Exploring Uncertainty-aware Class-wise Thresholds for Recognition Model's Uncertainty Detection

1st Jihyun Hwang
School of Electronics and Telecommunications Research Institute
University of Science and Technology
Daejeon, Republic of Korea
aribae@etri.re.kr

2nd* Minsu Jang
Social Robotics Research Section
Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea
minsu@etri.re.kr

***Abstract*—This study addresses the current limitations in artificial intelligence models, specifically their ability to accurately predict uncertainty, a crucial aspect for improving trustworthiness and safety. We introduce a technique for estimating uncertainty in zero-shot image classification using the vision-language model, CLIP. The standard method for quantifying uncertainty in classification models is to measure the entropy from normalized output value, which is then compared to a predetermined threshold to ascertain the certainty of the model's output. We propose a method for adapting this approach for vision-language model-based zero-shot image classification to enhance uncertainty prediction through class-wise entropy thresholding. Experimental results reveal that class-wise entropy thresholding surpasses the previous approach. Additionally, it is demonstrated that our approach is successful for detecting out-of-distribution instances.**

***Index Terms*—classification, uncertainty, out-of-distribution, cosine-similarity**

I. Introduction

Existing artificial intelligence technologies do not have the ability to distinguish between what they know and what they do not know, so they do not have flexibility in various unlearned situations, and cannot grow on their own because they cannot judge the uncertainty of knowledge. If artificial intelligence is aware of the uncertainty of knowledge and can grow and learn knowledge by resolving uncertainties, it will be able to respond flexibly to situations that have not been learned. To this end, when artificial intelligence encounters Out-Of-Distribution(OOD)[7] information, there is a need for a method to determine its own inference uncertainty.

The vision-language model is developing into a key technology used in a variety of applications in the field of contemporary computer vision. Through the relationship between images and text, these models are capable of carrying out a variety of tasks like image classification, search, and creation. Particularly non-language models, which set themselves apart from conventional models and perform well, include Contrastive Language-Image Pretraining (CLIP). Firstly, the CLIP is a versatile model. Models for object or image recognition that are currently in use were created with a specific task in mind. On the other hand, by comprehending and articulating the relationship between images and text, CLIP can perform well in a variety of domains. It also has the integration of language and vision. By combining the learning of image and text data, this model can comprehend the semantic relationship between an image and a text. This enables the model to categorize or retrieve images using the image's textual descriptions. The model can recognize and categorize novel objects that are not understood by using textual information in an open-world environment where unknown classes or objects appear through the integration of vision and language [1]. As a result, we want to take advantage of CLIP's multi-purpose model and vision-language integration capabilities to deal with the zero-shot classification problem. Traditional approaches set thresholds to measure uncertainty using confidence values. However, such confidence-based approaches could differ from the characteristics of the model since CLIP measures similarity by taking advantage of the distinction between images and text. Therefore, by resetting thresholds in CLIP to take into account distance-based similarity measurements, we would like to provide a method for measuring uncertainty that is more appropriate for zero-shot classification problems.

II. Approach

A. Uncertainty-aware CLIP based zero-shot image classification

The interaction between images and text enhances classification performance, and the CLIP model excels in solving image classification problems. Studies have reported that CLIP achieves high performance in image classification tasks, even under zero-shot conditions. To carry out classification using CLIP, a set of textual prompts is needed, each corresponding to a specific class. To classify an input image img, the initial step involves calculating E_{img}, the image embedding for the input image, and $E_{text}(j)$, the textual embedding for each prompt j. Subsequently, the cosine similarity s_j between E_{img} and $E_{text}(j)$ for $j = 1, 2, \ldots, K$ is determined using equation 1 and is normalized by 2 to get the final similarity score sim_j. Finally, the image is classified by selecting the prompt that has the highest similarity and identifying the corresponding class.

$$s_j = \frac{E_{img} \cdot E_{text}(j)}{\|E_{img}\| \times \|E_{text}(j)\|} \tag{1}$$

$$sim_j = \frac{s_j + 1}{2} \quad for\ j = 1, 2, \ldots, K \tag{2}$$

In the context of zero-shot image classification, the uncertainty of the classification is estimated by calculating the entropy of the distribution of similarities between the input image and the prompts. The uncertainty is directly proportional to this entropy: as the entropy increases, so does the uncertainty in the classification. The entropy is computed using the equation 3, where img indicates a given image.

$$H(img) = -\sum_{j=1}^{K} sim_j \log sim_j \tag{3}$$

B. Class-wise entropy thresholds method

The confidence thresholding technique measures the uncertainty of a classification model by asserting that a prediction is uncertain if the entropy of the output predictions exceeds a predetermined entropy threshold [10]. Generally, a single entropy threshold is applied even in multi-class classification scenarios. However, this approach has a limitation: it fails to account for the variability in prediction probabilities across different classes.

