Context-Aware Multi-Task learning

Huiyu Wang
Chair for Data Processing, Technical University of Munich
huiyu.wang@tum.de

***Abstract*—In autonomous driving tasks, we consider concept of context as weather, road types, period, etc. Due to different definitions of context in previous works, which led to divergence of research directions and confusion of readers, a unified and easy-to-understand definition of context was needed. In this paper, a more general definition of context in the field of autonomous driving is given. Compared with single-task classification, multi-task learning (MTL) uses correlation between context to better extract and classify different types of context. Model is trained on ONCE dataset and demonstrates competitive multi-classification ability and strong scalability.**

***Keywords*—Context, Classification, Multi-task learning.**

I. Introduction

In autonomous driving tasks, community expects to extract as much information as possible from sample images. Apart from detected subject objects, context has attracted attention as one of the factors in autonomous driving tasks. It plays a different role in diverse scenarios. Although accurately extracted context can greatly help model's understanding of scene, thereby improving performance of model, exact definition of context has not been universally recognized. Furthermore, many context-aware research directions are somewhat divergent. Therefore, definition of context is indicated in task of autonomous driving after combining many theoretical and practical fields of research: context is considered to be a scene or background other than detection subject, such as weather, time, location, illumination, which can greatly affect model's behavior and positioning of recognized subject. Since background of autonomous driving containa multiple contexts at the same time, multi-task classification is used to classify contexts. Considering correlation between different features, MTL [1] is adopted to train model. In this work, definition of context is shown. Model is trained to extract and classify context on ONCE dataset [2]. Factors are analyzed which affect context in combination with experiments. Main contributions are as follows:

Definitions of context and extraction method of context are compared in previous papers. Three different classification methods are introduced and appropriate loss function and reweight method are chosen in combination with ONCE dataset. Lastly experiments are designed and results are analysed.

II. Related Works

This work includes following aspects: definition of context and its extraction.

A. Context definition

Much prior works have a broader and general definition of Context. L. Wolf and S. Bilesch mention that any information that might be relevant to object detection, categorization and classification tasks, but not directly due to the physical appearance of the object, as perceived by the image acquisition system [3]. T. M. Strat tells us that any and all information that may influence the way a scene and the objects within it are perceived [4]. This work shows that context includes local pixel, 2D scene gist, 3D geometric, etc. [5]. C. Galleguillos and S. Belongie introduce that context is an appearance information, based on visual cues. It can successfully identify object classes up to a certain extent. Context information, based on the interaction among objects in the scene or on global scene statistics, can help successfully disambiguate appearance inputs in recognition tasks [6]. Recently context has been increasingly mentioned in the field of autonomous driving. Definition of context may vary based on problem of interest. Similar parts of the same autonomous driving image are considered to be their context [7], and training set samples with better diversity distribution are selected accordingly. Probability distribution of classes in the spatial neighborhood of the detection object is considered to be context [8]. It is believed that a piece of object and behavior has a context under a certain spatial and temporal relationships [9]. Compared with above definitions, our definition of context is more specific and easier to express in autonomous driving tasks. Difficulty of detecting context has also become one of the criteria for evaluating context.

B. Context detection

To detect context, people take different approaches: in this work [7], methods such as k-means++, core-set, and sparse modeling are used to describe similarity between images. A group of pictures with similarity is considered to have a similar context. S. Agarwal et al. tells us that probability distribution obtained by classifying different objects in the target neighborhood by classifier is used as context, and pictures with similar probability distribution are considered to have similar context [8]. In this paper [9], a set of similar behavior pictures is found by calculating mutual information between corresponding behaviors of different pictures. These images are considered to have similar contexts. Correspondingly, weather, time, road conditions and other labels in dataset are treated as context to perform multi-task classification of images, and same image can have multiple contexts. Since the labels are known and design structure is simple, our definition can more easily meet the needs of context in the field of autonomous driving.

TABLE I. FOR L TARGET VARIABLES (LABELS) , EACH OF K VALUES.

	$K = 2$	$K > 2$
$L = 1$	binary	multi-class
$L > 1$	multi-label	multi-task

III. METHODOLOGY

This paper compares differences between multi-class, multi-label, and multi-task, and then determined MTL as classification method. In addition, this work analyzes cross entropy loss commonly used in classification tasks, and finally uses reweighting [10] to adjust sample imbalance.

A. Comparison of different categories

Different classification methods are compared to choose method that meets our classification requirements.

Multi-class classification [11] means a classification task with more than two classes. Multi-class classification makes assumption that each sample is assigned to one and only one label. Activations and loss functions for this training task are Softmax [12] and categorical cross-entropy [13]. Reason for using Softmax is that its output values are correlated, and the sum of its probabilities is always 1. Once probability of one category is increased, the probability of other categories must be correspondingly reduced. After that each sample can only be marked one definition of categories.

