Ensuring Surjectivity in Zero Shot Learning

Rishabh Singh Yannick Strümpler Batuhan Yardim Nicolas Zobernig

Abstract

In this project, we investigate the effectiveness of our a novel approach to Zero-Shot Learning. Our method extends the usual approach of learning a mapping between semantic and visual embedding space by ensuring it to be surjective. By imposing surjectivity, we demonstrate our models generalize better to unseen classes. Moreover, we show that using triplet loss leads to better performance than traditional L2 loss. We compare our method to different baselines on the most commonly used datasets for Zero-Shot Learning.

1. Introduction

Visual Recognition has recently gained interest within both industry and research community, especially after the proposal of deep CNN models. Traditional recognition methods pursue supervised approach, thereby requiring significant corpus of annotated data for good performance. Major drawbacks of such approaches are the expense and prohibition in collecting large, diverse samples for training a generic model. This becomes cumbersome when fine-grained annotations are needed. Thus, it is imperative to develop recognition methods which can identify novel classes during test time with limited or even no instances from that class available during training.

Zero-Shot Learning (ZSL) ([5]) attempts to classify images of previously unseeen classes by leveraging additional information about them learnt during training on seen classes. In fact, ZSL is motivated by the human cognitive learning mechanism in identifying unknown classes. For instance, a person can easily recognise a "zebra" if he/she has only seen "horses" before and learnt that "zebra" is like a "horse" with black and white stripes. On similar lines, ZSL attempts to discover an intrinsic mapping between visual features and semantic features from observed objects and generalize it to unobserved categories.

We attempt to improve ZSL in various ways in our project. A majority of the methods so far fail to explicitly utilize the structure of unsupervised semantic vectors directly (such as word embeddings from NLP), which can carry valuable information. Moreover, in ZSL literature the semantic space is usually constructed on the basis of annotated visual attributes (such as "zebras are striped") whereas attempts to utilize word embeddings have been limited so far, and thus ZSL still requires additional supervision. Moreover, the domain shift across seen and unseen classes hinders the generalization ability of the learned visual-semantic mapping. Finally, the recognition performance is affected by inter-class and intra-class variations. In this work, we present a novel approach to zero-shot learning based on the encoder-decoder paradigm. We highlight our contribution in the following:

- In order to alleviate the domain shift and generalization problem, we enforce our cross-domain mapping to be surjective from image embedding space to the semantic space, so that every semantic label gets mapped to by some (unseen) image. We achieve this by proposing a novel loss function and a two-way architecture.
- We use the "triplet loss" from metric learning literature to favor compact clusters for samples of the same class and a better distinction between different classes. This explicitly enforces the visual-semantic mapping to behave nicely during test time.
- We enforce the mapping from images to their visual embeddings to be bijective to complement our theory. For this we propose using the i-RevNet [11] as a bijective feature extractor.
- We increase the information content of the semantic space by leveraging word and attribute vectors and their combinations. Word embeddings do not require additional human annotation and are available online from NLP models. Thus using them incurs almost no additional cost, and they can perform competitively.

2. Related Work

Our work builds on the intersection of Visual Recognition and Natural Language Processing. Here we briefly discuss existing works that are closely related to our approach.

Embedding Space. Current methods majorly use deep CNN models to learn the visual representations of images. The type of the network varies widely, however the most commonly used ones are ResNet [9], VGG [20] and GoogleNet [23]. None of the existing works enforce any constraints in visual space. Attributes are the most widely used representation for semantic space [13]. However, annotation of these attributes is expensive and cumbersome. Semantic word vector space has started gaining popularity especially in large-scale zero-shot learning recently [21][7]. The word vector space proves to be better scalable as no manually defined ontology is required and any class name can be represented as a word vector for free. People have also explored direct learning from textual descriptions of categories such as Wikipedia articles [4], sentence descriptions [16]. Karessli *et al*. [12] even made attempts to exploit human gaze as the auxiliary information, presenting that human gaze is indeed class specific. The final goal would be to alleviate the task of annotation from experts.

