Journal of Universal Computer Science, vol. 31, no. 1 (2025), 93-110
submitted: 13/2/2024, accepted: 13/8/2024, appeared: 28/1/2025 CC BY 4.0

Zero-shot Learning for Subdiscrimination in Pre-trained Models

Francisco Dominguez-Mateos
(University King Juan Carlos, Madrid, Spain
https://orcid.org/0000-0003-0909-7585, francisco.dominguez@urjc.es)

Vincent O'Brien
(South East Technological University, Calow, Ireland
https://orcid.org/0000-0002-8209-6661, vincent.obrien@setu.ie)

James Garland
(South East Technological University, Calow, Ireland
https://orcid.org/0000-0002-8688-9407, james.garland@setu.ie)

Ryan Furlong
(South East Technological University, Calow, Ireland
https://orcid.org/0000-0002-5072-7254, ryan.furlong@setu.ie)

Daniel Palacios-Alonso
(University King Juan Carlos, Madrid, Spain
https://orcid.org/0000-0001-6063-4898, daniel.palacios@urjc.es)

Abstract: In deep metric learning (DML) high-level input data are represented in a lower-level representation (embedding) space, such that samples from the same class are mapped close together, while samples from disparate classes are mapped further apart. In this lower-level representation, only a single inference sample from each known class is required to accurately discriminate between classes. To this end, embeddings trained for a specific task may contain additional feature information which can be used to go a level deeper into the discrimination task, i.e. allowing for feature sub-discrimination. This study takes an embedding trained to discriminate faces (identities) and uses the inherent feature information within the embedding to differentiate several attributes such as gender, age, and skin tone, without any additional training. This study is split into two cases; intra class discrimination where all the embeddings considered are from the same identity/individual but with minor attributes such as beard/beardless, glasses/without glasses and emotions; and extra class discrimination where the embeddings represent different identities/people with more prominent attributes such as male/female, pale/dark tone, young/older. In the intra class sub-discriminant scenario, the inference process distinguishes common attributes and several artefacts of different identities, achieving 90.0% and 76.0% accuracy for beards and glasses, respectively. The system can also perform extra class sub-discrimination with a high accuracy rate, notably 99.3%, 99.3% and 94.1% for gender, skin tone, and age, respectively. To sum up, this work investigates the sub-discriminative capabilities of DML models by clustering discriminative features evident within the structure of DML embeddings.

Keywords: Machine learning, Unsupervised learning, Deep metric learning, Zero-shot learning, One-shot learning, N-shot learning
Categories: H.3.1, H.3.2, H.3.3, H.3.7, H.5.1
DOI: 10.3897/jucs.120860

1 Introduction

Deep metric learning has proven to be highly adept at processing non-linear data, which in turn has led to the successful development of few-shot learning (FSL) [Wang et al., 2020], one-shot learning (OSL) [Santoro et al., 2016] and zero-shot learning (ZSL) [Wang et al., 2019] techniques. When applied to images, deep metric learning models first learn discriminative features from the images and then create embeddings representing the images. These embeddings are created such that when measured using a relatively simple Euclidean distance, the embeddings of similar images are closer, and the embeddings of dissimilar images are farther apart. In this paper, we want to answer the following questions:

- Do embeddings created by deep metric learning models exhibit *hierarchical coherence*?
- Can we use this *hierarchical coherence* to perform sub-discrimination?

Regarding the first question, we define the term *hierarchical coherence* as a relationship between embeddings that exhibits two properties. The first, coherence, is where the relationships between embeddings follow a logical order, and this order is consistent throughout the latent space. The second property is where the relationships are hierarchical, whereby certain image attributes have a greater impact on the embedding relationships than others. The second question asks if we can exploit the *hierarchical coherence* to perform sub-discrimination. Sub-discrimination is the ability to perform discrimination on attributes within and between classes; these attributes were not labelled during the training process, and the network was not trained to discriminate them. We consider this sub-discrimination a form of ZSL as the network was never trained to perform this type of attribute discrimination. To do this, we perform an experimental study where we take a pretrained network designed to discriminate faces and apply our sub-discrimination technique to differentiate between several attributes such as gender, age, and skin tone, without additional training. The attributes are split into extra and intra-class. *extra-class attributes* are defined as distinguishable facial features between different identities, e.g. age, gender, and skin tone. *intra-class attributes* are distinguishable within a single identity, namely the presence of beards or glasses and the expression of different moods and emotions, e.g. happy, angry, sad and neutral. This experimental study aims to provide insight into the *hierarchical coherence* within the embedding created for facial recognition and show that it can be used to discriminate additional attributes without the need for training.

The rest of the paper is organized as follows: Section 2 provides background material and discussions on related work, such as N-shot learning and deep metric learning. Section 3 describes the experimental set-up for performing discriminative studies of embeddings from high accuracy models. Section 4 details the results and an evaluation of the study. Section 5 describes the application of the results to perform reliable sub-discrimination on unseen images. Finally, Section 6 discusses future work and Section 7 provides a conclusion.

