

ZSpeedL - Evaluating the Performance of Zero-Shot Learning Methods using Low-Power Devices

Cristiano Patrício and João C. Neves

NOVA-LINCS

University of Beira Interior, Portugal

cristiano.patricio@ubi.pt, jcneves@di.ubi.pt

Abstract

The recognition of unseen objects from a semantic representation or textual description, usually denoted as zero-shot learning, is more prone to be used in real-world scenarios when compared to traditional object recognition. Nevertheless, no work has evaluated the feasibility of deploying zero-shot learning approaches in these scenarios, particularly when using low-power devices.

In this paper, we provide the first benchmark on the inference time of zero-shot learning, comprising an evaluation of state-of-the-art approaches regarding their speed/accuracy trade-off. An analysis to the processing time of the different phases of the ZSL inference stage reveals that visual feature extraction is the major bottleneck in this paradigm, but, we show that lightweight networks can dramatically reduce the overall inference time without reducing the accuracy obtained by the de facto ResNet101 architecture. Also, this benchmark evaluates how different ZSL approaches perform in low-power devices, and how the visual feature extraction phase could be optimized in this hardware.

To foster the research and deployment of ZSL systems capable of operating in real-world scenarios, we release the evaluation framework used in this benchmark (<https://github.com/CristianoPatricio/zsl-methods>).

1. Introduction

Object recognition approaches are restricted to the identification of a finite number of classes observed during training. On the contrary, humans are capable of identifying unknown objects simply using a textual description. Based on this observation, zero-shot learning (ZSL) emerged as a more realistic alternative to traditional object recognition, where the goal is to build computational models capable of identifying objects solely with a semantic description of the class, and without samples in the training phase.

In the last years, researchers have successfully exploited the advances on machine learning to boost the performance of this learning paradigm. At the same time, the effort put on creating standard evaluation protocols and benchmarks fostered the advances on the problem of ZSL. Thus, is not surprising that soon ZSL methods will be integrated into industrial solutions or end-user applications. Nevertheless, no work has evaluated the feasibility of deploying state-of-the-art ZSL methods in real-world scenarios, particularly the trade-off between accuracy and inference time.

In this paper, we introduce the first benchmark on the accuracy and inference time of ZSL methods, intended to study the feasibility of deploying these methods in real-world scenarios using low-power computational devices. This benchmark evaluates the impact of the two major phases of the processing chain of zero-shot learning, illustrated in Figure 1. The experimental results show that visual feature extraction is the bottleneck in the processing chain of these approaches, and thus this work is also concerned on assessing the impact that lighter CNN architectures have in the overall speed and accuracy of ZSL approaches. Also, our experiments show that is possible to reduce the size of the architecture used without significantly impacting the ZSL method's accuracy. This benchmark also reports the number of frames per second analyzed when using different low-power hardware devices, providing insights about the workability of these approaches in real-world scenarios. Finally, to foster the research on zero-shot learning and the development of ZSL-based solutions, we release the source-code for reproducing all the experiments performed, as well as the instructions for optimizing and deploying ZSL methods in low-power devices.

Accordingly, the major contributions of this paper are the following: (1) An extensive benchmark across four widely known ZSL datasets regarding the accuracy and processing time of state-of-the-art ZSL methods using visual features obtained from different CNN architectures. (2) A comparative analysis of the processing time of ZSL algorithms using lightweight architectures when run in different low-power