To mitigate this limitation, we propose to establish distinct entropy threshold values, th_c, for each class. These thresholds are optimized according to the entropy distribution specific to each class. In zero-shot classification scenarios using CLIP, the uncertainty of a classification is evaluated by comparing $H(img)$ to th_c for a given class c, where sim_c is the highest similarity score. A classification is deemed uncertain if $H(img)$ surpasses th_c. This complete procedure is depicted in Figure 1.

Our approach not only provides the model with the flexibility to handle challenging classes, but it also enhances overall classification accuracy, outperforming the results obtained with a single universal threshold. Hence, in this paper, we introduce a method for assessing classification result uncertainty while also establishing independent entropy thresholds for each class.

C. Grid search

We applied the grid search method to determine class-wise entropy thresholds. The most optimal parameter combination is chosen by methodically examining all possible parameter combinations, which is one of the numerous classic methods for optimization. Grid search has the benefit of exploring every possibility by evaluating every combination within a specified range. This enables us to determine the ideal hyperparameter values for enhancing the model's functionality. We systematically assess each combination's impact on the model's performance while taking the interaction between various hyperparameters into account. The performance and generalizability of the model are improved by determining the most optimal combination while taking into consideration all other possible combinations. By utilizing the grid search approach, we have determined optimal threshold values for each class. In this process, we extracted and analyzed a list of entropies calculated for all samples within each class. Entropy list quantifies the level of uncertainty associated with each sample. Specifically, from this list, we identified the entropy value that minimizes both the count of correctly classified samples exhibiting uncertain characteristics and the count of incorrectly classified samples lacking uncertainty. This entropy value was then chosen as the threshold to assess the uncertainty of the recognition model's classification results for each class. This approach is illustrated in Algorithm 1.

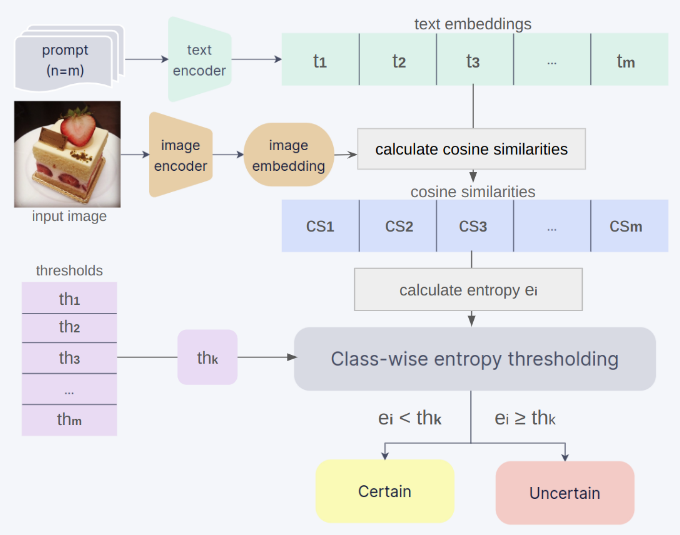
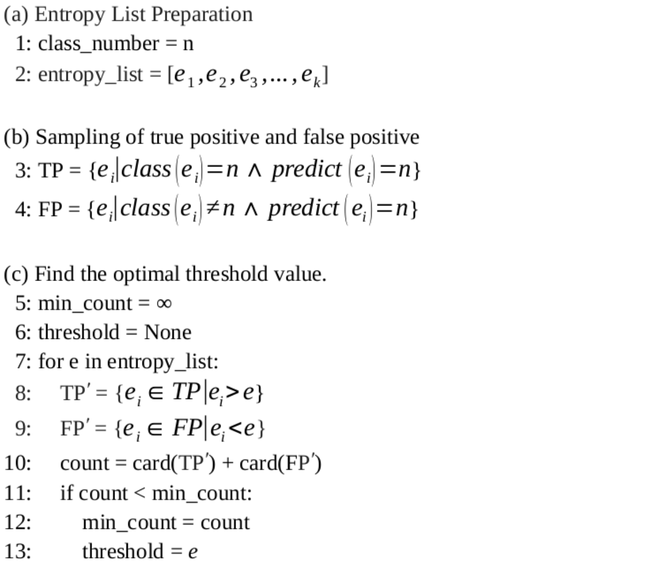


