Zero-shot Classification at Different Levels of Granularity

Matías Molina
Universidad Nacional de Córdoba, Argentina.
matias.molina@unc.edu.ar

Abstract

Zero-shot classification (ZSC) is the task of learning predictors for classes not seen during training. The different methods proposed in literature are evaluated over specific datasets with their specific class partitions, but little attention has been paid to the impact of the dataset granularity when ZSC is performed. The novelty of this work is to generate synthetic datasets by controlling their granularity level to analyze the ZSC performance afterwards. Moreover, it presents an approach that allows us to preserve the visual and semantic structures. The experiments show that ZSC performance exhibits strong differences depending on the data granularity and it reveals the relevance of both visual and semantic spaces when performing ZSC.

1. Introduction

In recent decades, image classification systems have improved dramatically. This has been achieved by the availability of large amounts of labeled images (used to train the classification models) in addition to the remarkable increase in computational capacity. While the amount of data is sufficiently large, the performance of the classifiers can vary depending on how many samples of a given category are available. In practice, the image datasets exhibit a long tail phenomena, namely, many instances are observed for a small set of categories but few instances for a large number of classes. In this context, paradigms such as few-shot learning have emerged, which aims at learning classifiers with few training samples. Zero-shot learning is the task of learning classifiers for categories for which no instances have been seen during training [9].

The general approach to address the zero-shot classification problem is to use a set of known categories to train the model and adapt it to the (unknown) target categories. To achieve this, it is necessary to introduce semantic information that allows us to describe both training and query categories. From the beginning of the zero-shot learning paradigm, different types of side information have been explored taking the form of visual attributes [9], word-embeddings [11], class hierarchies [1], etc. In literature, visual attributes are shown as the most effective output code representation. Since all the categories are represented by the same source of information, it is possible to learn how the known classes are related to the unknown.

As could be seen, in ZSC not only is the image representation (visual space) important but also the class representation (semantic space). One of the most common approaches to address this task are based on learning a projection between both visual and semantic representation spaces [1, 2, 16]. The idea is to project one space onto the other in order to compare how compatible are the different concepts for a given query image. Based on this approach, the proposal of this study is to use a bilinear compatibility model based on a structured support vector machine model SSVM-multiclass [19] similar than [2] but adding different penalty terms to preserve the geometric structure of the visual or semantic spaces (or both).

The most common procedure in ZSC literature is to compare the different methods using some of the typical datasets. The most used datasets are Animal with Attributes (AWA) [9], Caltech-UCSD Birds (CUB) [20] and Sun Attributes (SUN) [14]. These datasets have the particularity of providing visual attributes to describe its categories. While AWA is considered a *coarse-grained* dataset, CUB and SUN are considered *fine-grained*. The *granularity* of these datasates are conceptually defined by considering the meaning of the categories (for instance, AWA is composed of different types of animals and CUB different species of the same animal). In this context, AWA is a coarse-grained dataset since its categories are more distinguishable concepts compared with CUB. Despite that, there are no specific studies with respect to the granularity relationships between them.

In this work we don't have the intention of comparing the performance of zero-shot classifiers over the different literature datasets since they have been extensively studied. Instead, we pretend to create synthetic datasets of different granularity levels with the intention of analyzing the impact

of ZSC applied on one dataset with different granularities.

As the ZSC setting needs a suitable semantic space, an appropriate proposal is to use one of the literature datasets as base to create the synthetic data. The data generation is based on the Caltech-UCSD Birds (CUB) [20] dataset and the synthetic data are created by introducing a parameter to define the granularity level. The goal is to obtain different variants of one dataset, where each one of them differs in its granularity. Then, use them to compare the impact of the visual and semantic space when ZSC is performed.

In summary, the contribution of this study is to create synthetic datasets to analyze how ZSC works on different levels of granularity and also to explore the relevance of the visual and semantic space.

In the following section, we will introduce the model proposal, details of the data generation as well as metrics for measuring the dataset granularities.