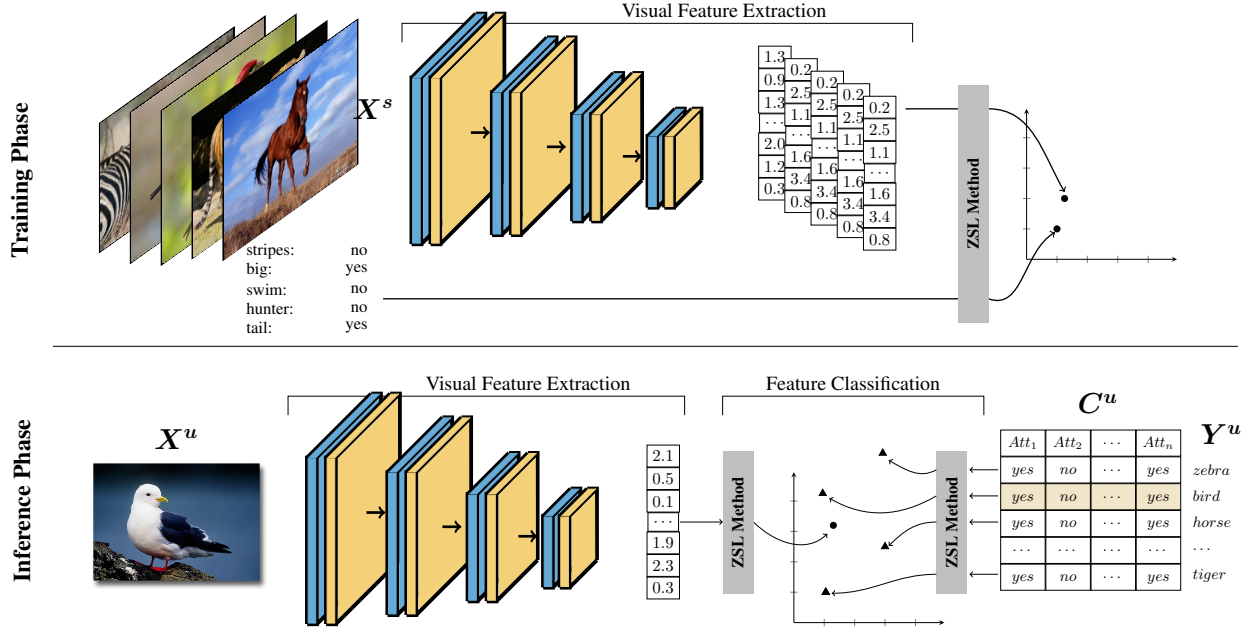


Figure 1. **Processing chain of zero-shot learning.** Visual feature extraction allows representing images through a compact descriptor. These visual features and its corresponding semantic embedding are fed, during the training phase, to ZSL methods for learning a function that reduces the distance between the two representations. In the inference phase, the visual features of the query image are transformed by the ZSL method, and the class of the images is determined by the closest semantic embedding of the test set.

devices. (3) An open-source evaluation framework for analyzing the accuracy/speed trade-off in the problem of ZSL.

The remainder of this paper is organized as follows: section 2 summarizes the most relevant works in the scope of our work. Section 3 provides the details of the experiments performed. In section 4 we discuss the obtained results and the conclusions are given in section 5.

2. Related Work

2.1. ZSL Problem

Zero-shot learning regards the problem of identifying classes not observed during the training phase. Formally, given a set of images $X = X^s \cup X^u$, and the corresponding labels of seen (s) and unseen (u) data, $Y = Y^s \cup Y^u$, the goal of ZSL is inferring $X^u \rightarrow Y^u$, without having access to Y^u during training. The only ancillary information provided is the semantic representation $C = C^s \cup C^u$ describing each class, which should be used as a proxy to map the visual data to the unseen classes. For inferring this mapping, most works follow one of two major strategies: 1) learning a projection function capable of relating both encodings in a common low-dimensional space; 2) generating the visual features from the semantic embedding for unseen classes and train a standard classifier to distinguish between classes. The work of Lampert *et al.* [1] was precursor of projection-based methods where a binary classi-

fier was trained to predict the attributes of an image given its visual features. In the test phase, an image was classified by measuring the probability of the predicted attributes match the semantic representation of unseen classes. Later, the idea of learning a linear compatibility matrix between visual features and semantic information was explored in [2]. The use of inner product similarity allowed to devise a simple system of linear operations capable of determining the similarity between the image features and unseen classes. This idea was explored in several works where different loss functions [3, 4] and regularization terms were proposed [5]. Also, several works improved the original idea by using non-linear compatibility functions [6, 7]. A different strategy was projecting both visual features and semantic embeddings into a common sub-space. Zhang and Saligrama [8] proposed the learning of a class-independent function to map both representations to a latent representations. Changpinyo *et al.* [9] relied on the notion of a weighted graph to align the semantic space with the model space, composed of classifiers for visual recognition, such that the coordinates of the model space are a projection of the graph vertices from the semantic space. Zhang *et al.* [10] proposed the first deep end-to-end ZSL model by training simultaneously the CNN used for visual feature extraction and a set of fully connected layers for transforming the semantic representation to the visual space. They found that this model coupled with the least square loss between

the two embedded vectors is less prone to suffer from the hubness problem.

In spite of the advances obtained with projection methods, they failed to achieve reasonable results in the generalized setting of zero-shot learning, where both seen and unseen classes are available in the test phase. This is explained by the bias of the mapping function towards seen classes, and to address this issue, Xian *et al.* [11] suggested the idea of generating the visual features of unseen classes conditioned on semantic information, and rely on a supervised classifier to distinguish between the target classes using the features extracted from the query image. Following this idea, several approaches have been proposed using different models for generating features [11], the use of cycle-consistency loss [12], or the use of feedback loops for refining generated features [13].

2.2. Visual Feature Extraction

ZSL methods rely on visual feature extraction to represent the images in a compact manner. While initial ZSL approaches used hand-crafted techniques for feature extraction, the use of Convolutional Neural Networks (CNN) has become the standard procedure for obtaining a compact description of a test image. Considering this, the research on ZSL has been primarily focused on devising strategies to learn the mapping from visual representations to semantic embeddings. As described in section 2.1, these strategies can be broadly divided into two major groups: 1) projection-based methods; 2) generative methods. The computational burden of each family is analyzed in Section 4.1.

2.3. ZSL Computational Cost

Most contributions on ZSL disregard the visual feature extraction phase by using precomputed features from standard CNN architectures. The reduced size of these features significantly decreases the processing time of the feature classification in the ZSL processing chain, even when the classification algorithm has a high temporal complexity. For this reason, most works have focused solely on reporting accuracy, and few works have analyzed inference time of the proposed models. In [14] and [15], the authors perform an analysis on the complexity of the learning strategy. Ji *et al.* [16] report the processing time during both the training and inference phase, and provide the complexity of the proposed algorithm. Pan *et al.* [17] evaluate the inference time of the proposed method both on CPU and GPU, as well as the time required by competing approaches. In other work [18], the authors report the processing time of the proposed method when using different datasets, as they noticed differences according to the number of unseen classes.