Zero Shot Learning. Early ZSL used a two-stage approach where the attributes of any given image are assigned in the first stage, then its class label is inferred by searching the class-attribute table using the nearest neighbor classifier. Direct Attribute Prediction (DAP) and Indirect attribute prediction (IAP) adopt the hidden layer of attributes as variables decoupling the

images from the layer of labels [14]. These methods suffer from distribution difference between the intermediate and target tasks. Recent works in ZSL focus on learning a mapping function between semantic and visual embedding space. These methods can be classified into three distinct groups. The first group learns a projection function from visual to semantic space $f : \mathcal{V} \rightarrow \mathcal{S}$ [13]. The second group learns the projection in the reverse direction $f : \mathcal{S} \rightarrow \mathcal{V}$ [2], and the third group learns an intermediate space where both visual and semantic vectors are projected into and a joint embedding is learnt [17].

The typical nature of the function learnt across spaces varies widely across methods. Some methods tackle ZSL by learning a linear mapping between the two embedding spaces [7, 13, 17]. These are extended with a non-linearity using a neural network to learn the mapping [22]. [1] proposed to build a bilinear compatibility function across the visual and the semantics via a ranking loss. [13] proposed an auto-encoder paradigm where an encoder aims to project a visual feature vector into the semantic space while the decoder exerts an additional reconstruction constraint. [2] builds on the same strategy and preserve semantic relations in the embedding space using cosine-similarity. Marginalized Latent Semantic Encoder (MLSE) gave current SOTA results for conventional ZSL. MLSE [3] is learned on the augmented seen visual features and the latent semantic representation. Meanwhile, latent semantics are discovered under an adaptive graph reconstruction scheme based on the provided semantics. Generative networks are increasingly being used to visually generate unseen categories from text descriptions and train a supervised classifier with hallucinated features [6, 28].

Generalised Zero-Shot Learning. The conventional ZSL aims to predict label of a test sample among unseen classes only. Generalized Zero-shot Learning (GZSL) extends the conventional zero-shot learning task to cases with both seen and unseen classes at test time [18]. Some methods use label embeddings [7] whereas others learn latent representations for images and classes[27] to tackle GZSL. We hypothesise that the surjectivity of the learnt mapping aids in performance for GZSL.

3. Model and Methods

Methodology. A typical approach in ZSL literature learns a parametrized function (i.e. a neural network) $f_\theta : \mathcal{V} \rightarrow \mathcal{S}$ that maps visual embeddings to their corresponding semantic embeddings. We follow this framework while also learning an inverse function $g_\theta : \mathcal{S} \rightarrow \mathcal{V}$. The spaces $\mathcal{S}$ and $\mathcal{V}$ denote the sets of semantic and visual embeddings of the whole dataset. We divide the sets into $\mathcal{S}^u, \mathcal{S}^s$ and $\mathcal{V}^u, \mathcal{V}^s$ where the super script u denotes that the set only contains embeddings of unseen classes and s denotes that the set only contains embeddings of seen classes. At training time we have no access to $\mathcal{V}^u$, i.e., the image embeddings of the unseen classes. We assume that visual embedding features are generated by a deep CNN F. In this approach, during test time, one maps an image to visual embedding space via F and then use one of the learned functions to relate the visual embedding of the test image to the semantic embeddings of all possible test classes and finds the closest match.

Model. The general goal in ZSL is to learn a mapping between the visual embedding space and the semantic embedding space. The architecture we chose is a symmetric nonlinear autoencoder where the encoder transforms the semantic representation to the visual representation $g_\theta : \mathcal{S} \rightarrow \mathcal{V}$ and the decoder reconstructs the semantic representation from the visual embedding $f_\theta : \mathcal{V} \rightarrow \mathcal{S}$. Unlike conventional autoencoders our latent representation is not arbitrary but should actually be equal to the visual embedding. For this the models must enforce additional constraints, explained below.

Classification. Classification, given the models, is performed through nearest neighbour search in either $\mathcal{S}$ or $\mathcal{V}$. In the former case we map the test example $\mathbf{v}_{test} \in \mathcal{V}^u$ to semantic space and find the nearest neighbour:

$$\begin{aligned} \hat{\mathbf{s}}_{test} &= f(\mathbf{v}_{test}) \\ \mathbf{s}^\star &= \underset{\mathbf{s}_{test} \in \mathcal{S}^u}{\arg\min} \|\hat{\mathbf{s}} - \mathbf{s}\|_2 \end{aligned}$$

To perform classification in visual space in some experiments we first map all possible semantic vectors to visual space and then find the nearest neighbor:

$$\begin{aligned} \hat{\mathcal{V}}^u &= g_\theta(\mathcal{S}^u) \\ \mathbf{v}^\star &= \underset{\hat{\mathbf{v}} \in \hat{\mathcal{V}}^u}{\arg\min} \frac{\mathbf{v}_{test} \cdot \hat{\mathbf{v}}}{\|\mathbf{v}_{test}\| \|\hat{\mathbf{v}}\|} \end{aligned}$$

As suggested in [2] we used cosine similarity instead of L2 loss for the nearest neighbour search in visual space.

