UTS Submission at the TRECVID 2020 Disaster Scene Description and Indexing Task

Qi Rao, Linchao Zhu, Yi Yang
The ReLER Lab
University of Technology Sydney, Australia
rao@student.uts.edu.au, {linchao.zhu,yi.yang}@uts.edu.au

Abstract

In this paper, we summarize the technical details applied in our submission of TRECVID 2020 Disaster Scene Description and Indexing (DSDI) task [1] . Our main effective improvements include three parts: data augmentation strategies according to the intrinsic characteristics of the LADI dataset, mixture of experts classifiers, and an ensemble strategy with multiple models. We split the dataset into train and validation set. The hyper-parameters are searched on our validation set. Performance on the validation set verify the effectiveness of our strategies.

1. Introduction

It is important to respond quickly to disasters for public safety. The rapid detection of disasters will be helpful in proposing effective responses, and recently has drawn increasing attention in industries and academia. Recently, Liu et al. [6] released the Low Altitude Disaster Imagery (LADI) Dataset for public safety's needs. LADI provided detailed division of disaster scenes and distinguished it from normal scenes. The goal is to recognize future disasters based on known disaster scenes from the Atlantic Hurricane and spring flooding seasons since 2015. Most of images were collected from aerial perspective.

2. Dataset

In our submission, we only used the provided LADI dataset [6] to train our models, which was proposed in 2019. Different from standard scene classification task, this special dataset has two unique characteristics: (1) most data were collected from a aerial perspective; (2) it focuses on the fine-grained division of disaster scenarios. The purpose is to conduct early detection and alarm for future disasters scenes. Compared with traditional scene recognition task, it brings extra difficulties in recognition. First, the limitation of single perspective results in context missing. For example, flooding scenario can be easily confused from natural water scenario through an airborne perspective. Second, the captured images are large in scale, which contributes to that the detailed objects are relatively small. However, in some cases, these detailed objects make one disaster scenario be different from others, which also differentiate the disaster from natural normal scenarios. For example, the scenario "flooding" and "ocean/river is difficult to distinguish from each other without the information from detailed context. Third, because of the expensive collection costs of airborne captured images, the data scale is limited and not enough to train robust classification models with high performance in a standard way.

In order to utilize limited data, and evaluate model's performance after training, we divided the full LADI dataset into a training set and a validation set, which were randomly selected into two parts with rate of 10:1.

3. Our Framework

Our framework includes a feature extraction module and a feature enhancement strategy.

3.1. Feature Extraction

Deep neural networks present outstanding representation ability on visual recognition tasks [5, 2, 8, 7]. We use several deep neural networks backbones to extract features from images, including ResNet-18, ResNet-50, ResNet-152 [3], Inception [9], SENet [4], Polynet [10]. Main results are listed in Table 2. All backbone models are pre-trained on ImageNet, then they are adapted to LADI dataset by fine-tuning.

3.2. Model Structures

We have tried several strategies to improve the performance. We found some critical factors that may influence the performance, which includes data augmentation, mixture of experts, and the loss function.

Model	mAP
ResNet-18	37.44
ResNet-50	37.18
ResNet-152	38.12
Inception-v4	37.52
SENet	38.39
PolyNet	37.56

Table 1. Overall mean AP results on LADI validation set with different backbones.

Data augmentation. Data augmentation is used to overcome the limitation of data scale and improve the final recognition performance. We observe that images in LADI dataset are mostly in an airborne perspective, that makes a large rotation angle reasonable during the data preparation procedure. We adjust the rotation angle from 20 degrees, which are commonly applied in a visual recognition task, to 360 degrees, for more comprehensive augmentation. Experiments are conducted on the same backbone ResNet-18 with a classifier. The classifier is a fully-connected layer, where its input size is the feature dimension, and the output size is the number of categories. Model trained with 20 degrees rotation receives 37.02 mean AP on validation set, while model trained with 360 degrees rotation receives 37.44 mean AP on validation set. Since LADI images are large in size, we resize and crop the input in data preparation procedure. And we add randomness for augmentation. The resize and crop size are 256 and 224 respectively. We use five-crop augmentation during testing

Mixture of experts. For better classification, we apply multiple classifiers in a single model and vote for the final result. Every classifier is a fully-connected layer, which receives the extracted feature vector as input and produces a score vector with category size. The final result vector is an average of all score vectors produced by classifiers. Experiments show that this strategy brings 0.8 mean AP gains compared with single classifier, from 36.64 mean AP to 37.44 mean AP.

Loss function. During training, scores are feeding into a sigmoid function to produce probabilities for each category. Since it is a multi-label task, a binary cross entropy loss is explored in our framework, In details, the binary cross-entropy loss is defined as

$$l(x, y) = -[y \cdot \log x + (1 - y) \cdot \log(1 - x)], \quad (1)$$

where x notes the produced probability, y notes the label.

4. Model Ensemble and Submission

In submission, we utilize ensemble strategy across multiple models. Models trained with different hyper-parameters such as learning rate, number of experts, data augmentation random factors, etc, are complementary to different scenarios. Ensembling all models that are with relative high performance on validation set can receive better performance. We select the top four results on validation set produced by different model combinations. Results show that the highest mean AP on validation set reach to 40.03 mean AP, over 3 points improvement compared with single model baseline. We use these four combinations of model selection to produce our final submission. The final test dataset is of about 5 hours videos. Each video can be regarded as a sequences of image frames. We feed each frame into our model and produce a score vector. Mean pooling of scores in all frames within a video contributes a video-level result. We use the four combinations of video selection to produce four sets of video results as our final submission.

Submission	combinations	selection	result
UTS.run.3	A+B+C+D+E+F+G	fully	0.281
UTS.run.2	A+B+C+D+E+F+G	partly	0.279
UTS.run.1	C+D+E+F+G	fully	0.227
UTS.run.4	C+D+E+F+G	partly	0.222

Table 2. Specifications and result of our final submission. UTS.run.3 and UTS.run.1 select fully 1000 entries for each category, while UTS.run.2 and UTS.run.4 select partly entries for each category. A and B are SENet50 and Resnet152 models trained on full LADI dataset (training and validation set), C, D, E, F, G are Inception, SENet50, Resnet50 pretrained on places365, Resnet50 pretrained on ImageNet, Resnet152 models trained on LADI training set.

References

[1] G. Awad, A. A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, A. Delgado, J. Zhang, E. Godard, L. Diduch, J. Liu, A. F. Smeaton, Y. Graham, G. J. F. Jones, W. Kraaij, and G. Quénot. Trecvid 2020: comprehensive campaign for evaluating video retrieval tasks across multiple application domains. In *Proceedings of TRECVID 2020*. NIST, USA, 2020. 1

[2] 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*, pages 770–778, 2016. 1

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1

[4] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 1

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*,

volume 25, pages 1097–1105. Curran Associates, Inc., 2012. 1

[6] J. Liu, D. Strohschein, S. Samsi, and A. Weinert. Large scale organization and inference of an imagery dataset for public safety. In *2019 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–6, Sep. 2019. 1

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1

[8] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[9] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *ICLR 2016 Workshop*, 2016. 1

[10] X. Zhang, Z. Li, C. C. Loy, and D. Lin. Polynet: A pursuit of structural diversity in very deep networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3900–3908, 2017. 1