2. Related work

One of the earliest definitions of learning without labeled data was presented in 2008 by Larochelle *et al*. under the name of *zero-data learning* [10]. At 2009 Lampert *et al*. [9] present the zero-shot learning definition and address it by using visual attributes. The methods presented at [9] are Direct Attribute Prediction (DAP) and Indirect Attribute Prediction (IAP), both algorithms are based on two stages. For instance, DAP learns an image classifier by combining a class-attribute predictor with an attribute-image predictor. Another popular family of methods are based on linear mappings between the visual and semantic space. Convex Combination of Semantic Embeddings (ConSE) [13] defines an implicit projection by defining a convex sum of semantic embeddings. Deep Visual Semantic Embedding (DeViSE) [4] is based on Ranking SVM [8], it maps the visual features to the word-embedding category representations. Attribute Label Embedding (ALE) [1] uses a bilinear compatibility function and a weighted approximation to the ranking objective function [23]. Similarly, Structured Joint Embedding (SJE) optimizes a ranking loss (based on the SSVM-multiclass [19]) by giving more importance to the top of the list. Embarrassingly Simple Approach to Zero-Shot Learning (ESZSL) [16] propose a linear model that could be solved by a closed-form.

More recently, models based on generative adversarial networks (GANs) have gained popularity. These methods generate synthetic (or "fake") images to obtain training instances of the target categories [18, 25]. Although these methods report the most competitive results, it is important to notice that they need to use the semantic representation of the unknown categories for training, which is a different characterization from the original zero-shot learning setting. This characterization is identified as class-transductive instance-inductive [21].

Beyond these general aspects of the models, different approaches have been proposed by realigning the geometric structures of the visual and semantic spaces. For instance, [7] learns a new representation (class-prototypes) for the visual and semantic spaces that are projected into a new common space where the similarities between the images and concepts are performed. The method aims to improve the semantic description by exploiting information from the visual space. The authors suggest an advantage of using the visual space as the embedding space for classification using nearest-neighbor. Similarly, [22] shows that optimizing visual space is beneficial for zero-shot learning. Also, [17] proposes to analyze the contribution of the visual and semantic information for ZSC showing a preference over the visual space to a greater or lesser extent. In [15] the authors propose to improve the unsupervised word-embedding by realigning them using the visual space as reference. The approach uses a triplet loss formulation in order to disambiguate unsuitable meanings. This method works as a preprocessing, not being specifically a ZSC algorithm. Another approach [26] is to build a graph to describe the neighbors structure of the semantic and visual spaces and learn how to project both spaces into a new space to remove the visual-semantic ambiguity.

3. Preliminaries

In zero-shot classification (ZSC) we are given a training set $\mathcal{D}^{\text{tr}} = \{(x_i, y_i)\}_{i=1}^{N}$ of image-label pairs, $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}^{\text{tr}} \subset \mathcal{Y}$, sampled from a known set of visual categories. The goal is to learn a mapping $f : \mathcal{X} \to \mathcal{Y}$ from $\mathcal{D}^{\text{tr}}$ to classify samples over a different set $\mathcal{Y}^{\text{ts}} \subset \mathcal{Y}$, where $\mathcal{Y}^{\text{tr}} \cap \mathcal{Y}^{\text{ts}} = \emptyset$. If $\mathcal{Y}^{\text{tr}}$ and $\mathcal{Y}^{\text{ts}}$ are not disjoint sets, the problem is known as generalized zero-shot learning. In addition we assume that each class y has a semantic representation $\mathbf{y} \in \mathbb{R}^e$ (*e.g.* visual attributes) and each image has its feature vector $\mathbf{x} \in \mathbb{R}^d$. We use a bilinear scoring function $F : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$:

$$F(x, y; W) = \mathbf{x}^T W \mathbf{y} \tag{1}$$

that measures the compatibility between an input image x and a concept y. The parameters of this function ($W \in \mathbb{R}^{d \times e}$) are learned using $\mathcal{D}^{\text{tr}}$. After training the prediction are defined by

$$\hat{y} = \underset{y \in \mathcal{Y}^{\text{ts}}}{\operatorname{argmax}} F(x, y; W). \tag{2}$$

Using this function, Akata *et al*. [2] propose to optimize the following loss function:

$$\ell_{sje}(X,Y;W) = \frac{1}{N}\sum_{n}\max_{y\in\mathcal{Y}}\{0, \mathbb{I}[y \neq y_n] + F(x_n, y; W) - F(x_n, y_n; W)\} \quad (3)$$

with $\mathbb{I}[b]$ = 1 if b is true and 0 in the other case.