Triplet Loss. For learning a meaningful mapping between the two spaces suitable for classification we use a triplet loss function as opposed to a traditional L2 metric, which we argue is better suited for nearest neighbour queries. During training, for each labeled visual-semantic vector pair $(\mathbf{v}, \mathbf{s})$, we obtain the

$$\mathbf{s}_n = \underset{\mathbf{s}' \in \mathcal{S}, \mathbf{s}' \neq \mathbf{s}}{\arg\min} \|\mathbf{s} - \mathbf{s}'\| \tag{1}$$

and minimize the loss

$$\ell(\mathbf{v}, \mathbf{s}, \mathbf{s}_n) = \Big[\|f_\theta(\mathbf{v}) - \mathbf{s}\|^2 - \|f_\theta(\mathbf{v}) - \mathbf{s}_n\|^2 + m\Big]_+, \tag{2}$$

where m is a constant margin. This loss corresponds to finding the closest negative semantic vector and computing the triplet loss, similar to the idea of mining hard triplets in face detection literature [19]. In our case, mining difficult triplets is easy, since we have a small number of possible semantic vectors. Note that for using nearest neighbour classification, an image vector v is classified correctly if and only if the semantic vector corresponding to the ground truth is closer to $f(v)$ than the closest negative semantic vector. The objective (2) actually directly optimizes over this possibility, plus a margin.

Surjection loss. Since during training ZSL models can learn to ignore unseen classes completely, we introduce a novel surjection loss based on the idea that a function is surjective if and only if it

has a right inverse. This can be viewed as autoencoder reconstruction loss in a specific direction. The novelty we introduce here is to use the semantic embeddings of both the seen and unseen set to compute the reconstruction loss during training. The general idea of the reconstruction loss is to limit the choice of mapping functions such that the composition $f_\theta \circ g_\theta$ is close to the identity. This leads to the following differentiable loss function:

$$\mathcal{L}_{\mathbf{s}} = \mathbb{E}_{\mathbf{s}\sim\mathcal{S}}\left[\|f_\theta(g_\theta(\mathbf{s})) - \mathbf{s}\|^2\right], \tag{3}$$

which we dub "surjection loss". In other words, this function maps to all possible semantic class embeddings. The relevance of this becomes clear at test time where we are given an image embedding of an unseen class $\mathbf{v} \in \mathcal{V}^u$ and try to predict the correct semantic embedding $\mathbf{s} \in \mathcal{S}^u$. If f_θ is only surjective on $\mathcal{S}^s$, i.e., $f_\theta : \mathcal{V} \to \mathcal{S}^s$ and the correct semantic embedding $\mathbf{s} \notin \mathcal{S}^s$ we cannot find the the correct embedding by computing $\hat{\mathbf{s}} = f_\theta(\mathbf{v}) \in \mathcal{S}^s$.

Note that for true generalizability, we would like the mapping $f_\theta \circ F$ from images to semantic vectors to be surjective. Regular CNNs used by existing methods can still violate surjectivity on $\mathcal{S}$ by mapping images to a low dimensional manifold in $\mathcal{V}$. To prevent this and complement our theory, we use a provably bijective feature extractor CNN variant (named i-RevNet) F so that the composition $f_\theta \circ F$ is surjective.

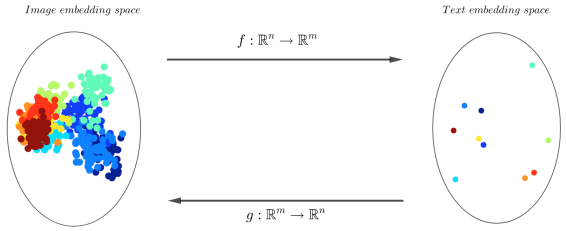

Figure 1. Representation of the proposed idea. We argue f should be made surjective by learning the right inverse g.

An abstract outline of our method can be seen in Figure 1. Our idea suggest that f_θ -the mapping from visual features to semantic features- should be surjective, and this is achieved by learning a right inverse g_θ. This can for instance force the orange visual vectors -which are near red the red class' visual vectors- to be mapped to the orange sample in semantic space -which is also nearby the red semantic vectors- to satisfy surjectivity. This enforces the model to learn the mapping for orange vectors even though no training samples for it are given.