Nevertheless, no work has specifically evaluated the inference time of the overall processing chain of ZSL, nei-

ther the impact of using different architectures for the visual feature extraction phase. Also, the evaluation on low-power devices has not been considered yet. To the best of our knowledge, our work is the first benchmark on ZSL inference time in low-power devices, providing a comparative analysis of how different CNN architectures impact the speed/accuracy trade-off of these approaches.

3. Evaluation Methodology

This section details the experimental protocol, including a description of the testbed state-of-the-art ZSL methods, the statistics of used datasets, and the adopted evaluation protocols.

3.1. Methods

We have selected six state-of-the-art ZSL methods, including ESZSL [5], SAE [19], DEM [10], f-CLSWGAN [11], TF-VAEGAN [13], and CE-GSZL [20]. The selected approaches cover the two major strategies in ZSL: (1) projection-based methods (ESZSL [5], SAE [19], and DEM [10]), and (2) generative methods (f-CLSWGAN [11], TF-VAEGAN [13], and CE-GSZL [20]).

DAP [1] was a seminal work in the topic of ZSL, where a binary classifier was trained to predict the attributes of an image given its visual features. In the inference phase, an image is classified by measuring the probability of the predicted attributes match the semantic representation of unseen classes. Regarding its computational cost, this approach depends mainly on complexity of the classifier used to estimate the attributes of an image given its visual representation and the number of test classes.

ESZSL [5] is a representative approach of the use of inner product similarity for determining the similarity between the image features and image classes. Similarity scores for test classes are inferred through a set of linear transformations, encoded by matrix multiplication operations. The reduced complexity of these transformations, and the possibility of parallelizing matrix multiplication, ensures a short inference time.

SAE [19] treats ZSL as an encoding-decoding problem by using an auto-encoder (AE) to transform the visual features to the semantic space and subsequently recover the same features from the class semantic representation. This strategy significantly outperformed state-of-the-art, and allowed to classify an image either by inferring its semantic representation from the visual features, or transforming the semantic representation of target classes to the visual feature space and then use k-NN in the feature space. The use of a linear AE allows a very fast attribute estimation. However, the overall inference time is mainly dependent on the size of unseen classes due to the use of k-NN.

DEM [10] was one of the first end-to-end ZSL mod-

els where the CNN and the feature projection function were jointly optimized. This strategy significantly increases training time, but the classification inference phase only depends on the size and number of fully connected layers.

f-CLSWGAN [11] introduced the disruptive idea of casting ZSL as a traditional supervised classification problem, by training a GAN for generating the visual features of any class from their semantic representation, and use a standard supervised classifier to distinguish between the unseen classes. During the test phase, inference time is only dependent on the classifier used (softmax classifier in the original work).

TF-VAEGAN [13] constitutes an improvement over traditional generative approaches, since it increases the semantic consistency of generated features by enforcing the decoding of generated features to be similar to original semantic embeddings. This semantic embedding decoder is also used during the inference stage for providing the feature classifier with both extracted features and their decoded semantic embeddings, which slightly increases the inference time over f-CLSWGAN.

CE-ZSL [20] is a hybrid approach combining a projection model and a generative model. The projection model is used to estimate the semantic embeddings of generated visual features, in order to compare these embeddings with the original ones using comparator network trained using contrastive learning. Contrary to other generative approaches, CE-ZSL performs the classification in the semantic space, where a softmax classifier is trained using both seen semantic information and unseen synthetic embeddings. For this reason, as in the case of f-CLSWGAN, inference time depends solely on the classifier used.

3.2. Datasets

The evaluation of the six state-of-the-art ZSL methods considered in this study is carried out on the four most popular ZSL datasets, namely, *Animals with Attributes 2* (AWA2) [21], *Caltech-UCSD-Birds* (CUB-200-2011) [22], *SUN Attributes* [23] (SUN), and *Attribute Pascal and Yahoo* (aPascal-aYahoo) [24]. The datasets statistics are presented in Table 1.

Table 1. Statistics of the datasets used in this benchmark.

Dataset	No. Classes	No. Instances	No. Attributes
CUB-200-2011 [22]	200	11788	312
SUN Attributes [23]	717	14340	102
AWA2 [21]	50	37322	85
aPascal-aYahoo [24]	32	15339	64

The visual features for all the images in the dataset are extracted using the top-layer pooling units of the ResNet-101, MobileNet, MobileNetV2, Xception, and EfficientNetB7 pre-trained on ImageNet-1K without fine-tuning. Moreover, we adopt the Proposed Split (PS) [21] to ensure that test classes are disjoint from the ones used to train the

CNN model. For semantic embeddings, we use the class-level attributes provided by [21] for AWA2 (85-dim), SUN (102-dim), CUB (312-dim), and APY (64-dim). We conduct all experiments under the inductive setting, in which only labeled instances of seen classes are considered at the training phase [21].

3.3. Hardware

Experiments were performed in a desktop computer and two small low-power devices, namely a Raspberry Pi 4 Model B and a Jetson Nano Developer Kit. The hardware specifications are given in table 2.

Table 2. Hardware Specification.