4. Approach

This work has two contributions: first, the definition of the model to optimize the classifier by preserving the visual (semantic) structures. And second, the synthetic data generation with parametric granularity.

4.1. Model

The model proposal is to extend the loss function of the Eq. (3) with the aim to control the structures of the semantic and visual spaces. This extension is made by adding a new term that encodes the relationship between both spaces. Intuitively, the encoding takes the form of a graph where the vertices are categories and the edges represent the distances between them. In the case of the semantic space, the definition is straightforward because each category is represented by a vector. In order to have a single representation for each visual category, we averaged the instances of each class. Then, we can define two matrices that encode the similarity between the categories in the visual and the semantic space:

$$G^X = \bar{X}\bar{X}^T \quad (4)$$

$$G^Y = \Phi\Phi^T. \quad (5)$$

with, $\bar{X}_y = \text{avg}\{\mathbf{x}|x \in \mathcal{X}, class(x) = y\}$ and $\Phi_y^T = \mathbf{y}$, $y \in \mathcal{Y}$. In the same manner,

$$G^{XW} = \bar{X}W(\bar{X}W)^T = \bar{X}WW^T\bar{X}^T \quad (6)$$

$$G^{WY} = \Phi W^T(\Phi W^T)^T = \Phi W^T W\Phi^T \quad (7)$$

are the matrices that define the graph to represent the visual (semantic) space after projecting it to the semantic (visual) space. G_{ij}^X is the similarity between the classes i and j in the visual space but G_{ij}^{XW} is the similarity of i and j when they are projected to the semantic space and one more time against the visual space. Similarly for G_{ij}^Y and G_{ij}^{YW}.

We can also encode the structure of both spaces jointly:

$$H_{\odot}^{XY} = G^X \odot G^Y, \quad H_{\odot}^{XWY} = G^{XW} \odot G^{WY} \quad (8)$$

with $\odot$ the Hadamard product, *i.e.* $(A \odot B)_{ij} = A_{ij}B_{ij}$[1]. The idea behind this is to weigh those similarities that are important enough in both spaces simultaneously.

Finally, our loss function is defined by:

$$L(X,Y,\Phi;W) = \ell_{sje}(X,Y,\Phi;W) + \lambda\ell_{\psi}(X,\Phi;W) \quad (9)$$

With λ an hyperparameter that trade-off between ℓ_{sje} and the preservation term ℓ_{ψ}. The latter could be defined in different flavors:

$$\ell_{\text{mul}} = ||H_{\odot}^{XY} - H_{\odot}^{XWY}||_F^2 \quad (10)$$

$$\ell_{\text{mul}(\mathbf{x}\rightarrow\mathbf{y})} = ||G^{XW} - G^Y||_F^2 \quad (11)$$

$$\ell_{\text{mul}(\mathbf{y}\rightarrow\mathbf{x})} = ||G^X - G^{WY}||_F^2 \quad (12)$$

Notice that while the Eq. (10) aims to preserve the structures of both spaces, Eq. (11) uses the semantic space as a reference to align the visual space and Eq. (12) uses the visual space as an anchor to drive the structure of the semantic space.

4.2. Synthetic data generation

To be able to analyze the impact of ZSC on different granularities, it is proposed to create synthetic datasets. Since the ZSC problem requires that the categories could be represented by an auxiliary source of information (*e.g.* visual attributes) we must consider this fact when synthesizing. Thus, we need to create datasets with variable granularity but with a suitable semantic representation. The proposed idea is to manipulate the CUB [20] dataset in order to variate the granularity while maintaining the given attribute representation of the categories.

It is possible to generate datasets of different granularities using the Guyon [5] method[2]. This algorithm creates clusters of points normally distributed (for a given variance) about vertices of an n-dimensional hypercube and assigns an equal number of clusters to each class. We can adapt this mechanism in order to maintain the original geometric relationship of the visual features of CUB instead of using the hypercube vertices. The adaptation works as follow:

0. Use the average of each class of a given dataset as the input centroids.
1. After normalizing the centroids. Scale them by some scalar γ.
2. Generate an specific number of random points around each centroid normally distributed by a standard deviation σ.