4. Experiments

4.1. Datasets and Embedding Vectors

aPY. [5] The *Objects with Attributes* dataset contains images from the PASCAL VOC 2008 [10] challenge and also images scraped from Yahoo, each object annotated with a bounding box and visual attributes. The classes are typically quite simple such as dog, zebra, building, chair. The 64 visual attributes are on a per instance basis, so as is common practice for this dataset, the per instance features were averaged to obtain the per class attributes. The dataset contains 32 object classes in total, we use 20 of these during training and the rest for evaluation.

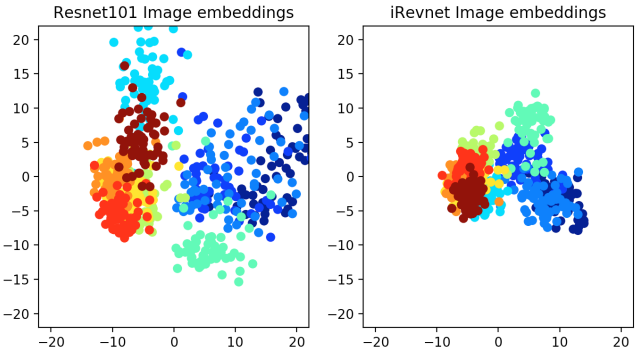

Figure 2. Embeddings of 9 different classes of the CUB Dataset, projected down to 2 dimension using PCA. *Left:* Traditional Resnet101 [8] image embeddings. *Right:* iRevNet [11] image embeddings.

AwA2. The *Animals with Attributes 2* Dataset [25] consists of 37'322 images of 50 different animal species. Each class is annotated with 85 attributes describing different properties of the animal, like its physical appearance, social behavior, typical habitat and more. We use 40 classes (*seen* classes) for training and the remaining 10 (*unseen* classes) are used in testing.

CUB. [24] The *Caltech-UCSD Birds-200-2011* dataset contains 200 classes of bird species annotated with 312 visual attributes and bounding boxes that only contain the bird are provided. We split the dataset into 150 seen and 50 unseen classes.

SUN. The *Scene Understanding* dataset [26] consists of around 14'340 images from 717 scene categories. Each image is annotated annotated with 102 attributes such as 'camping', 'farming', 'sports' etc. We use 645 *seen* classes in training an evaluate on the remaining 72 *unseen* classes.

Visual embedding. For the visual embeddings, we use the features of the iRevNet network [11], where we extract the data in the last layer before the fully-connected layers. We use the proposed splits from [25] in order to obey to the Zero-Shot Learning paradigm and avoid using test classes that have been used in training the network. In Figure 2 we have the visual embeddings for various classes visualized via PCA.

Word embedding. We use word embeddings of the class label obtained from GloVe [15] with 6B tokens. For classes with a composite label (*i.e.* more than one word) we construct a weighted average of each word. The specific weights we use are class-specific but oftentimes the first words correspond to descriptive attributes and are thus less relevant (*e.g.* for the class *blue whale* we assign more weight to *whale* as it is more relevant).

4.2. Baseline Results

SAE. Our first baseline from literature is the semantic autoencoder, which similar to ours in that it learns two functions in an autoencoder setup. However, SAE proposes learning a linear map (as in PCA), and the decoder structure is reversed com-

pared to ours. In other words, SAE incorporates an autoencoder in $\mathcal{V} \rightarrow \mathcal{S} \rightarrow \mathcal{V}$ compared to our $\mathcal{S} \rightarrow \mathcal{V} \rightarrow \mathcal{S}$. Note that within our framework SAE corresponds to learning an injective function (since they learn a left inverse) as opposed to a surjective one. However, this implies every semantic vector can be assigned to a single visual vector even though images of the same object can vary, potentially hurting their performance.

Linear projections. Even though in our approach we learn an inverse mapping and a highly non-linear forward map, a straightforward simplification of our model would be only learning a single linear transformation from one space to the other minimizing an L2 loss. For this baseline, we only learn the network f_θ minimizing the L2 loss with the target semantic vector, i.e. $\|\mathbf{s} - f_\theta(\mathbf{v})\|^2$.