Fig. 1. Class-wise entropy thresholding method architecture. k is index of top-1 cosine similarity

Algorithm 1 Grid search method for fine class-wise entropy thresholds



```
(a) Entropy List Preparation
  class_number = n
  entropy_list = [$e_1, e_2, e_3, \dots, e_k$]

(b) Sampling of true positive and false positive
  TP = {$e_i | class(e_i) = n \wedge predict(e_i) = n$}
  FP = {$e_i | class(e_i) \neq n \wedge predict(e_i) = n$}

(c) Find the optimal threshold value.
  min_count = ∞
  threshold = None
  for e in entropy_list:
      TP′ = {$e_i \in TP | e_i > e$}
      FP′ = {$e_i \in FP | e_i < e$}
      count = card(TP′) + card(FP′)
      if count < min_count:
          min_count = count
          threshold = $e$
```

The results of the uncertainty judgment in III-B and III-C were analyzed using three thresholds. The three thresholds are **grid search class-wise thresholds** obtained through Algorithm 1 , their average value, **mean of class-wise thresholds single threshold**, and **grid search single threshold** obtained by applying Algorithm 1 to the prediction results of the CLIP model for the entire dataset. Regarding this, entropy values derived from train data were set as the entropy samples. This allowed for the measurement of the model's classification result's uncertainty and the use of the data to determine the optimal threshold value. The determined thresholds were evaluated using the test dataset. The determined thresholds were evaluated using the test dataset.

III. Experiments and Results

A. Datasets

In this experiment, experiments were conducted using four different datasets. Through this, we evaluated the performance of a recognition model and introduced class-wise entropy thresholds to determine uncertainty, and sought ways to further improve uncertainty detection performance by utilizing it. We used two datasets (CIFAR10[3] and Food101[4]) showed better performance of the CLIP model compared to the existing pre-trained ResNet50 model and the CLIP model as a result of the CLIP paper, and two standard Zero-Shot Classification(ZSC) datasets (Stanford Dogs[5] and Caltech-UCSD Birds-200-2011[6]).

TABLE I
The Statistics of The Two Datasets Used

	Dataset	
	CIFAR10	***Food101***
# of classes	10	101
# of images	6000	101000

The number of images and classes for each dataset is shown in Table I, and Table II shows the performance of the CLIP model for each dataset. The classification result was derived by applying the CLIP model to each dataset. When the dataset was split into train and test, the matching distinction was preserved, but when it wasn't, the data was split into train and test in an 8:2 ratio.

TABLE II
CLIP Model Accuracy

	Dataset	
	CIFAR10	***Food101***
# of images	6000	101000
# of correct predictions	7926	12434
# of incorrect predictions	2074	7766
Accuracy(%)	79.206	61.554

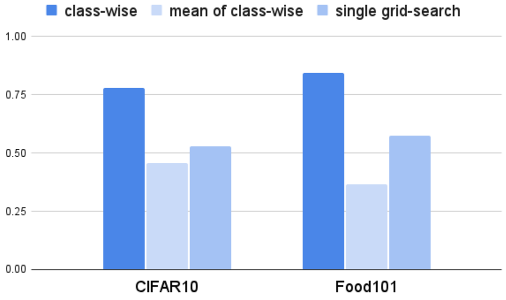
B. Class-wise entropy thresholding test

A comparative analysis was conducted to evaluate the efficacy of the three threshold values that were obtained. The aim was to evaluate the possibility for applying the uncertainty threshold to reveal instances of uncertainty within the subset of instances where the model failed to correctly classify. The purpose of this comparison was to demonstrate how effectively the newly introduced uncertainty threshold could detect ambiguous predictions (Table III).