	Desktop	Raspberry Pi 4B	Jetson Nano Dev Kit
CPU	Intel® Core™ i7-10700U CPU @ 2.90GHz × 16	Broadcom BCM2711 Quad core Cortex-A72 (ARM v8) 64-bit SoC @ 1.5GHz	Quad-core ARM A57 @ 1.43 GHz
GPU	-	-	128-core Maxwell
RAM	32GB	4GB	4 GB
Storage	1TB SSD	64GB microSD	64GB microSD
OS	Pop.OS! 20.10 64-bit	Raspbian	Ubuntu 18.04.5 LTS

3.4. Evaluation Protocols

The performance of ZSL methods in the restricted and generalized settings is measured using the standard evaluation protocol proposed in [21]. The multi-way classification accuracy (MCA) is adopted to assess the average per-class accuracy in the restricted setting. In the generalized setting, the average per-class classification accuracy is determined on training (Y^s) and test (Y^u) classes, and the harmonic mean is obtained by $H = \frac{2*acc_{Y^s}*acc_{Y^u}}{acc_{Y^s}+acc_{Y^u}}$, where acc_{Y^s} and acc_{Y^u} denotes the accuracy of seen and unseen classes, respectively.

4. Results

4.1. ZSL Methods: Inference Time

This section analyzes the time consumed by different ZSL methods for classifying a single test image, disregarding the visual feature extraction phase. This analysis intends to evaluate the impact of the different feature classification strategies in the overall inference stage of ZSL methods. Moreover, the impact of the size of visual features and the use of low-power devices in the processing time is also studied. The results obtained are reported in Table 3.

As expected, ZSL approaches are extremely fast when disregarding the visual feature extraction phase. The low-dimension of both visual features and semantic representation of the classes grants a reduced number of computations, even when the algorithm used has polynomial time. In general, ZSL methods are capable of classifying visual features of an image in less than 1ms when using CPU, and in less than 25ms when using low-power devices. Regarding the comparison between projection methods and generative methods, it is not possible to conclude which strategy

Table 3. Processing time (in milliseconds) of feature classification on CPU, Raspberry Pi 4B (R-PI 4B) and Jetson Nano.

Method	Visual Features Dimension											
	512			1,024			2,048			4,032		
	Desktop	R-PI 4B	Jetson Nano	Desktop	R-PI 4B	Jetson Nano	Desktop	R-PI 4B	Jetson Nano	Desktop	R-PI 4B	Jetson Nano
DAP	0.74±0.00	18.72±0.45	5.11±0.05	0.43±0.00	19.05±0.31	5.24±0.15	0.45±0.01	18.82±0.32	5.20±0.10	0.44±0.02	18.61±0.35	5.28±0.11
IAP	0.74±0.00	18.61±0.20	5.11±0.03	0.44±0.00	18.85±0.07	5.28±0.13	0.45±0.02	18.70±0.08	5.24±0.09	0.43±0.00	18.49±0.06	5.19±0.11
SAE	0.12±0.00	1.11±0.13	1.21±0.05	0.12±0.00	1.35±0.09	1.57±0.02	0.13±0.00	1.66±0.07	1.88±0.05	0.13±0.00	2.29±0.07	2.42±0.04
ESZSL	0.03±0.00	0.81±0.09	0.49±0.05	0.03±0.00	1.08±0.10	0.75±0.06	0.04±0.00	1.40±0.15	0.97±0.07	0.06±0.00	1.89±0.05	1.37±0.02
DEM	0.83±0.05	4.15±0.03	9.43±0.49	1.54±0.12	7.35±0.12	13.33±0.32	3.11±0.14	26.74±0.80	24.28±1.47	6.25±0.23	44.69±1.28	40.29±2.94
F-CLSWGAN	0.71±0.06	3.95±0.07	2.02±0.12	0.83±0.09	4.40±0.09	1.98±0.09	0.96±0.10	5.72±0.12	1.79±0.05	1.26±0.11	7.03±0.14	2.55±0.13
TF-VAEGAN	1.40±0.03	17.89±0.87	10.49±1.90	1.97±0.05	26.83±1.65	12.02±0.39	3.08±0.06	44.95±2.53	10.23±0.76	5.35±0.10	77.28±3.23	19.42±4.76
CE-GZSL	0.75±0.09	4.08±0.22	2.70±0.09	0.82±0.09	4.38±0.24	1.76±0.16	0.95±0.06	5.33±0.13	1.71±0.09	1.29±0.10	6.84±0.30	2.60±0.06

is faster. Instead, the difference lies in the type of models used, as the methods based on deep learning models have a higher processing time.

Table 4. Processing time of the visual feature extraction phase on CPU, Raspberry Pi 4B (R-PI 4B) and Jetson Nano. Execution time is presented in milliseconds with the format $avg \pm std$.