4.3. Results

In this section we evaluate our method and conduct a detailed ablation study. We start by evaluating the impact of using the surjectivity constraint. We then provide a comparison between L2 and triplet loss, before finally reporting on the obtained Zero-Shot Classification results.

4.3.1 Surjectivity evaluation

Here, we show the influence of using the surjectivity loss as an additional regularization term in our cost function. In Figure 3 we perform a parameter sweep across the weight coefficient of the surjection loss and report the test accuracy it leads to. We can see a slight improvement in accuracy when choosing the ideal coefficent, which portrays the positive impact it has in the ZSL problem, roughly corresponding to a 3-5% improvement. As expected, a too high parameter leads to poor performance due to high bias.

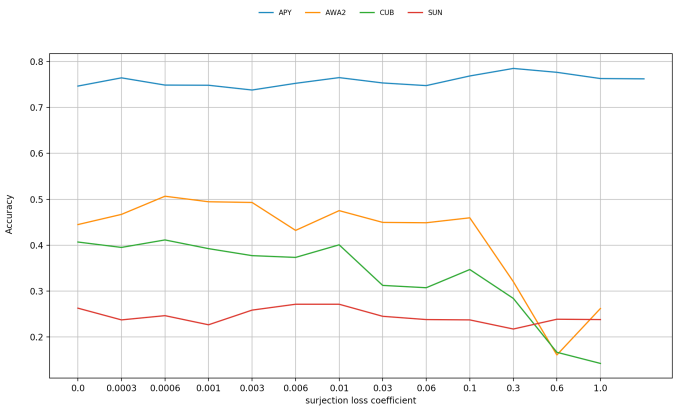

Figure 3. Evaluation of surjectivity regularization

4.3.2 Triplet Loss

Since we deviate from the conventional method of using the L2 Loss for training, we assess the significance of using the Triplet Loss on our performance. We tune the hyperparameters to the ideal case for both Triplet and L2 Loss and depict their influence on top-1 prediction accuracy in Figure 4. One can see that triplet loss is invariably better than L2 loss in all datasets, proving its advantage over L2 loss as we argued for nearest neighbour queries.

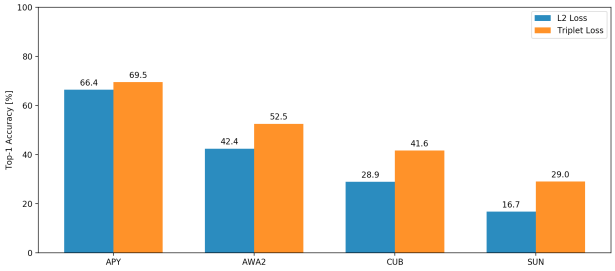

Figure 4. Test accuracy evaluation of triplet loss compared to L2 loss.

4.3.3 ZSL Evaluations

We test with the proposed split from [25] on aPY, AwA2, CUB and SUN datasets. The results in this section are using triplet loss and NN search in the semantic space. In Table 1 we present classification results in the conventional ZSL case, *i.e.* we evaluate the performance only on unseen classes, when using both the word embeddings of the class label obtained from GloVe and the class attributes provided in the datasets. We don't evaluate SAE since it doesn't support joint semantic embeddings. Our method clearly works better than linear projections. In Table 2 we show a similar scenario, but solely using the word embeddings as additional information source. We see better performance than both the baselines in majority of the datasets. Similarly, Table 3 presents the results only using attributes. Here the performance is better for AwA, CUB and SUN because they are attribute-based datasets, hence attributes as semantic vectors are more effective than the word embeddings. In APY we believe surjective loss works especially well, since the classes are rather simple names (commonly occurring in the English language) and distinguishable. Thus surjection naturally leads to better generalization, whereas in other datasets intertwined embeddings do not combine that well with surjective loss, although still we get better performance.

	Datasets			
	aPY [5]	**AwA2** [25]	**CUB** [24]	**SUN** [26]
Linear	29.0	51.4	25.0	10.5
SAE [13]	-	-	-	-
Ours	**69.5**	**52.5**	**41.6**	**29.05**

Table 1. Zero Shot Top-1 Accuracy (ZSL) using **attributes on top of word vectors**. Note that SAE can not be evaluated in this setting due to model restrictions.

4.3.4 ZSL with Visual Space Queries

Finally, we compare performing nearest neighbour queries in visual space and semantic space (both using triplet loss). One can

	Datasets			
	aPY [5]	**AwA2** [25]	**CUB** [24]	**SUN** [26]
Linear	46.8	20.4	9.5	7.5
SAE [13]	9.26	**37.67**	3.94	14.1
Ours	**78.2**	27.3	**12.2**	**20.67**

Table 2. Zero Shot Top-1 Accuracy (ZSL) using **only word vectors** of the class label.