[1] Notice that $\odot$ could be replaced by any element-wise operator.

[2] https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_classification.html

3. Normalize all the vectors to the unit sphere.

If we repeat this process using different scale factors γ and maintaining the same σ, we obtain datasets of different granularities. Namely, if $\gamma_1 > \gamma_2$ and using the same σ, the clusters generated by γ_1 have less intra-class distance and bigger inter-class distance with respect to the clusters generated by γ_2. Thus, γ defines the class separation. The idea is shown in Figure 1.

As a result we can manipulate any well known dataset used in ZSC literature obtaining different granularities and keeping the structure of the visual centroids and using the original semantic space which is crucial for the knowledge transfer in ZSC.

5. Experiments

Our experiments are based on the optimization of the Eq. (9) and using the different options as in the Eq. (10), (11) and (12). The synthetic datasets are based on the CUB dataset and ResNet101 [6] features. It means that the visual space is given by the features extracted from the ResNet101 model and the average per class vectors are used to initialize the Guyon algorithm described in Section 4.2. The semantic space is given by the default class description, namely, it is defined by the visual attributes of the CUB dataset.

ZSC performance is measured by the average per-class top-1 accuracy [24]. The model is trained using SGD. At validation time, the learning rate and the hyperparameter λ (Eq.(9)) are searched in the set $\{10^{-1}, 10^{-3}, 10^{-5}, 10^{-7}\}$, taking the values that maximize the validation accuracy. As a preprocessing, L2-normalization is applied on both the visual and semantic representations. Regarding the train/test class split, it uses the proposed split by Xian *et al.* [24] and also four different random splits.

For a sanity check, we measure the granularity of the dataset $\mathcal{D}$ by the average standard deviation as follow:

$$\bar{\sigma}(\mathcal{D}) = \frac{1}{C}\sum_{c=1}^{C}\bar{\sigma}(c), \bar{\sigma}(c) = \frac{1}{d}\sum_{i=1}^{d}[\sigma(c)]_i, \qquad \sigma(c) = \text{std}\{\mathbf{x}|class(x) = c\}. \tag{13}$$

with C the number of classes. The insight behind this is to measure how compressed the clusters are. Thus, if the class separator γ increases, $\bar{\sigma}(c)$ must decrease. Also, we use the RSM (Revised Silhouette with Medoids index), RankM (Ranking with medoids index) [3]:

$$RSM(\mathcal{D}, \delta) = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta(x_i, c')}{\delta(x_i, c_{x_i})} \tag{14}$$