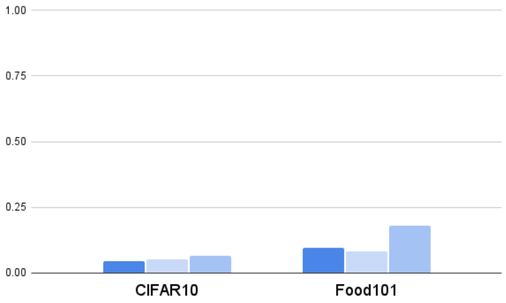
TABLE III
Uncertainty Prediction Performance

	Dataset	
	CIFAR10	***Food101***
class-wise thresholds	**0.779**	**0.845**
mean of class-wise thresholds single threshold	0.456	0.367
grid search single threshold	0.529	0.576

Fig 2 is a visualization of the results. As a result of the analysis, the proportion of the case where class-wise entropy thresholding was applied was higher than that of the case where the model was classified as incorrect (a graph of Fig 2).



(a) Uncertain imcorrect classification (↑)



(b) Uncertain correct classification (↓)

Fig. 2. Uncertainty detection performance

C. OOD dataset test

Further experimental analysis was performed to verify the effectiveness of threshold setting on the problem of not addressing classification on out-of-distribution(OOD) images not given at the prompt, which is a limitation of the CLIP model mentioned in this work. In-Distribution(ID) and OOD were excluded from the prompt by designating 20% of the same class as OOD in the train and test datasets, respectively, and then calculating the prediction results of the CLIP model to obtain thresholds and compare the results of applying them(Table IV).

The results of OOD (Fig 3) confirmed that class-wise entropy thresholds showed the highest performance. Based on

TABLE IV
UNCERTAINTY PREDICTION PERFORMANCE IN OOD DATASETS

	Dataset	
	CIFAR10	***Food101***
class-wise thresholds	**0.779**	**0.845**
mean of class-wise thresholds single threshold	0.456	0.367
grid search single threshold	0.529	0.576

TABLE V
UNCERTAINTY PREDICTION ACCURACY(F1-SCORE)

dataset	**Class-wise entropy thresholding test**		**OOD dataset test**	
	CLIP	***Our***	***CLIP***	***Our***
CIFAR10	0.796	**0.956**	0.671	**0.956**
Food101	0.607	**1**	0.516	**1**

these results, it is judged that the possibility of use can be increased by supplementing the uncertainty detection performance of recognition models through the class-wise entropy thresholding method.

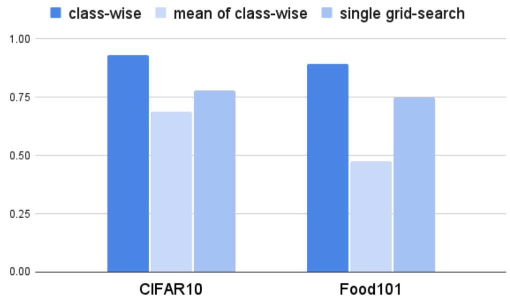


Fig. 3. Uncertainty detection performance in OOD datasets (↑)

Our class-wise entropy thresholding method we applied showed the effect of actually seeing the improvement of the accuracy of the model. Below are example images of determining uncertainty through the method (Fig 4), and Table V shows how the F1 score improved in each part. If the model's prediction results were deemed uncertain, the correct labeling was given, and the results showed an improvement in accuracy across all datasets.

IV. DISCUSSIONS

Our class-wise entropy thresholding method, the method used in this study, shows higher performance as a result of synthesizing the entire set of results. This leads to the conclusion that the application of threshold values for each class has a beneficial impact on the model's performance improvement. However, it becomes apparent that there is a significant bias in the basic characteristics of the images as one delves deeper into the details of ZSC datasets. The CLIP model's classification performance noticeably declines as a result of this bias. The expected improvement remained elusive despite efforts to reduce this bias by using a threshold, and the resulting uncertainty measurements failed to reach the desired level of significance. As a result, due to their potential to reduce the reliability of the model, it becomes essential to take a cautious and meticulous approach when navigating the landscape of these datasets. Thus, it is key to set out on a research adventure in search of methods to simultaneously improve the model's functionality and the dependability of the class-wise entropy thresholding approach. It is expected that these studies will not only contribute to a broader understanding of dataset utilization within machine learning models but will also provide invaluable guidelines to uphold and ensure the unwavering reliability of the resulting models by addressing and supplementing the prevailing bias within the ZSC dataset through focused future research endeavors.

