1. Introduction

**Student Number: 181345**

The classic paradigm of object recognition requires well-labelled training data to be learned in order to identify an image within that category. However, as the number of object categories continuously increases, it is nearly impossible to manually label all the data. “Humans can distinguish 30,000 basic object classes” [1] and more sub-classes. The conventional supervised learning pipeline consists of data residing in a feature space; once the class labels are annotated for this data, it is possible to distinguish between the different classes. However, if unseen categories are fed into the feature space, the classifier would not be able to classify the unseen data as it does not possess such model to do so. In contrast to the conventional supervised learning pipeline, zero-shot learning (ZSL) aims to recognize “novel visual categories without labelled training samples” [2].

To undertake this problem, 85 binary attributes are introduced, where objects are categorized based upon a high-level description of not only the visible features but also “properties that are not directly visible but related to visual information, such as animal habitat” [3].

These binary attributes will serve as the semantic properties to classify objects using a class description, which is also known as attribute-based classification.

Furthermore, the Animal with Attribute 2 dataset was used for the zero-shot learning system where 40 animal classes were labelled as the training data and the 10 remaining as the testing data. Therefore, the goal was to train the attribute classifiers to predict the unseen test images utilizing different image feature extraction methods.

## 2. Methods

Two images feature extraction methods were used in this investigation: KAZE features and a pre-trained convolutional neural network (CNN) dubbed AlexNet at the ‘fc7’ layer.

2.1. KAZE

KAZE features represent a novel “multiscale 2D feature detection and description algorithm in nonlinear scale spaces” [4]. It has been first introduced as an alternative to SIFT (scale-invariant feature transform) to solve the issue of the inherent “reduction in localization accuracy” present in SIFT. SIFT detects and describes features by approximating the Gaussian scale space of an image: the advantage of selecting rougher scales result in the reduction of noise and “the emphasis of more prominent structures” [5]. Although being the simplest option to build a scale space representation of an image, gaussian blurring used in SIFT would not disregard or miss the natural boundaries present in the object and smoothes it to the same degree. In contrast, KAZE utilizes a nonlinear diffusion filter to find the most dominant orientations present in an image. Given the most dominant orientation, each sample is weighted with a Gaussian centered at the interest point.

2.2. AlexNet (‘fc7)

AlexNet is CNN that is trained on “more than a million images from the ImageNet database [6]”. A pretrained AlexNet was also used to extract the features of the images with the Animal with Attributes 2 (AwA2) dataset. The ‘fc7’ later represents the layer before the classification layer within the CNN, it can used as a “generic image descriptor for recognition tasks [7]” in this case, for attribute classification.

3. Procedure

In order to create a complete direct attribute prediction model, all the images labeled as training were used to extract the features. Once the features were extracted, 85 linear SVC classifiers were trained with the extracted features, these classifiers would then be used to compute the probability of the attribute being present in each test image and the probability of the test class being present in the individual test image. For each feature extraction method, the outputs differed: the Kaze feature vector had 2048 dimensions whereas the CNN features with a pretrained AlexNet at the ‘fc7’ layer has 4096. However, the manner in which the probabilities were computed to achieved each method’s accuracies were identical.

4. Results

Upon classifying the unseen test images using each classifier the overall mean accuracy of the classifier trained with AlexNet features yielded 33.26% accuracy. In contrast to the model trained with the KAZE features yielded a mean accuracy of 10.38%, the accuracy results are shown in figure1. An conventional train/test split was used for cross validation for the training of the classifiers.

Before training, it was expected that the model trained with the KAZE features would perform poorly compared to the the model trained with pretrained CNN features. Given that AlexNet outperformed any other model during ImageNet competition [8], this may have contributed to the high accuracy of the CNN model. In addition, although the entire training image dataset was used for training, the images were down sampled by 50% reduce the memory size of images and certain features of the images extracted had to be removed due to the wrong dimensionality cases. The latter would not have been a direct contribution to the accuracy of the KAZE trained model.



Figure 1. Overall accuracy of the models trained with CNN features and Kaze features. ‘ts’ represents the test-size parameter used for cross validation split.

Furthermore, even when comparing the individual attribute classifier accuracies, the model trained using the CNN features yielded a greater accuracy for the majority of the classifiers. For instance, as represented in figure 2, the CNN feature trained model yielded a higher accuracy overall when classifying the individual attribute quadrupedal. This specific attribute was chosen for comparison as it was one of the attributes with the most positive values within the predicate matrix.



Figure 2. Individual attribute accuracies of “quadrupedal” between CNN feature trained model and KAZE feature trained model.

## 5. Discussion

Although the models did not yield the highest accuracies from classification due to potential reasons outlined earlier, it is important to discuss the nature of the dataset. The given Animal with Attribute 2 dataset contains images with different dimensions alongside an imbalanced number of image classes. Furthermore, given the 85 attributes to classify each class trait, some attributes might have been ambiguous and not specific which is paramount in the attribute-based classification.

By exploiting the semantic representation shared between a labelled supplementary dataset and a target dataset with different classes with no annotation, ZSL is able to classify novel objects without existing training data. However, there might exist certain inherent issues underlying the nature of ZSL. First introduced by Hospedales et Al. the *projection domain shift problem* could contribute to a flaw within ZSL. Due to having “disjoint and potentially unrelated classes” [9] when deploying the trained model to the unseen dataset, the model could be biased towards the target class.

To overcome this problem, Hospedales et Al present a “transductive multi-view embedding” framework which encompasses a multi-layer classification using not only attributes, but also word vectors and multiple features at their best performance.

## 5. Further workings

## The accuracy of the ZSL model implanted could have improved if the both CNN and KAZE features were combined to train the model. Each feature would classify the best properties and combine into an ensemble model to produce the most robust model.

# References

1. Transductive Multi-View Zero-Shot Learning - IEEE Journals & Magazine, ieeexplore.ieee.org/document/7053935/.
2. Transductive Multi-View Zero-Shot Learning - IEEE Journals & Magazine, ieeexplore.ieee.org/document/7053935/.
3. Lampert, Christoph H., et al. “Attribute-Based Classification for Zero-Shot Visual Object Categorization.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, 2014, pp. 453–465., doi:10.1109/tpami.2013.140.
4. Alcantarilla, Pablo Fernández, et al. “KAZE Features.” *Computer Vision – ECCV 2012 Lecture Notes in Computer Science*, 2012, pp. 214–227., doi:10.1007/978-3-642-33783-3\_16.
5. Alcantarilla, Pablo Fernández, et al. “KAZE Features.” *Computer Vision – ECCV 2012 Lecture Notes in Computer Science*, 2012, pp. 214–227., doi:10.1007/978-3-642-33783-3\_16.
6. “Deep Network Designer.” Pretrained AlexNet Convolutional Neural Network - MATLAB Alexnet - MathWorks United Kingdom, uk.mathworks.com/help/deeplearning/ref/alexnet.html.
7. “Deep Network Designer.” Pretrained AlexNet Convolutional Neural Network - MATLAB Alexnet - MathWorks United Kingdom, uk.mathworks.com/help/deeplearning/ref/alexnet.html.
8. “Deep Network Designer.” Pretrained AlexNet Convolutional Neural Network - MATLAB Alexnet - MathWorks United Kingdom, uk.mathworks.com/help/deeplearning/ref/alexnet.html.
9. Transductive Multi-View Zero-Shot Learning - IEEE Journals & Magazine, ieeexplore.ieee.org/document/7053935/.