PLSA-BASED ZERO-SHOT LEARNING

Wai Lam Hoo and Chee Seng Chan

Centre of Image and Signal Processing
Faculty of Computer Science & Information Technology
University of Malaya, 50603 Kuala Lumpur, Malaysia
{*wailam88@siswa.um.edu.my; cs.chan@um.edu.my*}

ABSTRACT

Current zero-shot learning methods relied on attributes to describe the unseen class characteristics, using the learned seen class model. However, these approaches required extensive attribute labels on each object class, and a well-defined, attributes relationship between the seen and unseen class with the aid of human knowledge. In this work, we avoid these with a novel learning process using the probabilistic Latent Semantic Analysis (pLSA). We replace the attributes with topic model and extend the representation as a mapping algorithm to object classes, so that zero-shot learning would be possible. With this, less annotated class information is required to achieve similar performance. Evaluations on three public datasets had shown the effectiveness of our proposed method.

***Index Terms*—** Zero-shot learning, pLSA, object detection, object recognition

1. INTRODUCTION

In a real time situation, there is a large amount of object classes that required to be recognised by humans. However, in the event of an unknown object class, humans tend to employ existing object class that one understands and finds a relationship between them to describe the unknown class. As an example in Figure 1, an unknown class 'mule' can be described by using the 'horse' and 'donkey' class, respectively. This is how zero-shot learning arises.

Current zero-shot learning approaches have been very much focused on using attributes as the image feature descriptions [1, 2, 3]. These approaches had shown promising results. However, they required extensive attribute labels on each object class, and a well-defined, attributes relationship between the seen and unseen class. This is impractical as it requires exhaustive human interventions. Besides that, there are object classes that are difficult to be described using such attributes relationship.

The research work is partly sponsored by High Impact Research (HIR) under grant no. UM.C/625/1/HIR/MOHE/FCSIT/08 and UM Bright Sparks Programme.

(a) Horse (b) ? (c) Donkey

Fig. 1: Zero-shot learning example. To learn the unknown class (b), we relate it to known class (a) horse and (c) donkey.

In this paper, we attempt to tackle these problems by employing the pLSA model and relate the unseen class by a novel mapping algorithm using the signature topics set. Our proposed approach, first, build a Bag-of-Words (BoW) representation from the seen class image features. Then, a pLSA model is learned from the BoW. Finally, we deduce the signature topic mapping for the unseen classes based on the learned pLSA model, and so zero-shot learning can be performed.

1.1. Contributions of the work

Zero-shot learning receives wide attention in recent years. Palatucci et. al [4] investigated this problem in missing word classes that semantically related, following by Lampert et. al [5] that studied unseen object class based on attributes. Unlike [5], Parikh and Grauman [1] extended the attributes' performance in zero-shot approaches with the introduction of relative attributes, while [3, 6] focused on large-scale dataset problems. Although attributes seem a very efficient way to perform zero-shot learning, it has two major limitations. In the former, the dataset needs to be intra-class, while in the latter, binary or relative relationship between all classes need to be defined. One must note that such a process requires very extensive human supervision efforts. In here, our **first contribution: topic model -** we replace the attributes with topic model so that less annotated class information is required to achieve similar performance [1, 5]. To our very best knowledge, this is the first attempt to use topic model in zero-shot learning.

Topic models are widely applied in image classification [7]. The topic models are particularly effective when pairing with BoW representation, where the models group ambiguous codewords together and generate a topic distribution over a codebook. One of the most popular topic models is pLSA [8] which serves as a mid-level clustering method and tries to find the relationship of codewords. The codewords are grouped together for common yet meaningful representations. Besides pLSA, Latent Dirichlet Allocation (LDA) [9] is also a well-known variant of topic models. Our **second contribution: pLSA -** we employ pLSA to replace the attributes during the zero-shot learning. We choose pLSA because it doesn't require any prior information compares to LDA model. We also further extend the topic model representation as a mapping algorithm to object classes, so that zero-shot learning would be possible. With this, the requirement of human supervision is reduced, as well as handling the inter-class variation problem.