Architecture	Processing Time			Feat. Dim.	Size (MB)
	Desktop	R-PI 4B	Jetson Nano		
MobileNet	25.57±3.17	310.52±9.80	29.58±3.14	1024	16
MobileNetV2	27.59±3.38	297.63±8.54	33.04±20.11	1280	14
InceptionV3	33.80±2.81	609.23±3.54	143.28±29.19	2048	92
ResNet50V2	38.07±3.13	887.86±5.27	160.87±1.78	2048	98
NASNetMobile	39.67±2.35	370.10±5.79	111.81±20.40	1056	23
ResNet50	40.25±3.15	968.05±17.31	164.69±3.28	2048	98
Xception	43.43±3.29	1081.18±11.07	155.75±23.21	2048	88
ResNet101V2	54.13±3.09	1655.37±15.72	-	2048	171
DenseNet201	54.79±2.86	1404.16±39.72	-	1920	80
ResNet101	57.46±2.99	1639.46±76.71	-	2048	171
VGG16	59.73±4.10	2046.51±15.03	221.75±16.96	512	528
VGG19	69.15±2.42	2557.60±6.12	-	512	549
EfficientNetB7	86.54±0.85	2436.62±106.28	-	2560	256
NASNetLarge	95.67±2.98	2115.31±16.84	-	4032	343

4.2. Visual Feature Extraction: Inference Time

Considering that, in general, ZSL approaches are extremely fast during the inference stage (refer to section 4.1), this section is devoted to the analysis of time consumed by the visual feature extraction phase, which is usually performed using CNNs. For this, we evaluate several CNN architectures with respect to the time consumed for obtaining the features of the top-layer pooling units, and using different hardware devices. The results are reported in Table 4.

The results show that visual feature extraction is significantly slower when compared with the inference time of ZSL methods, being the bottleneck in overall inference stage of ZSL. Also, the results evidence that the consumed time varies largely over different architectures, and only lightweight models are capable of providing an acceptable running time in a low-power computational devices such as Raspberry Pi. However, these lightweight architectures are hardly used in ZSL applications, since the vast majority of the works adopt the ResNet101 architecture as the *de facto* model for benchmarking the accuracy of ZSL methods. For this reason, it is particularly important to understand the impact of using other architectures in the accuracy of these methods, providing additional insights about their speed/accuracy trade-off regarding the architecture used for feature extraction.

4.3. ZSL Accuracy: Impact of CNN Complexity

Based on the observation that lightweight CNNs are capable of reducing significantly the time required to extract visual features, we argue that is particularly important to assess how the use of features obtained from lightweight models impacts the accuracy of ZSL methods. For this, the ZSL methods considered in this study are re-trained using the features extracted from networks with varying complexity. To ensure a fair evaluation, the hyper-parameters of each approach are adjusted during training using a validation set. Finally, each method is evaluated in the four datasets considered in this benchmark (AwA, CUB, APY and SUN) under the restricted and generalized (GZSL) settings. The results are depicted in Figure 2, and evidence that the features generated by lightweight architectures allow ZSL approaches to attain competitive results when compared with the *de facto* ResNet101 architecture, typically used for assessing the accuracy of ZSL methods. Also, it can be observed that, in general, lines are nearly horizontal, meaning that the performance of ZSL methods does not consistently improve with the complexity of the architecture used for feature extraction. These results suggest that the inference phase of ZSL methods can be speed up without compromising the accuracy of ZSL.

To provide a thorough evaluation of ZSL methods using different architectures, we also report in Table 5 the performance of each method using the standard metrics of ZSL, i.e., multi-way classification accuracy (MCA) in the restricted setting, and in the generalized setting the accuracy of the seen (S) and unseen (U) classes, as well as their harmonic mean (H).

4.4. ZSL Speed/Accuracy Trade-off

While section 4.1 analyzed the computational cost of the different phases of ZSL inference stage separately, it is particularly important to perceive how the processing time of the overall inference stage and the accuracy of ZSL methods is impacted by the use of different CNN architectures and the use of low-computational devices. Accordingly, we evaluate the performance of ZSL methods according to the MCA using four datasets commonly used in the field of ZSL (AwA, CUB, APY and SUN) under the restricted setting. The overall inference time