2 Background & Related Works

The field of machine learning has seen enormous growth in numerous application spaces due to the ability of machine learning models to outperform traditional approaches. The

success of deep learning in areas such as computer vision, natural language processing and speech recognition has been principally due to the ability to leverage large amounts of data to develop models capable of meaningful representation [Górriz et al., 2020]. However, accessing large high-quality datasets is not always possible, and the lack of such datasets can present a significant barrier to developing machine learning models. This issue is not an isolated case of deep learning; other research areas suffer from the same restriction [Tan et al., 2015, Hu et al., 2016, Hasrul et al., 2012]. With this in mind, methods aimed at solving the problems associated with limited data availability have emerged; namely, few-shot learning [Wang et al., 2020], one-shot learning [Santoro et al., 2016], and zero-shot learning [Wang et al., 2019]. These learning methods are commonly utilized in computer vision tasks, where employing an object categorization model still gives appropriate results even with zero, one or few training samples. These methods may use generative [Rezende et al., 2016, Lake et al., 2015] or discriminative models [Vanesa Sancho, 2011, Fan et al., 2014, Lin et al., 2015, Parkhi et al., 2015] to achieve their goal. This work studies one of the most common discriminative approaches, deep metric learning, where a model generates a vector in a low dimensional space with the help of a similarity metric.

DML proposes to train an artificial neural network based on a non-linear feature encoder that embeds the extracted features that are semantically similar and close to one another and maps dissimilar features further away from each other using an appropriate distance metric. DML improves on the linear constraints of metric learning approaches, whereby kernel tricks [Cortes and Vapnik, 1995] are required for the non-linear classification task. The DML approach uses the non-linear activation functions of neural networks to solve the problem of non-linearity. DML approaches have achieved exceptional results on various tasks ranging from face verification, and recognition to three-dimensional (3D) modelling, a broad review of these studies can be seen in [Kulis, 2012]. Specifically, with regards to image recognition, it can be seen that on several hard, fine-grained datasets [Russakovsky et al., 2015, Nilsback and Zisserman, 2008, Parkhi et al., 2012, Russakovsky and Fei-Fei, 2012] where the examples are difficult to distinguish, e.g. images of objects from the same class, deep metric learning classification can eclipse the state-of-the-art [Karlinsky et al., 2019].

The DML objective can be viewed as a solution to the one-shot learning problem, whereby one instance of a class is required to classify a new unknown instance from the same class. In this study, we manipulate the one-shot learning capabilities of embeddings generated by DML models to perform near zero-shot learning, where salient attribute information unknown during the training process is used to sub-discriminate known classes.

The *hierarchical coherence* investigated in this study is displayed in an early work by [Hadsell and LeCun, 2006]. Here, the authors demonstrate a mapping test of horizontally translated Modified National Institute of Standards and Technology (MNIST) [LeCun et al., 1998] digits to a 2D output manifold. This manifold exhibits clustering with respect to the translations despite the system not being trained to do so. The network is trained to map digits of the same class close to one another and digits of different classes further apart; however, the authors note that coherent clusters are formed with respect to the five translations, and each cluster is well organized for class discrimination. This clustering can be viewed as a result of the *hierarchical coherence* produced when the system learns a representation of the input data. More recently, ZSL approaches focus on creating novel classifiers that are generalised enough to accurately classify unseen classes from a large number of independent datasets. A recent approach known as a Simple framework for Contrastive Learning of visual Representations (SimCLR) [Chen

et al., 2020] displays the same coherence demonstrated by [Hadsell and LeCun, 2006]. In their work, the authors conduct the pretraining in an unsupervised manner, whereby the network is given no labeled instances of the training data. The network is trained to map an image close to an augmented version of the same image and further away from images that are not augmented versions of themselves. The SimCLR model achieves state-of-the-art classification results on the ImageNet dataset and even outperforms a strong supervised baseline despite the lack of class labels. The SimCLR model can discriminate between disparate classes due to the coherence displayed in the new latent space learned by the network during the training stage. & ddddd Another recent example, Contrastive Language-Image Pre-training (CLIP) [Radford et al., 2021], has shown remarkable zero-shot classification results on numerous datasets. During training, CLIP takes images scraped from the Internet and passes them through an image encoder. The text associated with each image, generally in the form of a caption, tag etc., is passed through a text encoder. The model is trained to map similar image embeddings and word embedding close to one another and dissimilar embeddings further apart. At inference, new unseen images are passed through the pretrained model and are represented as embeddings; labels are taken from the specific classification task and are also represented as embeddings. Zero-shot classification is achieved by using the label embeddings as prompts and classifying the images based on the prompt they lie closest to. CLIP represents the current state-of-the-art for zero-shot classification and even outperforms a fully supervised linear classifier. In contrast to the approach taken by [Radford et al., 2021], whereby the authors seek to create a universal zero-shot model, the work we propose focuses on examining the ZSL capabilities of pretrained DML models. Our method utilizes clustering to manipulate the *hierarchical coherence* present within the structure of embeddings generated by DML models. We restrict our study to a DML face recognition model. However, if all DML models exhibit similar *hierarchical coherence*, then our method can be tailored to specific applications by substituting the pretrained models.

