algorithm, the MAE result is only 0.40. In conclusion, Semi-supervised learning is a completely feasible method for the Image Quality Assessment problem with large amounts of unlabeled data. More broadly, this method can be completely applied to other problems where the amount of unlabeled data is much larger than the amount of labeled data.
In our future, we can involve augmenting the labeled data through manual labeling, create additional semi-supervised learning algorithms. Moreover further enhancement of results can be achieved through extensive fine-tuning of hyperparameters

Acknowledgment

This research is funded by the Hanoi University of Science and Technology (HUST) under project number T2022-PC-052.

6. REFERENCES

[1] H.R.Sheikh, M. F. Sabir and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," IEEE Transactions on image processing, jourvol 15, number 11, pages 3440–3451, 2006

[2] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli and F. Battisti, "Tid2008-a database for evaluation of full-reference visual quality assessment metrics," Advances of modern radioelectronics, jourvol 10, number 4, pages 30–45, 2009.

[3] N. Ponomarenko, O. Ieremeiev, V. Lukin andothers, "Color image database tid2013: Peculiarities and preliminary results," inEuropean workshop on visual information processing (EUVIP) IEEE, 2013, pages 106–111

[4] V. Hosu, H. Lin, T. Sziranyi and D. Saupe, "Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment," IEEE Transactions on Image Processing, jourvol 29, pages 4041–4056, 2020.

[5] KADID-10k: A Large-scale Artificially Distorted IQA Database, Hanhe Lin, Vlad Hosu, and Dietmar Saupe

[6] K. Sohn, D. Berthelot, N. Carlini andothers, "Fixmatch: Simplifying semisupervised learning with consistency and confidence," Advances in neural information processing systems, jourvol 33, pages 596–608, 2020.

[7] Adamatch: A Unified approach to semi supervised learning and domain adaptation David Berthelot , Rebecca Roelofs , Kihyuk Sohn , Nicholas Carlini , Alex Kurakin

[8] CoMatch: Semi-supervised Learning with Contrastive Graph Regularization Junnan Li Caiming Xiong Steven C.H. Hoi

[9] FREEMATCH: SELF-ADAPTIVE THRESHOLDING FOR SEMI-SUPERVISED LEARNING Yidong Wang12, Hao Chen3 , Qiang Heng4, Wenxin Hou5, Yue Fan6, Zhen Wu7, Jindong Wang1, Marios Savvides3, Takahiro Shinozaki2, Bhiksha Raj3, Bernt Schiele6, Xing Xie1

[10] Revisiting Consistency Regularization for Semi-Supervised Learning Yue Fan Anna Kukleva Bernt Schiele

[11] Zheng, Mingkai, et al. "Simmatch: Semi-supervised learning with similarity matching." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[12] Su, Shaolin, et al. "Blindly assess image quality in the wild guided by a self-adaptive hyper network." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[13] Zhang, Weixia, et al. "Blind image quality assessment using a deep bilinear convolutional neural network." IEEE Transactions on Circuits and Systems for Video Technology 30.1 (2018): 36-47.

[14] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

[15] Re-IQA: Unsupervised Learning for Image Quality Assessment in the Wild Avinab Saha Sandeep Mishra Alan C. Bovik Laboratory of Image and Video Engineering, The University of Texas at Austin