Semantic Diversity Learning for Zero-Shot Multi-label Classification

Avi Ben-Cohen Nadav Zamir Emanuel Ben Baruch Itamar Friedman Lihi Zelnik-Manor
DAMO Academy, Alibaba Group
{avi.bencohen, nadav.zamir, emanuel.benbaruch, itamar.friedman, lihi.zelnik}
@alibaba-inc.com

1. More Experiments

1.1. Backbone Variations

In our experiments we use TResNet-M [2] as a backbone for our visual model, due to its efficiency and reported high accuracy on several competitive computer vision datasets. To further extend our analysis and comparison with prior works we also explore two popular backbone architectures, VGG19 [3] and ResNet50 [1] in Table 1. We report results using our approach as well as adding a comparison to Fast0Tag [4] loss function with our E2E training scheme as a baseline. As can be seen, using our approach with VGG19 as a backbone, the results in terms of mAP for both zero-shot and generalized zero-shot are superior compared to prior works but lower than our current backbone, while using ResNet50 as a backbone improves over VGG19 in all metrics. Best results are achieved using TResNet-M backbone. In addition it can also be seen that the results in terms of mAP for tag-based image retrieval using different backbone variations are higher than current prior works, suggesting that our training scheme extends and may improve the quality of various model architectures.

Table 1: Results using alternative backbones on NUS-WIDE test set. We report the results in terms of F1($K = 3$), F1($K = 5$), and mAP for ZSL and GZSL tasks. Best results are in bold.

Backbone	Method	Task	F1($K = 3$)	F1($K = 5$)	mAP
VGG19 [3]	Fast0Tag [4]	ZSL	24.2	22.2	20.2
		GZSL	11.7	13.0	6.6
TResNet-M [2]	Fast0Tag [4]	ZSL	25.7	23.3	21.6
		GZSL	15.4	16.6	9.7
VGG19 [3]	Ours	ZSL	29.0	26.5	24.2
		GZSL	16.8	19.0	9.9
ResNet50 [1]	Ours	ZSL	30.0	27.6	24.4
		GZSL	17.7	20.1	11.2
TResNet-M [2]	Ours	ZSL	**30.5**	**27.8**	**25.9**
		GZSL	**18.5**	**21.0**	**12.1**

(a) (b)

Figure 1: Examples of (a) a diverse image, and (b) a less diverse image with an average cosine distance of 0.7, and 0.4 respectively.

2. Reproduciblity

To support future research in the field, we currently work to publish our trained models and share a reproducible training code on GitHub.

3. Additional Qualitative Results

To visualize images with high or low semantic diversity, we have computed the average cosine distance (1-cosine similarity) per image, on 10K images taken from NUS-Wide validation set. The mean and std of the cosine distance was 0.53 ± 0.20. In Figure 1 we show a more diverse image and a less diverse image as an example.

We present in Figure 2 additional qualitative results using our proposed method for several sample images from NUS-WIDE test set. It can be seen that in several cases the unseen tags (marked by asterisks) are ranked in the top-10. In addition, while some of the unseen tags are incorrect based on the ground truth annotation, in most cases there exists a noticeable semantic relation between these tags to the image.

graffiti
art
London
mural
England
urban
green
paint
war
politics

officers
protesters
riot
politics
__police__
London
men
roadblock
protest
soldier

firefighter
demonstration
France
Canada
riot
action
officers
winter
sport
__protest__

wildlife
nature
deer
moose
elk
Canada
wild
animals
park
Alaska

football
crowd
cheering
soccer
baseball
red
England
game
parade
basketball

bride
__wedding__
Hawaii
sea
bravo
beautiful
white
couple
groom
dress

Figure 2: Qualitative results showing the top-10 tags retrieved using our proposed method. Bold text represents the correct tags according to the provided ground truth in NUS-WIDE test set. Asterisks mark unseen tags.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[2] Tal Ridnik, Hussam Lawen, Asaf Noy, Emanuel Ben Baruch, Gilad Sharir, and Itamar Friedman. Tresnet: High performance gpu-dedicated architecture. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1400–1409, 2021. 1

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

[4] Yang Zhang, Boqing Gong, and Mubarak Shah. Fast zero-shot image tagging. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5985–5994. IEEE, 2016. 1