where c_{x_i} is the centroid of the cluster to which x_i belongs

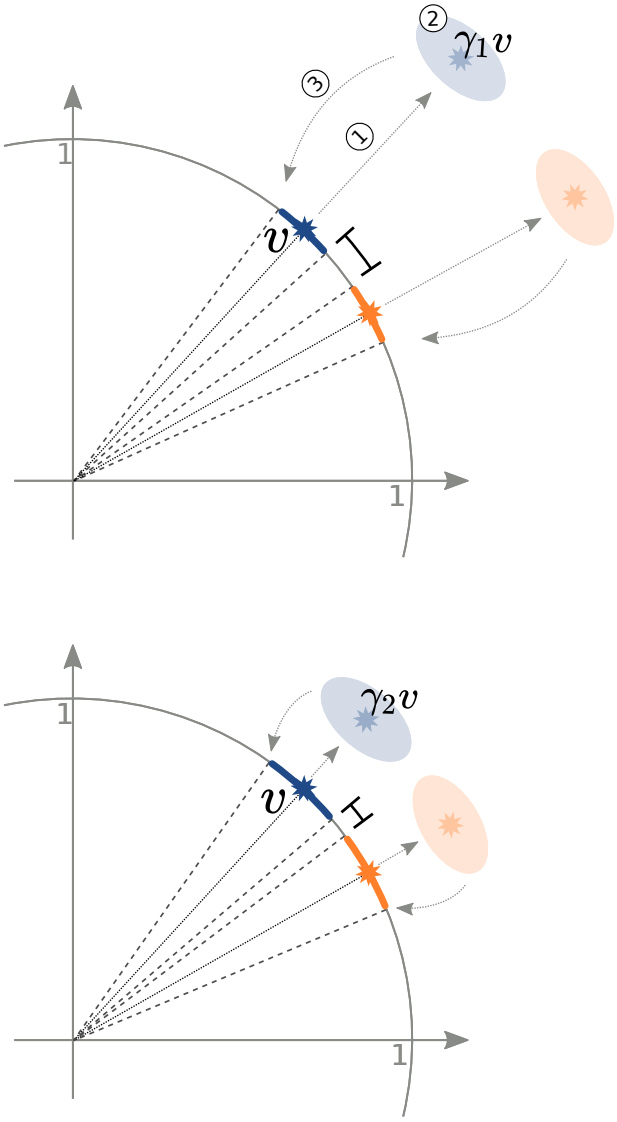

Figure 1. Data generation with different granularities. The image shows two different datasets. The space at the right has a finer granularity than the left space. The centroids (stars) are scaled by γ_1 (left) and γ_2 (right) and the clusters are generated for the same standard deviation (shadow). Then, all vectors are normalized to the unit sphere. As $\gamma_1 > \gamma_2$ the separation of the resulting clusters is greater on the left space, generating a more coarse-granularity respect to the other space.

(*i.e.* the ground-truth class of x_i) and c' is the centroid closest to x_i: $c' = \text{argmin}_{c \neq c_{x_i}}\, \delta(c, x_i)$, for a given distant δ.

$$RankM(\mathcal{D}, \delta) = 1 - \frac{C}{n(C-1)}\sum_{i=1}^{n}\left(1 - \frac{1}{R_{ic}}\right), \tag{15}$$

where R_{ic} is the rank of the x_i's class centroid among all class centroids. In addition, we run a SVM linear clas-

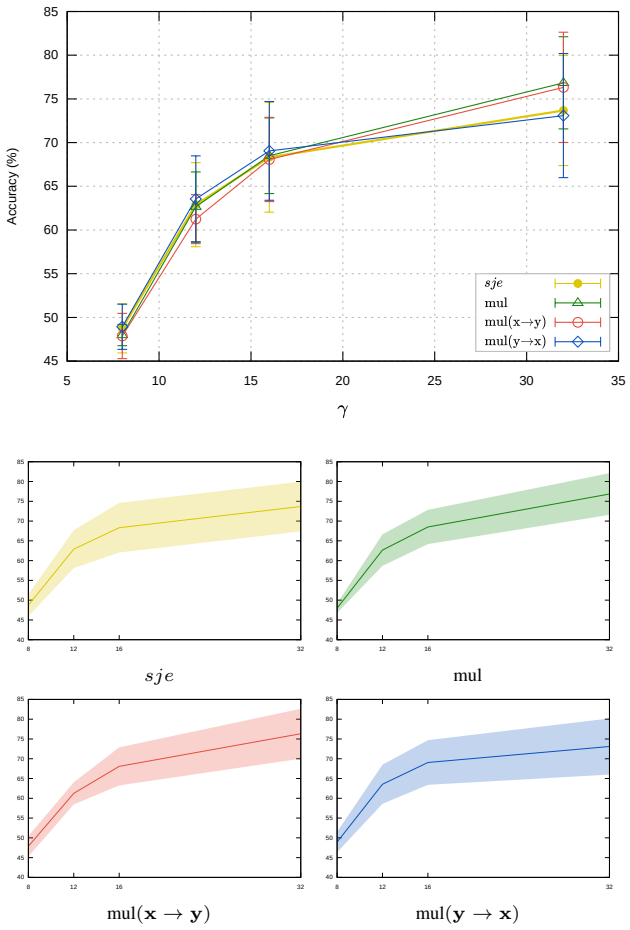

Figure 2. Top: Average per-class top-1 accuracy and std. dev. of the methods. Bottom: Average per-class top-1 accuracy and std. dev. of each proposal separately.

sifier to check if the classification accuracy improves as the class separator grows. The SVM is trained and tested with a 70/30 class split over the complete dataset.