2. METHODOLOGY

2.1. BoW model and pLSA

We first build a BoW model to represent images. Specifically, we use the Pyramid Histogram of Gradient (PHOG) [10]. However, we didn't concatenate all the HOG descriptor found. Instead, we put all the features in a codebook learning mechanism using the Random Forest (RF) similar to [11]. Therefore, we can obtain a set of HOG descriptors that quantize shape information locally and globally by the nature of the PHOG. The RF codebook can learn image shapes as a whole, as well as local patch characteristic.

Then, we learned the pLSA model using the built BoW. The topic-specific image distribution $p(z|d)$ is learnt as

$$p(w|d) = \sum p(w|z)p(z|d), \tag{1}$$

where w denotes codewords, d denotes images, and a latent topic z is the mid-level information across images under different object categories d using the quantized BoW information w.

2.2. Zero-shot learning via signature topics

In this section, we discuss given the learnt pLSA model, how to relate unseen classes using the seen classes' information by Coarse Class (CC) and Fine Class (FC) relationship, respectively. Following that, we shows in detail how zero-shot learning is performed using the novel signature topic and coarse-to-fine relationship.

2.2.1. Coarse Class and Fine Class - their relationship

In zero-shot learning, we have object classes that is 'seen', that is availability of training samples for the classes. Also, we have unseen classes where there are no training images available during the training phase. Our task here is how to define unseen classes' model using the seen classes' information a.k.a. zero-shot learning. In particular, we utilise pLSA to relate the seen and unseen class. Adopting [12], we define

Definition 1 Fine Class (FC) as a specific object class that need to be classified and is a subset to one of the Coarse Class.

Definition 2 Coarse Class (CC) as a large concept class that shares a conceptual similarity, either physical or biological, within its own FC.

Both the FC and CC can be expressed as to Eq 2:

$$\forall_{c_{FC} \in c_{CC}} \;\; c_x \sim c_y; \;\; c_x, c_y \in c_{FC} \tag{2}$$

where c denotes class $c \in C$ and $\sim$ relates both classes in some manner. For example, CC can be electrical devices and its associated FC are refrigerator, washing machine, etc.

2.2.2. Signature topic mapping

Employing the learnt pLSA model, firstly, we denote a seen class as $s \in S$ and an unseen class as $u \in U$, where $S, U \in C$. Secondly, we introduce a novel mapping algorithm that uses m number of topic sets from all the topic sets, M to represent each seen class s ($m \ll M$). We denote this as the signature topic set ST_s:

$$ST_s = \arg\max_{ST_m} \sum_{m \in M} p(ST_m|d), \tag{3}$$

where size of M is 2^n and n is number of latent topics. Taking $n = 3$ as example, the size of M is 8 ([0 0 1], [0 1 0], [1 0 0], [0 1 1], [1 0 0], [1 0 1], [1 1 0], [1 1 1]), where 1 indicates the signature topic(s) and vice versa.

Therefore, for each unseen class u, we can employ a pair of ST_s to predict the ST_u by:

$$ST_{S_x} \sim ST_u \sim ST_{S_y}, \tag{4}$$

where class S_x and S_y are random seen classes picked from the same CC. ST_u is inferred as the union of the ST_s pairs to achieve zero-shot learning. Ideally, if ST_{S_x} is [0 0 1] and ST_{S_y} is [1 0 0], then ST_u is [1 0 1].

Finally, class c' can be predicted by evaluating $p(ST_{c'}|d_t)$:

$$p(c'|d_t) = \frac{p(ST_{c'}|d_t)}{\sum_{c \in C} p(ST_{ST_c}|d_t)}. \tag{5}$$

where t are the test images.