	Datasets			
	aPY [5]	**AwA2** [25]	**CUB** [24]	**SUN** [26]
Linear	38.4	47.6	26.5	26.5
SAE [13]	5.5	48.88	21.44	33.68
Ours	**65.3**	48.4	36.6	44.2
MLSE [3]	46.2	**67.8**	**64.2**	**62.8**

Table 3. Zero Shot Top-1 Accuracy (ZSL) using **only attributes** of the class label. MLSE [3] is the current SOTA in conventional ZSL.

see in Table 4 that semantic queries almost always work better. This could be due to the fact that varying visual features can correspond to the same set of semantic features, thus queries through the many-to-one function f_θ in semantic space distinguishes classes better. One explanation of the better performance of queries in visual space for SUN dataset can be the large number of classes (717) as compared to the rest three datasets (50, 64 and 200 classes).

	Datasets			
Query	**aPY** [5]	**AwA2** [25]	**CUB** [24]	**SUN** [26]
Visual	54.9	41.4	34.2	**53.52**
Semantic	**69.5**	**52.5**	**41.6**	29.05

Table 4. Zero Shot Top-1 Accuracy (ZSL) using **attributes on top of word vectors** of the class label for queries performed on two spaces.

4.3.5 Generalized ZSL

For the purpose of experimenting with how well our model could perform in the so-called "generalized ZSL" (GZSL) scenario, we conducted additional tests on the APY dataset. The GZSL task is typically much more challenging due to the fact that there is a severe sample imbalance between the classes tested. Therefore, GZSL algorithms usually overfit to only predict classes they have seen before, whereas we expected the surjectivity loss to improve this. To keep the report compact, we do not report detailed quantitative evaluations. We found the accuracy of our model to be 56.2% on the GZSL task using word embeddings only, compared to 13.7% for SAE. Moreover, the confusion matrix (Figure 5) indicates that the surjectivity constraint actually forces the model to predict unseen labels. Without the surjectivity constraint, the cyan box is almost empty (see Appendix). One can also observe expected patterns such as confusing goats (seen) with zebras (unseen) and boats (seen) with jet skis (unseen) which are due to the underlying difficulty of the dataset.

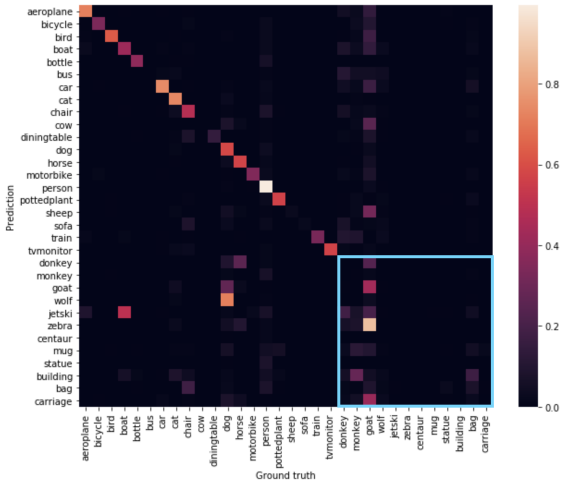

Figure 5. The confusion matrix evaluated on the APY dataset. The first 20 classes are seen at test time. 12 classes (outlined with a cyan box) are not included in the training set.

5. Discussion and Conclusion

In our experiments we validated that our proposed methods actually improved performance in ZSL and performed detailed ablation studies of various components of our approach. Almost all our claims were verified: surjection loss improved predictive accuracy and generalization to unseen classes, and triplet loss significantly increased classification performance compared to L2 loss. In fact, our models can compete with the algorithms using hand-annotated visual attributes by using only word embeddings from NLP models. This is a major advantage of our approach: annotation is expensive, whereas there are many general purpose word embeddings available freely. Finally even though we use two networks during training, the inverse network exists only to ensure surjective training and can be discarded afterwards. Thus our method introduces no computational cost at test time.

Moreover, we observed that the SAE baseline results are heavily dependent on the choice of hyper-parameters and the train-val split, leading to major discrepancies between their reported performance and our reproduced values. Their method is sensitive to choices like CNN model used and the low performance we obtain has also been replicated by other papers implementing SAE [2].