All these metrics are summarized in Table 1. It shows that as the class-separator (γ) increases, the RSM increases and the SVM classification improves. Also, the average std.dev. ($\bar{\sigma}$) decreases. Which indicates that effectively the data generation works as expected. Notice that the RankM measure is not useful in this case, this comes from the fact that RankM measures the granularity of a given datasets based on how many points of any cluster are closer to another cluster. The last statement is interesting because it shows that not all granularity metrics are equally useful.

Table 2 shows the results of applying the different proposals (Eqs. (9)–(12)) and they are summarized in Figure 2. sje indicates the SJE method (Eq. (9) without extensions: $\lambda = 0$). $\text{mul}(\mathbf{x} \to \mathbf{y})$ indicates the SJE loss function extended with $\ell_{\text{mul}(\mathbf{x}\to\mathbf{y})}$ (Eq. (9)+(11)) and similarly for $\text{mul}(\mathbf{y} \to \mathbf{x})/\ell_{\text{mul}(\mathbf{y}\to\mathbf{x})}$ and $\text{mul}/\ell_{\text{mul}}$. Since we created

γ	$\bar{\sigma}$	rankm	rsm	SVM
8	0.0215	1	1.021	65.5
12	0.0211	1	1.030	91.1
16	0.0206	1	1.042	98.2
32	0.0178	1	1.117	100

Table 1. Different granularity measures for each datasets.

	γ	sje	mul	mul $(\mathbf{x} \to \mathbf{y})$	mul $(\mathbf{y} \to \mathbf{x})$
PS	8	56.26	55.14	55.01	55.31
	12	70.93	71.95	72.50	71.61
	16	76.77	78.74	78.44	78.10
	32	84.21	88.79	86.35	81.73
S_1	8	48.73	47.27	46.8	48.97
	12	60.53	62.56	59.58	62.09
	16	64.36	67.65	63.00	67.31
	32	70.8	77.65	75.01	70.06
S_2	8	52.51	49.22	51.66	52.38
	12	69.89	68.23	64.29	70.84
	16	77.43	74.68	74.00	77.43
	32	82.96	83.98	83.71	82.96
S_3	8	48.00	48.78	47.18	48.17
	12	62.00	60.86	63.00	61.36
	16	67.50	67.03	69.70	65.23
	32	71.90	73.13	78.00	66.42
S_4	8	48.73	47.27	46.8	48.97
	12	60.53	62.56	59.58	62.09
	16	64.36	67.65	63.00	67.31
	32	70.80	77.65	75.01	70.06
Avg (std)	8	48.75 (2.81)	47.98 (1.22)	47.87 (2.59)	48.93 (2.58)
	12	62.90 (4.81)	62.66 (3.99)	61.25 (2.80)	63.55 (4.94)
	16	68.32 (6.28)	68.50 (4.33)	68.07 (4.82)	69.05 (5.65)
	32	73.68 (6.30)	76.84 (5.27)	76.32 (6.31)	73.09 (7.10)

Table 2. Average per-class top-1 accuracy over the synthetic datasets of different class separators γ. S_i are random class splits and PS is the split proposed by Xian *et al*. [24]. Avg(std) over $\{S_i\}$.

the synthetic data using the CUB dataset, it is possible to run the classifiers with the class partition proposed by Xian *et al*. (PS) [24]. To analyze the variability we created four different random splits $(S_1, ..., S_4)$.

The results show that as the class separability γ grows, the classification improves as expected. In addition, the variability with respect to the class partition is higher when the granularity is more coarse, a similar conclusion is mentioned in [12]. This phenomena is a clear tendency independently of the method, which implies that ZSC is more stable over fine-grained datasets. Such observations suggest that the variability of the classification task in zero-shot setting can cause confusion when comparing methods using coarse-grained dataset. Therefore it may be crucial to consider granularity and variability as part of the evaluation protocol. Not paying attention to granularity (particularly in the coarse-grained cases) is to neglect the variability, which could bias the selection of a ZSC method in practice.