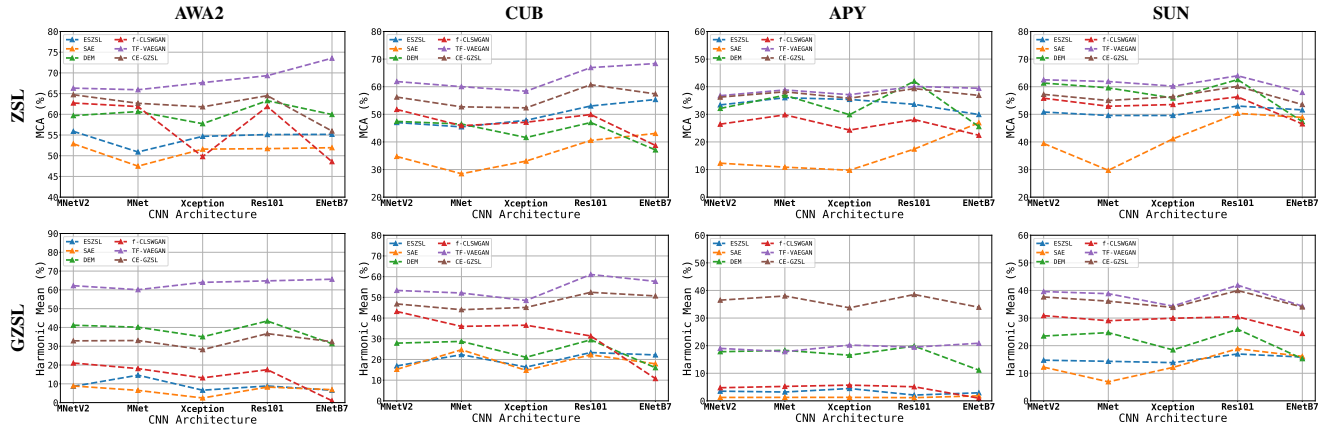


Figure 2. **Performance of ZSL methods with respect to the CNN architecture used.** Six state-of-the-art ZSL approaches were trained using visual features obtained from CNN architectures with varying complexity (networks complexity increases from left to right in x-axis). The multi-way classification accuracy (MCA) of the evaluated approaches in four ZSL datasets evidences that lightweight architectures do not significantly reduce the performance of ZSL approaches.

of ZSL methods is measured according to the frames per second (FPS) processed when using low-power computational devices. Also, to ensure a fair and comprehensive evaluation of the performance of these devices, the complexity of CNN models is reduced by using different network optimization techniques, such as layer fusion, matrix normalization, and the reduction of floating point precision (FP16/INT8). These optimizations are carried out using the NVIDIA TensorRT, allowing the creation of models compatible with the integrated GPU of Jetson Nano. Figure 3 depicts the results obtained, organized by dataset (columns) and the device used for inference (rows). As expected, lightweight networks significantly decrease the processing time of ZSL inference phase. However, the throughput of these networks does not exceed 4 FPS when using Raspberry Pi 4B, decreasing its applicability in real-world scenarios. In contrast, Jetson Nano is capable of delivering 30 FPS using lightweight networks, which can be explained by the optimizations performed using TensorRT and the use of an integrated GPU. Regarding the comparison between lightweight architectures and the ResNet101, the top-1 accuracy decreases 4% in average considering the four evaluated datasets, while the number of FPS increases from 0.6 to 3.3 in Raspberry Pi 4B, and 6.2 to 30.6 in Jetson Nano. This suggests that lightweight networks offer a compelling trade-off between inference time and accuracy, when compared to the *de facto* architecture used for benchmarking ZSL accuracy.

5. Conclusions

This work introduces the first benchmark on zero-shot learning regarding the processing time in the inference stage. The two major phases of the inference stage (vi-

sual feature extraction and feature classification) of zero-shot learning approaches are analyzed with respect to the impact in the overall inference time. The obtained results confirm that visual feature extraction is the major bottleneck in the pipeline of ZSL approaches, justifying thus the need for studying the impact of using different CNN architectures in the processing time of this phase, as well as in the speed/accuracy trade-off of ZSL methods. Accordingly, we measure the accuracy and processing speed of different ZSL methods when using architectures of varying complexity. Also, the evaluation was carried out using different hardware devices, to assess the feasibility of deploying ZSL methods in low-power computational devices. The results suggest that the use of lightweight architectures does not significantly decrease the accuracy of ZSL methods, while reducing dramatically inference time. Moreover, the analysis to processing time of low-power devices shows that standard single-board computers are not sufficient to operate in real-world scenarios (maximum of 4 FPS), but the use of specialized hardware, such as integrated GPUs, in this type of low-power devices can significantly reduce the processing time of visual feature extraction, enabling to perform the inference at 30 FPS. Finally, we make publicly available a framework for simultaneously evaluating the accuracy and inference time of ZSL methods, which we hope to foster the advances on the topic of ZSL.

Acknowledgements

The authors would like to thank the support provided by FCT strategic project NOVA LINC (UIDB/04516/2020).

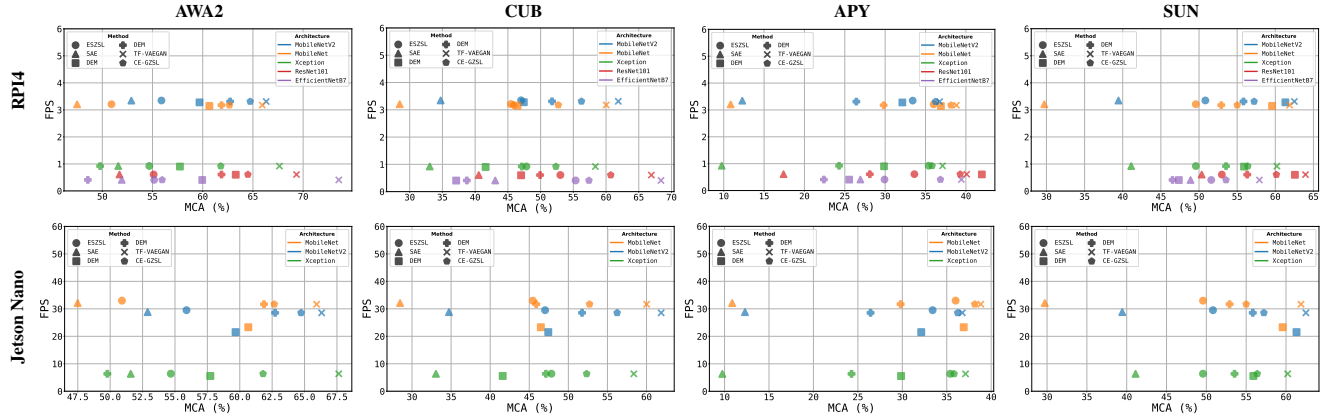


Figure 3. Accuracy/Speed trade-off of ZSL methods with respect to the CNN architecture used. The accuracy and the overall inference time of ZSL methods is evaluated with respect to the architecture used for feature extraction. While lightweight networks are clearly faster than remaining architectures, it is interesting to note that the accuracy does not significantly decrease. Also, the FPS obtained using Jetson Nano evidence that ZSL methods can be seamlessly deployed in low-power computational devices if lightweight networks are used and the device is equipped with specialized hardware (refer to 4.4).