3. EXPERIMENTS

In the experiments, we use 3 different public datasets: PubFig [2], Cifar-100 [12], and Caltech-256 [13] to evaluate the effectiveness of our proposed approach. In order to evaluate $p(c'|d_t)$, 1-vs-all classification is performed. Unless specified, all dataset features are extracted using PHOG with 3 pyramid levels, 180° angle and 20 bins. For the RF codebook, it is learnt using 10 trees and 100 leafnodes.

Table 1: Performance evaluation (%) of the proposed method in different numbers of unseen class, n.

Unseen class, n	1	2	3	4	5
PubFig	58.89 ± 4.94	54.35 ± 4.80	54.99 ± 4.93	51.65 ± 3.80	51.30 ± 5.32
Cifar-100	56.84 ± 0.57	54.85 ± 0.75	-	-	-
Caltech-256	52.14 ± 6.04	51.49 ± 6.42	51.86 ± 5.74	52.13 ± 6.49	51.32 ± 6.20

Table 2: Comparison of the proposed method with classical solutions in general inference task (no unseen class).

PubFig (%)			Cifar-100 (%)		
pLSA	[1]	[5]	**pLSA**	[14]	[15]
67.83 ± 0.91	62.00 ± 1	37.00 ± 1	**58.13 ± 0.30**	53.70 ± 1	54.80 ± 1

3.1. PubFig

Identical subset as in [1] is used where 8 random identities are extracted with each class consists of 100 images. The PHOG features are computed and RF codebook is built using 10 trees with 100 leafnodes each. The PHOG features are quantized to BoW space and build pLSA model using 11 topics, similar to the number of attributes used in [1]. We also employ the class relationship in [1] to find the unseen class topic set as Eq. 4. However, the optimum nearest seen classes pair between the unseen classes are chosen, and we assume the ($\succ$) relationship in [1] is similar to our ($\sim$) relationship.

Figure 2 shows our results and a comparison with relative attributes [1] and binary attributes using DAP [5] in terms of different number of unseen categories. From this, we notice that our proposed method has a better performance consistency ($\pm 7\%$) in compare to [1]($\pm 23\%$) and [5]($\pm 17\%$) respectively. This shows the ability and effectiveness of the proposed mapping algorithm to handle the intra-class variation problem as opposed to extensive attributes annotation [1, 5].

3.2. Cifar-100

Cifar-100 [12] has 100 classes and each class contains 600 images with 32x32 resolutions. The 100 classes are further grouped into 20 CC. Each CC has 5 FC, where n of them is(are) unseen. The resultant from this, the total unseen class is $n*20$. We picked 30 training images randomly, and the rests are testing images. In this dataset we use 10 topics, given the intuition that 10 major semantic topics exist in the CC which are 'mammals', 'size', 'trees', 'vehicles', 'food', 'household', 'insects', 'reptiles', 'people', and 'flower'. Besides, due to low resolution on these images, we only use 2 pyramid levels and 50 codewords per tree.

Generally, when using least seen classes during the learning process, poorer accuracy is expected. However interestingly in Table 1, the accuracy difference of our proposed

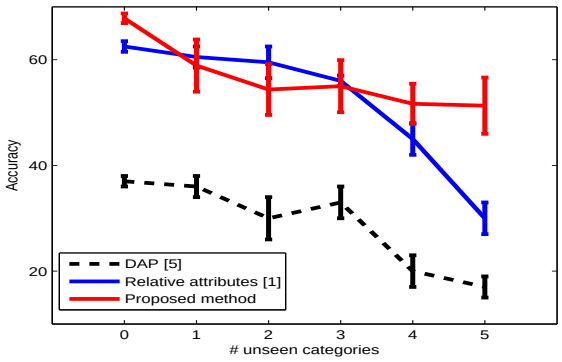

Fig. 2: Accuracy vs. Number of unseen categories: Comparison of the proposed method and [1, 5] in PubFig dataset.

method between $n = 1$ and $n = 2$ are only differed by a fraction of $\pm 2\%$ even the difference number of unseen classes between them is large ($\sim$ 20 classes). This has shown the robustness of the proposed method. Due to a limited number of FC in each CC, we were only able to perform zero-shot learning up to 2 unseen classes per CC, which is a total of 40 unseen FC. This is because we need at least two seen classes to model an unseen class. If we have more than 3 unseen classes in a CC, all the unseen classes will be modelled by the same seen classes, and this is invalid as it will lead to all the unseen class to have the same signature topic.