Multi-label classification [14] assigns to each sample a set of target labels. This can be thought as predicting properties of a data-point that are not mutually exclusive. The activations and loss function are Sigmoid [15] and binary cross-entropy. Sigmoid is used because it processes each raw output value separately, making results independent of each other.

Multiclass-multioutput classification [16] (also known as multi-task classification) means that a single estimator has to handle several joint classification tasks. Set of labels can be different for each output variable. And activations and loss function are Softmax and categorical cross-entropy. Obviously multi-task classification is the classification method suitable for our context detection.

Relationships of three categories are shown in Table I.

B. Cross entropy function

Cross-entropy can be used to define a loss function [1] in machine learning and optimization. True probability y_i is true label, and given distribution q_i is predicted value of current model:

$$L = \frac{1}{N}\sum_{i}^{n} L_i = -\frac{1}{N}\sum_{i}^{n}\sum_{c=1}^{M} y_{ic}\log(p_{ic}), \tag{1}$$

where M is number of categories, y_{ic} is sign function (0 or 1), if true category of sample i is equal to c, y_{ic} takes 1, otherwise y_{ic} takes 0. p_{ic} is the predicted probability that observed sample i belongs to class c.

Classification model is based on learning process of loss as shown in Fig. 1.

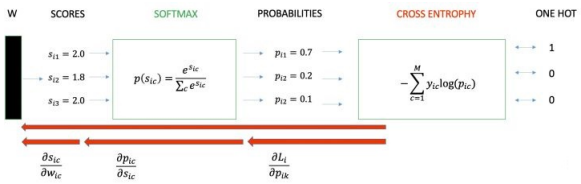

Fig. 1. Learning process of loss

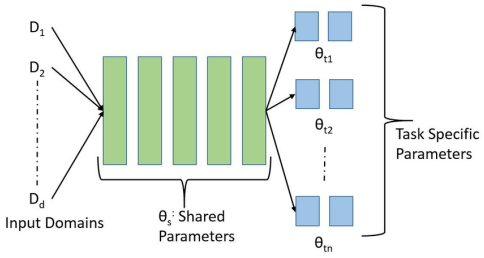

Fig. 2. Multi-task learning architecture [17]

Derivation process can be divided into three sub-processes:

$$\frac{\partial L_i}{\partial w_{ic}} = \frac{\partial L_i}{\partial p_{ik}} \cdot \frac{\partial p_{ik}}{\partial s_{ic}} \cdot \frac{\partial s_{ic}}{\partial w_{ic}}, \tag{2}$$

$$\frac{\partial L_i}{\partial w_i} = [\partial(s_i) - y_i] \cdot x_i. \tag{3}$$

Related to our task situation, our multi-task learning architecture is shown in Fig. 2.

C. Uniform sample distribution:

Sometimes classification problem of imbalanced data occur. A reweighting method is needed to make sample sampling tend to be balanced. There are many weighting methods, this work uses weighted cross entropy:

$$L = -\frac{1}{N}\sum_{i}^{n}\sum_{c=1}^{M} w \cdot y_{ic}\log(p_{ic}). \tag{4}$$

The idea of weighted cross entropy is to use a coefficient to describe importance of sample in loss. For a small number of samples, strengthen its contribution to loss, and reduce its contribution to loss for a large number of samples. Calculation logic of w is to assume that training dataset has M classes, number of samples in each class is n_i, and i ranges from 1 to M. Then there is a reciprocal way to calculate w. Find reciprocal of the number of M samples, and $1/n_i$ is coefficient of the corresponding category.

IV. EXPERIMENTS AND RESULTS

This paper compares the performance results of models with and without MTL, with and without reweighting. Experimental setup is firstly introduced and results are discussed.

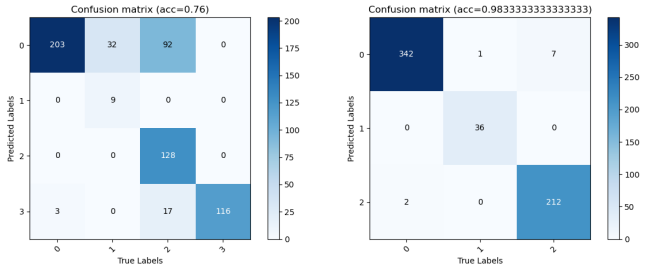

Fig. 3. Confusion matrix of period (left) and weather (right) with MTL and reweighting

A. Experimental setup

ONCE dataset [2] is used for this work. It is a large-scale autonomous driving dataset with 2D&3D object annotations. It contains 1 million LiDAR frames, 7 million camera images with diverse environments such as day/night, sunny/rainy, urban/suburban. 1500 images are randomly selected from ONCE dataset to ensure a uniform distribution across categories. Training set, validation set and test set are distributed in a ratio of 2:1:2. Each image contains two context, including three weathers and four time periods. Also our batchsize is 64 and epoch is 15.