Table 5. Accuracy of ZSL methods evaluated under the restricted (ZSL) and generalized (GZSL) setting on the AWA2, CUB, SUN and APY datasets using different CNN architectures for the visual feature extraction.

Methods	Architecture	AWA2				CUB				SUN				APY			
		ZSL		GZSL		ZSL		GZSL		ZSL		GZSL		ZSL		GZSL	
		MCA	U	S	H	MCA	U	S	H	MCA	U	S	H	MCA	U	S	H
ESZSL [5]	ResNet101	55.11	4.66	87.07	8.86	53.02	14.29	63.73	23.35	52.99	12.15	28.22	16.99	33.62	1.07	72.24	2.11
	MobileNet	50.91	8.04	79.32	14.60	45.45	14.22	52.13	22.34	49.58	10.49	22.64	14.33	36.00	1.65	66.23	3.21
	MobileNetV2	55.89	4.56	83.94	8.65	47.02	9.96	55.71	16.89	50.83	10.76	23.37	14.74	33.41	1.81	68.91	3.52
	Xception	54.68	3.44	86.67	6.61	47.83	9.55	56.23	16.33	49.58	9.44	25.93	13.85	35.39	2.32	69.75	4.48
	EfficientNetB7	55.16	3.42	87.41	6.57	55.35	13.45	63.37	22.19	51.60	11.25	26.98	15.88	29.92	1.50	69.62	2.93
SAE [19]	ResNet101	51.71	4.34	85.39	8.26	40.55	14.10	52.55	22.24	50.35	15.97	23.10	18.89	17.41	0.6	18.02	1.16
	MobileNet	47.49	3.40	77.62	6.51	28.47	15.87	55.74	24.70	29.72	5.69	8.76	6.90	10.86	0.69	8.32	1.27
	MobileNetV2	52.89	4.66	87.07	8.86	34.73	9.21	43.77	15.22	39.44	9.93	15.89	12.22	12.3	0.71	5.25	1.25
	Xception	51.59	1.25	87.64	2.47	33.06	8.84	44.32	14.74	41.11	9.65	16.32	12.13	9.75	0.71	6.68	1.28
	EfficientNetB7	51.94	3.59	85.70	6.89	43.07	10.59	57.61	17.90	48.89	12.71	22.05	16.12	26.93	0.93	53.70	1.83
DEM [10]	ResNet101	63.29	29.21	84.60	43.42	47.02	21.56	46.66	29.49	62.57	20.0	36.94	25.95	41.98	11.54	71.28	19.86
	MobileNet	60.66	26.59	81.87	40.14	46.48	20.21	49.61	28.72	59.58	18.61	36.78	24.72	36.91	10.59	67.87	18.31
	MobileNetV2	59.68	27.25	84.37	41.19	47.43	20.30	44.59	27.89	61.32	18.26	32.91	23.49	32.13	10.27	66.71	17.80
	Xception	57.73	21.98	86.09	35.02	41.6	15.78	31.66	21.06	55.9	13.89	27.52	18.46	29.86	9.61	59.53	16.54
	EfficientNetB7	59.95	19.14	87.04	31.38	37.1	11.49	26.41	16.02	47.36	12.15	20.58	15.28	25.54	6.37	44.10	11.14
f-CLSWGAN [11]	ResNet101	61.87	9.69	90.89	17.52	49.92	20.93	62.45	31.35	56.32	28.47	32.71	30.46	28.1	2.64	76.92	5.11
	MobileNet	61.87	10.07	89.46	18.10	45.89	27.13	53.39	35.98	52.92	33.13	25.85	29.04	29.81	2.71	76.14	5.24
	MobileNetV2	62.73	11.94	89.73	21.07	51.75	38.46	49.22	43.18	55.83	36.81	26.59	30.87	26.43	2.45	78.61	4.76
	Xception	49.78	7.11	89.96	13.17	47.13	27.78	53.39	36.54	53.54	29.03	30.89	29.93	24.29	2.99	64.41	5.71
	EfficientNetB7	48.56	0.50	90.12	1.0	38.73	5.96	55.39	10.76	46.53	25.69	23.26	24.41	22.41	0.5	75.82	0.98
TF-VAEGAN [13]	ResNet101	69.34	57.60	74.04	64.79	66.92	57.91	64.52	61.04	63.96	46.60	38.10	41.92	40.1	11.15	76.54	19.46
	MobileNet	65.93	52.61	70.06	60.09	60.02	48.14	56.76	52.10	61.88	45.63	33.80	38.83	38.83	10.24	68.57	17.82
	MobileNetV2	66.32	53.33	74.65	62.22	61.87	50.55	56.43	53.33	62.5	46.88	34.26	39.59	36.75	10.99	70.28	19.0
	Xception	67.64	53.90	78.76	64.00	58.37	44.52	53.24	48.49	60.21	39.51	30.35	34.33	37.11	11.81	72.20	20.21
	EfficientNetB7	73.55	54.92	81.62	65.66	68.39	51.43	65.77	57.72	57.92	36.04	32.75	34.32	39.44	12.05	78.25	20.88
CE-GZSL [20]	ResNet101	64.5	23.03	90.46	36.71	60.72	56.75	48.72	52.43	60.14	45.14	35.85	39.96	39.3	31.07	50.69	38.53
	MobileNet	62.66	20.39	86.98	33.04	52.71	39.89	49.17	44.04	55.0	40.14	32.83	36.12	38.15	29.40	53.67	37.99
	MobileNetV2	64.73	20.18	88.56	32.87	56.24	48.37	45.38	46.83	57.22	44.10	32.83	37.64	36.23	30.82	44.72	36.49
	Xception	61.8	16.71	89.80	28.18	52.34	39.22	53.14	45.13	56.39	38.33	30.31	33.85	35.82	11.81	72.20	20.21
	EfficientNetB7	55.96	19.76	90.35	32.43	57.39	50.24	51.14	50.68	53.54	37.15	31.36	34.01	36.85	25.87	49.19	33.91