Also, notice that there are differences when the method includes the structure of the visual (semantic) space as a guide for the semantic (visual) space. For instance, when considering the case of the greater separability (*i.e.* $\gamma = 32$) both $\mathrm{mul}(\mathbf{x} \rightarrow \mathbf{y})$ and mul improve around 3 points, which suggest that the semantic structure plays an important role. Beyond this particular case, another observation is that mul presents a better trade-off between the accuracy and the variability. This could indicate that it is not convenient to consider one of the spaces as more important than the other. Thus, it leads us to hypothesize that there must exist an optimal combination between both visual and semantic space. This conclusion is in line with [17]. These observations could be better seen in Figure 2.

6. Conclusion

In this work an approach to analyze the behavior of ZSC applied on synthetic datasets of different granularity levels is proposed. We discussed how to generate suitable datasets using the class centroids of the visual space of CUB to initialize the Guyon algorithm jointly with a scale factor. This allows us to construct the same visual space with different granularities maintaining the original semantic space.

The results show that for the fine-grained datasets the performance variability is lower compared with the coarse-grained cases, as seen in all the proposed methods. In addition, the results suggest that both visual and semantic space are relevant and they play different roles in the classification process which leads us to formulate the hypothesis of a possible optimal combination of the information provided for each space.

As a summary, the most important result throughout all the experiments is the observed correlation between the variability and the data granularity. This is a crucial observation, especially with coarse-grained data, because the high performance variability could bias in practice the selection of one method over another. Thus, as a methodological conclusion, this study suggests to include the dataset granularity as an important characteristic to obtain a more comprehensive evaluation protocol for ZSC.

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 819–826, 2013. 1, 2

[2] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2927–2936, 2015. 1, 2

[3] Yin Cui, Zeqi Gu, Dhruv Mahajan, Laurens Van Der Maaten, Serge Belongie, and Ser-Nam Lim. Measuring dataset granularity. *arXiv preprint arXiv:1912.10154*, 2019. 4

[4] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, pages 2121–2129, 2013. 2

[5] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3(null):1157–1182, Mar. 2003. 3

[6] 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. 4

[7] Huajie Jiang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Learning class prototypes via structure alignment for zero-shot recognition. In *Proceedings of the European conference on computer vision (ECCV)*, pages 118–134, 2018. 2

[8] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proc. of the eighth ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining*, pages 133–142, 2002. 2

[9] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958. IEEE, 2009. 1, 2

[10] Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In *AAAI*, volume 1, page 3, 2008. 2

[11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 1

[12] Matias Molina and Jorge Sanchez. Performance variability in zero-shot classification. In *LatinX in AI at Neural Information Processing Systems Conference 2020*. Journal of LatinX in AI Research, dec 2020. 6

[13] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination

of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013. 2

[14] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2751–2758. IEEE, 2012. 1

[15] Ruizhi Qiao, Lingqiao Liu, Chunhua Shen, and Anton van den Hengel. Visually aligned word embeddings for improving zero-shot learning. *arXiv preprint arXiv:1707.05427*, 2017. 2

[16] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, pages 2152–2161, 2015. 1, 2

[17] Jorge Sánchez and Matías Molina. Trading-off information modalities in zero-shot classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3841–3849, January 2022. 2, 6

[18] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 8247–8255, 2019. 2

[19] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6(Sep):1453–1484, 2005. 1, 2

[20] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1, 2, 3

[21] Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. A survey of zero-shot learning: Settings, methods, and applications. *ACM Tr. on Intelligent Systems and Technology (TIST)*, 10(2):1–37, 2019. 2

[22] Xinsheng Wang, Shanmin Pang, Jihua Zhu, Zhongyu Li, Zhiqiang Tian, and Yaochen Li. Visual space optimization for zero-shot learning. *arXiv preprint arXiv:1907.00330*, 2019. 2

[23] Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. 2011. 2

[24] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 4, 5

[25] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *Proc. of the IEEE Conf. on Computer Vision and pattern recognition*, pages 5542–5551, 2018. 2

[26] Li Liu Yang Long and Ling Shao. Attribute embedding with visual-semantic ambiguity removal for zero-shot learning. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 40.1–40.11. BMVA Press, September 2016. 2