3.3. Caltech-256

Caltech-256 dataset [13] consists of 30607 images grouped into 256 object classes and a background class. However, it doesn't have any CC concept. Therefore, we introduce a way to group the classes to their respective CC that follows Cifar-100 CC except some specific classes that need new CC concept. In Table 3, we show the distribution of selected Caltech-256 classes with 5 existing CC as in Cifar-100 and the introduction of 4 new CC, with a total up to 9 CC. Only 158 of the Caltech-256 classes are grouped here because the CC for those ungrouped object categories had very least FC members. For this dataset, total unseen class is $n * 9$.

From the results in Table 1, we observed the outcome is more fluctuated in comparison to PubFig and Cifar-100 results. This may cause by the FC in some of the CC are semantically related but with very low visual similarity, e.g. 'computer keyboard', 'computer monitor' and 'computer mouse',

Table 3: Coarse Class (CC) concept for selected Caltech-256 dataset.

CC	Caltech-256 class (FC)
household electrical devices	binoculars, boom-box, bread maker, calculator, CD, computer keyboard, computer monitor, computer mouse, floppy-disk, head-phones, iPod, joystick, laptop, light bulb, megaphone, microwave, palm-pilot, paper-shredder, PCI-card, photocopier, refrigerator, rotary-phone, toasters, treadmill, tripod, VCR, video-projector, washing machine
household furniture	bathtub, chandelier, chess-board, desk-globe, doorknob, ewer, flashlight, hammock, hot-tub, hourglass, mailbox, mattress, menorah, picnic table
large man-made outdoor things	buddha, Eiffel-tower, golden-gate-bridge, light-house, minaret, pyramid, skyscraper, smokestack, teepee, tower-Pisa, windmill
medium mammals	dog, duck, elk, goat, goose, llama, minotaur, penguin, porcupine, raccoon, skunk, swan, unicorn, zebra, greyhound
vehicles	blimp, bulldozer, cannon, canoe, car-tire, covered-wagon, fighting-jet, fire-truck, helicopter, hot-air-ballon, kayak, ketch, license-plate, motorbikes, mountain-bike, pram, school-bus, segway, self-propelled-lawn-mower, snowmobile, speedboat, steering-wheel, touring-bike, tricycles, wheelbarrow, airplanes, car-side
household daily items	beer-mug, chopsticks, coffee-mug, knife, spoon, stained-glass, paperclip, paper-shredder, coins, dice, drinking-straw, dumb-bell, fire-extinguisher, frying-pan, ladder, pez-dispenser, playing-card, roulette-wheel, screwdriver, Swiss-army-knife, tweezer, umbrella
sports	baseball-bat, baseball-glove, baseball-hoop, billiards, bowling-ball, bowling-pin, boxing-glove, football-helmet, Frisbee, golf-ball, skateboard, soccer-ball, tennis-ball, tennis-court, tennis-racket, yo-yo
wears	cowboy-hat, diamond-ring, eyeglasses, football-helmet, necktie, sneaker, socks, top-hat, t-shirt, human-wear, wielding-mask, yarmulke, tennis-shoes, saddle, stirrups
musical instruments	electric-guitar, french-horn, grand-piano, guitar-pick, harmonica, harp, harpsichord, mandolin, sheet-music, tambourine, tuning-fork, xylophone

which belong to CC = 'household electrical devices'. Following that, error rates will increase when unseen classes are 'computer mouse', but related to 'computer monitor', and 'computer keyboard' respectively.