References

- [1] C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 951-958, 2009. 2, 3
- [2] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 819-826, 2013. 2
- [3] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. In Advances in Neural Information Processing Systems (NIPS). Devise: A deep visual-semantic embedding model, pp. 2121-2129, 2013. 2
- [4] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained im-

- age classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2927-2936, 2015. 2
- [5] B. Romera-Paredes, and P. Torr. An embarrassingly simple approach to zero-shot learning. In Proceedings of the International Conference on Machine Learning (ICML), pp. 2152-2161, 2015. 2, 3, 7
- [6] R. Socher, M. Ganjoo, H. Sridhar, O. Bastani, C. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In Advances in Neural Information Processing Systems (NIPS), pp. 935-943, 2013. 2
- [7] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele. Latent embeddings for zero-shot classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 69-77, 2016. 2
- [8] Z. Zhang and V. Saligrama. Zero-shot learning via joint latent similarity embedding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6034-6042, 2016. 2
- [9] S. Changpinyo, L. Chao, B. Gong, and F. Sha. Synthesized classifiers for zero-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5327-5336, 2016. 2
- [10] L. Zhang, T. Xiang, and S. Gong. Learning a Deep Embedding Model for Zero-Shot Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2021-2030, 2017. 2, 3, 7
- [11] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata. Feature Generating Networks for Zero-Shot Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5542-5551, 2018. 3, 4, 7
- [12] Y. Xian, S. Sharma, B. Schiele, and Z. Akata. fvaegan-d2: A feature generating framework for any-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10275-10284, 2019. 3
- [13] S. Narayan, A. Gupta, F. Khan, C. Snoek, and L. Shao. Latent Embedding Feedback and Discriminative Features for Zero-Shot Classification. In Proceedings of the European Conference on Computer Vision (ECCV), 2020. 3, 4, 7
- [14] Q. Wang, W. Wu, Y. Zhao, and Y. Zhuang. Graph active learning for GCN-based zero-shot classification. Neurocomputing, vol 435, pp. 15-25, 2021. 3
- [15] Y. Liu, X. Gao, Q. Gao, J. Han, and L. Shao. Label-activating framework for zero-shot learning. Neural Networks, vol. 121, pp. 1-9, 2021. 3
- [16] Z. Ji, Y. Yu, Y. Pang, J. Guo, and Z. Zhang. Manifold regularized cross-modal embedding for zero-shot learning. Information Sciences, vol. 378, pp. 48-58, 2017. 3
- [17] C. Pan, J. Huang, J. Hao, and J. Gong. Towards zero-shot learning generalization via a cosine distance loss. Neurocomputing, vol. 381, pp. 167-176, 2020. 3
- [18] X. Li, M. Fang, H. Li, and J. Wu. Zero shot learning based on class visual prototypes and semantic consistency. Pattern Recognition Letters, vol. 135, pp. 368-374, 2020. 3
- [19] E. Kodirov, T. Xiang, and S. Gong. Semantic Autoencoder for Zero-Shot Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3174-3183, 2017. 3, 7
- [20] Z. Han, Z. Fu, S. Chen, and J. Yang. Contrastive Embedding for Generalized Zero-Shot Learning. arXiv preprint arXiv:2103.16173, 2021. 3, 4, 7
- [21] Y. Xian, C. Lampert, B. Schiele, and Z. Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. IEEE transactions on pattern analysis and machine intelligence (PAMI), vol. 41(9), pp. 2251-2265, 2018. 4
- [22] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset, 2011. 4
- [23] G. Patterson, and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2751-2758, 2012. 4
- [24] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1778-1785, 2009. 4
- [25] B. Zhao, Y. Fu, R. Liang, J. Wu, Y. Wang, and Y. Wang. A large-scale attribute dataset for zero-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR), pp. 0-0, 2019.