B. Results

MTL: In this experiment, it is expected whether effect of MTL improves classification performance. In order to realize structure of MTL, model structure of ResNet18 [18] is modified, retaining the first seven layers of backbone and adding a pooling layer and a corresponding linear layer of each context. The model is trained with pretrained parameters. Such a model structure enables two tasks to be trained on one model simultaneously, and features can be shared between them. Correspondingly, two separate ResNet18 models are trained , each of which only completed one classification task. By comparing training results of the two groups, it can be seen that although training set is small, MTL can still improve classification performance of the model. Fig. 3 compares with Fig. 5, and Fig. 7 compares with Fig. 9. Since the model after reweighting already has a high classification ability for Period, adding MTL has little improvement on period and weather context.

Reweighting: Also improvement of the model by reweighting still needs to be shown. A model that removes the reweighting function is designed. By comparing Fig. 3 and Fig. 7, Fig. 5 and Fig. 9, it can be seen that the model is improved more after adding reweighting. However, since weather feature is easier for the model to classify, improvement is not very obvious. Fig. 3 is confusion matrix of the model with MTL and reweighting, Fig. 7 is confusion matrix of the model with MTL but without reweighting, Fig. 5 is confusion matrix of the model with reweighting but without MTL, and Fig. 9 is confusion matrix of the model with neither MTL nor reweighting as a control.

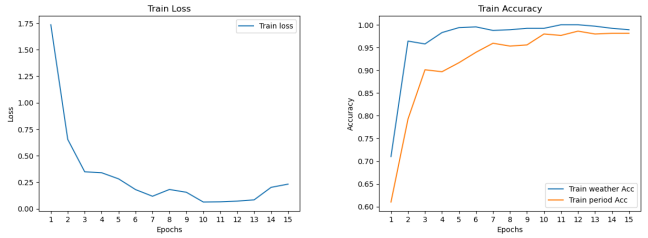

Fig. 4. Loss and accuracy with MTL and reweighting

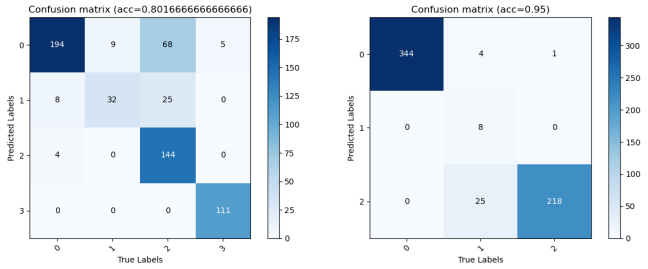

Fig. 5. Confusion matrix of period (left) and weather (right) with reweighting but without MTL

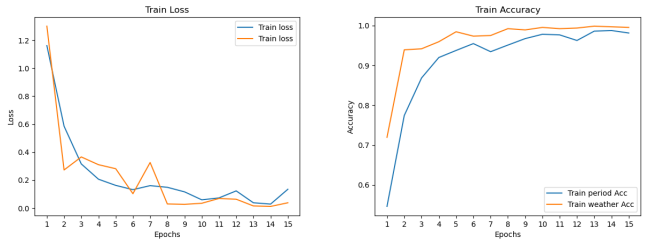

Fig. 6. Loss and accuracy with reweighting but without MTL

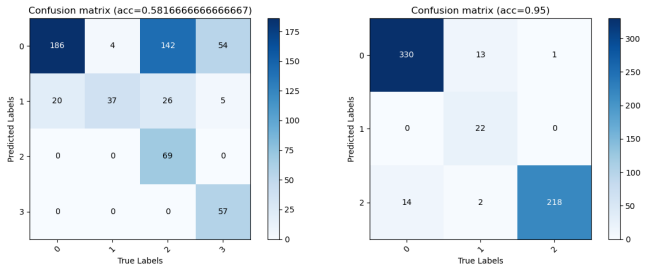

Fig. 7. Confusion matrix of period (left) and weather (right) with MTL but without reweighting