We believe our proposed method is a promising research direction, and leave the future development of the idea to future work. Namely, it could be interesting to use ideas from the generative adverserial networks literature for distribution matching to ensure a more general "almost everwhere" surjectivity. Moreover, there could be better ways to embed and query semantic vectors such as using hyperbolic spaces or the quadruplet loss function. Since our loss functions are universal, one could also attempt to apply them to the state-of-the-art to further increase performance.

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1425–1438, 2015.

[2] Yashas Annadani and Soma Biswas. Preserving semantic relations for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7603–7612, 2018.

[3] Zhengming Ding and Hongfu Liu. Marginalized latent semantic encoder for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6191–6199, 2019.

[4] Mohamed Elhoseiny, Babak Saleh, and Ahmed Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2584–2591, 2013.

[5] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1785. IEEE, 2009.

[6] Rafael Felix, Vijay BG Kumar, Ian Reid, and Gustavo Carneiro. Multi-modal cycle-consistent generalized zero-shot learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 21–37, 2018.

[7] Andrea Frome, Greg Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Neural Information Processing Systems (NIPS)*, 2013.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015.

[9] 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.

[10] Derek Hoiem, Santosh K Divvala, and James H Hays. Pascal voc 2008 challenge. In *PASCAL challenge workshop in ECCV*. Citeseer, 2009.

[11] Jörn-Henrik Jacobsen, Arnold Smeulders, and Edouard Oyallon. i-revnet: Deep invertible networks. *arXiv preprint arXiv:1802.07088*, 2018.

[12] Nour Karessli, Zeynep Akata, Bernt Schiele, and Andreas Bulling. Gaze embeddings for zero-shot image classification. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2017.

[13] E. Kodirov, T. Xiang, and S. Gong. Semantic autoencoder for zero-shot learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4447–4456, July 2017.

[14] 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.

[15] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[16] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 49–58, 2016.

[17] Bernardino Romera-Paredes and Philip H. S. Torr. An embarrassingly simple approach to zero-shot learning. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 2152–2161. JMLR.org, 2015.

[18] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1757–1772, 2012.

[19] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

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

[21] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943, 2013.

[22] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 935–943. Curran Associates, Inc., 2013.

[23] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[24] Catherine Wah, Steve Branson, Pietro Perona, and Serge Belongie. Multiclass recognition and part localization with humans in the loop. In *2011 International Conference on Computer Vision*, pages 2524–2531. IEEE, 2011.

[25] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning - A comprehensive evaluation of the good, the bad and the ugly. *CoRR*, abs/1707.00600, 2017.

[26] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492. IEEE, 2010.

[27] Hanwang Zhang, Xindi Shang, Wenzhuo Yang, Huan Xu, Huanbo Luan, and Tat-Seng Chua. Online collaborative learning for open-vocabulary visual classifiers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2809–2817, 2016.

[28] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1004–1013, 2018.

6. Appendix

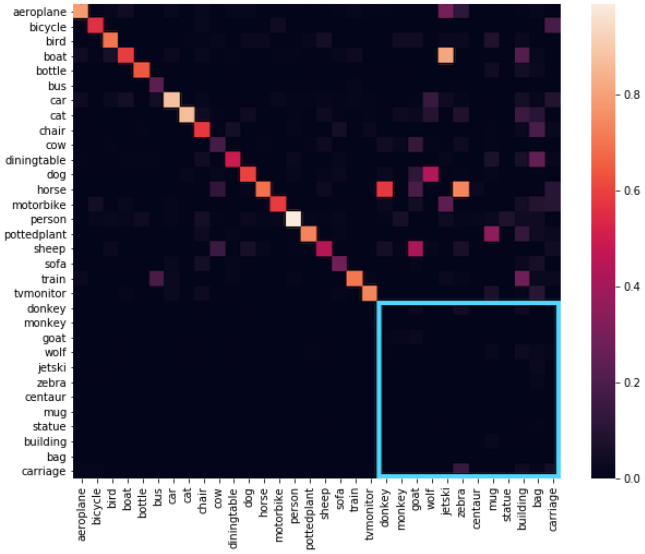

Figure 6. The confusion matrix evaluated on the APY dataset without the surjectivity constraint. The first 20 classes are seen at test time. 12 classes (outlined with a cyan box) are not included in the training set. Without surjectivity, the model completely fails to predict unseen classes.