3.4. General inference task

Though the work is focused on zero-shot learning, we also showed that the proposed method is capable to perform general inference tasks. According to Table 2, our proposed method outperforms the state-of-the-art methods [1, 5, 14, 15] in PubFig and Cifar-100 datasets, respectively. This shows the effectiveness and the robustness of the proposed method. We didn't perform inference on Caltech-256 dataset as there are only 158 classes are extracted; it is not comparable and fair to any state-of-the-art solutions.

3.5. Discussion

In the experiments, they are some cases where the predicted ST_u is redundant. That is, if a lower number of topics is chosen, the numbers of possible M will also be reduced. Hence, there is chances that different unseen class u will have the same ST_u. In order to solve this, all the experiments use good number of topics to minimize the redundancy of ST_c.

Based on the experiments in Caltech-256, we aware that classification accuracy is fluctuating due to the FC collection quality under each CC. FC within CC is grouped based on semantic relationship. However, these FC might be visually dissimilar. This problem is likely to be solved by introducing a middle-level class group to further assign the FC within CC to some high-visual similarity group, e.g. we can further group FC: 'head-phones', 'rotary-phones' and 'megaphone' in CC: 'household electrical devices' to be 'phones'. When we pick the random seen classes to model 'megaphone', 'head-phones' and 'rotary-phones' will have priority as the related seen class.

4. CONCLUSION

This paper presents a novel approach of zero-shot learning, where we learn unseen classes by modelling latent topics of seen classes. This is in contrast to attributes approach that requires extensive human intervention, and we achieve comparable performances. The main advantage of the topic model is it allows intra-class prediction while attributes do not. Latent topic model automatically finds the most relevant topics for images, and therefore, we can identify each object class with a signature topic set ST. Our future work includes introduce tighter relationship between the FC in the same CC. With this, we can expect to achieve better performance in zero-shot learning. We are also interested to look into relatively large-scale dataset, with considerably good resolution.

5. REFERENCES

[1] D. Parikh and K. Grauman, "Relative attributes," in *IEEE International Conference on Computer Vision, ICCV 2011*, nov., pp. 503 –510.

[2] N. Kumar, A.C. Berg, P.N. Belhumeur, and S.K. Nayar, "Attribute and simile classifiers for face verification," in *IEEE 12th International Conference on Computer Vision, ICCV 2009*. IEEE, pp. 365–372.

[3] M. Rohrbach, M. Stark, and B. Schiele, "Evaluating knowledge transfer and zero-shot learning in a large-scale setting," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011*. IEEE, pp. 1641–1648.

[4] M. Palatucci, D. Pomerleau, G. Hinton, and T. Mitchell, "Zero-shot learning with semantic output codes," *Advances in neural information processing systems, NIPS*, vol. 22, pp. 1410–1418, 2009.

[5] C.H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009.* IEEE, pp. 951–958.

[6] O. Russakovsky and L. Fei-Fei, "Attribute learning in largescale datasets," in *ECCV 2010 Workshop on Parts and Attributes*, 2010, vol. 1.

[7] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, and T. Tuytelaars, "A thousand words in a scene," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, pp. 1575 –1589, 2007.

[8] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning J.*, vol. 42, pp. 177–196, 2001.

[9] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2005*.

[10] A. Bosch, A. Zisserman, and X. Muoz, "Image classification using random forests and ferns," in *IEEE 11th International Conference on Computer Vision, ICCV 2007.* IEEE, pp. 1–8.

[11] F. Moosmann, E. Nowak, and F. Jurie, "Randomized clustering forests for image classification.," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, 2008.

[12] Alex Krizhevsky, "Learning multiple layers of features from tiny images," Tech. Rep., 2009.

[13] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," 2007.

[14] I. Goodfellow, A. Courville, and Y. Bengio, "Large-scale feature learning with spike-and-slab sparse coding," *International Conference on Machine Learning, ICML*, 2012.

[15] Y. Jia, C. Huang, and T. Darrell, "Beyond spatial pyramids: Receptive field learning for pooled image features," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2012*. IEEE, pp. 3370–3377.