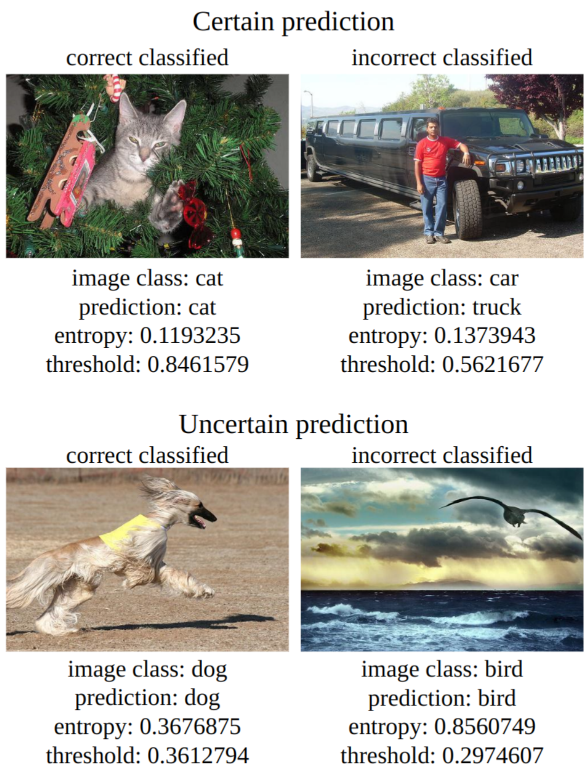


Fig. 4. Uncertainty detection example images

V. CONCLUSION

In this paper, we studied how to improve the model's uncertainty detection performance by applying class-wise entropy thresholds to classification problems using a recognition model. Experiments demonstrate that more robust classification performance can be achieved by introducing class-wise entropy thresholds to measure the uncertainty of the model's prediction results. Through this method, it was confirmed that improved classification performance can be obtained even in uncertain cases. Based on these results, it is expected that our method can contribute to increasing the likelihood of safe and reliable applications in real-world applications by helping to increase the reliability of artificial intelligence models. In addition, thanks to the nature of the CLIP model, which allows users to easily add new classes using prompts without additional learning, it seems that it can be effectively applied to active learning problems in the future. Through this, it seems that high performance can be achieved at a low labeling cost[8].

REFERENCES

[1] Radford, Alec, et al. ”Learning transferable visual models from natural language supervision.” International conference on machine learning. PMLR, 2021.

[2] Liashchynskyi, Petro, and Pavlo Liashchynskyi. ”Grid search, random search, genetic algorithm: a big comparison for NAS.” arXiv preprint arXiv:1912.06059 (2019).

[3] Krizhevsky, Alex, and Geoffrey Hinton. ”Learning multiple layers of features from tiny images.” (2009): 7.

[4] Bossard, Lukas, Matthieu Guillaumin, and Luc Van Gool. ”Food-101–mining discriminative components with random forests.” Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13. Springer International Publishing, 2014.

[5] Khosla, Aditya, et al. ”Novel dataset for fine-grained image categorization: Stanford dogs.” Proc. CVPR workshop on fine-grained visual categorization (FGVC). Vol. 2. No. 1. Citeseer, 2011.

[6] Wah, Catherine, et al. ”The caltech-ucsd birds-200-2011 dataset.” (2011).

[7] Hendrycks, Dan, and Kevin Gimpel. ”A baseline for detecting misclassified and out-of-distribution examples in neural networks.” arXiv preprint arXiv:1610.02136 (2016).

[8] Settles, Burr. ”Active learning literature survey.” (2009).

[9] Chicco, Davide, and Giuseppe Jurman. ”The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation.” BMC genomics 21.1 (2020): 1-13.

[10] Seung, H. Sebastian, Manfred Opper, and Haim Sompolinsky. ”Query by committee.” Proceedings of the fifth annual workshop on Computational learning theory. 1992.