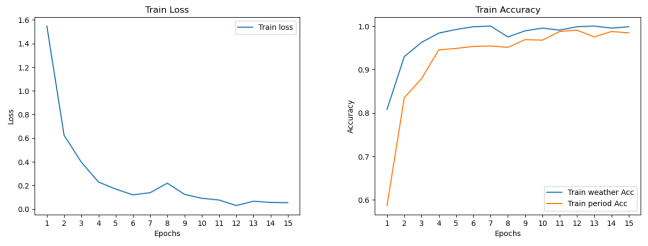

Fig. 8. Loss and accuracy with MTL but without reweighting

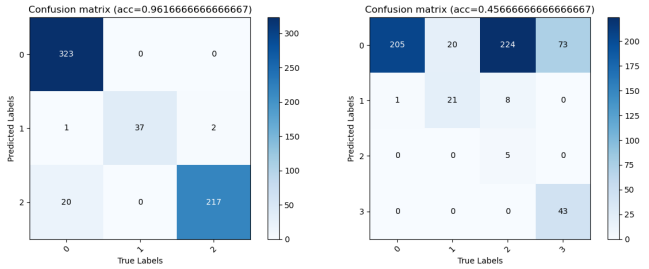

Fig. 9. Confusion matrix of period (left) and weather (right) without MTL and reweighting

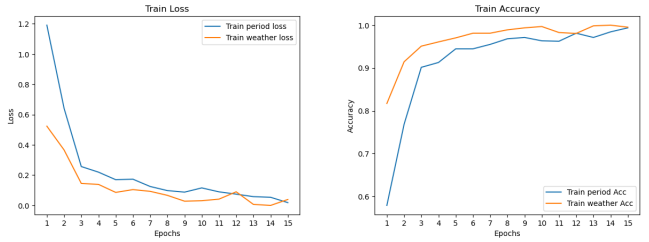

Fig. 10. Loss and accuracy without MTL and reweighting

V. Conclusion

Different views on definition of context can lead to divergent research directions. This work gives a general defination of context that it is considered as a scene or background other than a recognized subject, which can greatly affect the model's decision making and positioning of the detected subject. In addition, context based on ONCE dataset with the help of MTL is extracted and classification results are analyzed. The results prove that our model can effectively classify different context types for autonomous driving. Due to limited label types of the dataset, more context features cannot be consider . More work for examining other different contexts can be done in the future.

References

[1] Y. Lee, J. Jeon, J. Yu, and M. Jeon, "Context-aware multi-task learning for traffic scene recognition in autonomous vehicles," in *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2020, pp. 723–730.

[2] J. Mao, M. Niu, C. Jiang, X. Liang, Y. Li, C. Ye, W. Zhang, Z. Li, J. Yu, C. Xu *et al.*, "One million scenes for autonomous driving: Once dataset," 2021.

[3] L. Wolf and S. Bileschi, "A critical view of context," *International Journal of Computer Vision*, vol. 69, no. 2, pp. 251–261, 2006.

[4] T. M. Strat, "Employing contextual information in computer vision," *DARPA93*, pp. 217–229, 1993.

[5] O. Marques, E. Barenholtz, and V. Charvillat, "Context modeling in computer vision: techniques, implications, and applications," *Multimedia Tools and Applications*, vol. 51, no. 1, pp. 303–339, 2011.

[6] C. Galleguillos and S. Belongie, "Context based object categorization: A critical survey," *Computer vision and image understanding*, vol. 114, no. 6, pp. 712–722, 2010.

[7] E. Haussmann, M. Fenzi, K. Chitta, J. Ivanecky, H. Xu, D. Roy, A. Mittel, N. Koumchatzky, C. Farabet, and J. M. Alvarez, "Scalable active learning for object detection," in *2020 IEEE intelligent vehicles symposium (iv)*. IEEE, 2020, pp. 1430–1435.

[8] S. Agarwal, H. Arora, S. Anand, and C. Arora, "Contextual diversity for active learning," in *European Conference on Computer Vision*. Springer, 2020, pp. 137–153.

[9] M. Hasan and A. K. Roy-Chowdhury, "Context aware active learning of activity recognition models," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4543–4551.

[10] M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to reweight examples for robust deep learning," in *International conference on machine learning*. PMLR, 2018, pp. 4334–4343.

[11] M. Aly, "Survey on multiclass classification methods," *Neural Netw*, vol. 19, no. 1, p. 9, 2005.

[12] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.

[13] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Annals of operations research*, vol. 134, no. 1, pp. 19–67, 2005.

[14] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 3, no. 3, pp. 1–13, 2007.

[15] J. Han and C. Moraga, "The influence of the sigmoid function parameters on the speed of backpropagation learning," in *International workshop on artificial neural networks*. Springer, 1995, pp. 195–201.

[16] Y. Zhang and Q. Yang, "An overview of multi-task learning," *National Science Review*, vol. 5, no. 1, pp. 30–43, 2018.

[17] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "An all-in-one convolutional neural network for face analysis," in *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*. IEEE, 2017, pp. 17–24.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.