In [Jain et al., 2018], the authors propose a facial clustering method, whereby the authors use various models such as ResNet 50 initialized with weights pre-trained on the ImageNet [Deng et al., 2009] dataset for feature extraction before scaling the data, performing dimensionality reduction, and finally clustering the data for the facial recognition task. Similarly, we take embeddings generated by a pre-trained facial recognition model but apply clustering to identify what facial attributes are represented within the embedding structure before using this inherent information to perform near zero-shot facial attribute classification. Consequently, the work we propose focuses on sub-discrimination (classification of facial attributes), not discrimination (classifying identities).

3 Experimental Study

In order to examine the *hierarchical coherence* within embeddings generated by DML models, this study applies clustering to embeddings created for the facial recognition task as a means to observe the additional information contained within the embedding structure while also investigating whether a hierarchy of representation exists between this additional information, i.e., do certain attributes have a stronger representation within the embedding structure.

3.1 Deep Metric Learning Model

The model chosen for the experiment is the *dlib face recognition ResNet model v1* contained in the Dlib library [King, 2009]. This model was selected as it supplies an open source, easy to use facial recognition convolutional neural network (CNN) with a classification accuracy of 99.38% on the standard Labeled Faces in the Wild (LFW) face recognition dataset [Liu et al., 2015]. The *dlib face recognition ResNet model v1* is based on the ResNet 29 architecture [He et al., 2016]. The creator of the model trained it using approximately three million faces, and 7485 individual identities, any overlap with the LFW dataset was avoided. The work of [King, 2009] specifies that the datasets used to train the network include the Face Scrub dataset [Hong-Wei and Stefan, 2014], the VGG dataset [Parkhi et al., 2015] and images sourced from the Internet. The model performs facial recognition by mapping images of faces to a 128-dimensional space where images of the same identity are mapped near each other and images of different identities are mapped far apart in the new embedding space. [King, 2009] states that the network training started with randomly initialized weights and used a metric loss function that tries to project all identities into non-overlapping clusters of distance threshold radius 0.6. When using a distance threshold of 0.6, the model received its highest accuracy of 99.38% on the standard LFW face recognition benchmark.

The facial recognition task can be achieved by applying a discriminative classification algorithm such as K-nearest neighbor (KNN) onto the embeddings without being conditioned about the number of classes or, in our case, the number of identities. Consequently, it is possible to perform facial recognition using only one example, achieving one-shot learning. In this study, we adopt the K-Means and Mixture of Gaussians (MoG) algorithms to gain a better insight into attribute representation within the embedding structure. Our work shows that a *hierarchical coherence* exists within the embedding structure as K-Means successfully forms coherent clusters where salient intra and extra-class attributes can be accurately classified. As a result of this, embeddings generated for one purpose, in our case, facial recognition, can find an additional application in sub-discriminating intra and extra-class attributes through no additional retraining of the DML model.

If the pretrained model has been trained with low quality images, e.g., low pixel resolution, then our subdiscrimination, such as skin tone, will be less accurate. We noticed this ourselves when we ran an inference on a low-quality image of a person wearing thin-rimmed glasses; the model may be less accurate at detecting if the person is wearing glasses.

The dataset has been very well reviewed both automatically and manually. Since this is a state-of-the-art model, the clustering accuracy is good, but there are two issues that may have influenced the results. Firstly, we use the non-overlapping threshold of radius 0.6. However, it is very likely a higher radius could have given a lower classification rate but a better subdiscrimination clustering accuracy. Secondly, we have cleaned datasets, which potentially increases the subdiscrimination accuracy.

3.2 Experimental Procedure

Several tests were undertaken to examine the *hierarchical coherence* to determine if embeddings generated by the model would cluster images based on specific attributes without any additional training of the model. For these tests, datasets with specific class attributes were manually created before investigating how accurately the K-Means algorithm can cluster the resulting embeddings based on attribute discrimination. The

attributes used within the experiments are categorized as extra-class and intra-class. extra-class attributes are facial features that are distinguishable between different identities, namely gender, skin tone and age. intra-class attributes are noted as facial features which are distinguishable between one unique identity, namely emotions, the presence and absence of beards, and the presence and absence of glasses. The images used for the test datasets were selected from the *facial expressions* dataset available from the Muxspace GitHub repository [Rowe, 2016]. The *facial expressions* dataset consists of 13,718 unprocessed images and was chosen due to its diversity in terms of gender, age, and ethnicity. It also displays variety in terms of emotions for unique identities, making it suitable for some intra-class tests. To perform intra-class attribute testing, each dataset must contain only one unique identity. However, for some intra-class tests, the *facial expressions* dataset did not contain enough images of the same identity with specific attributes. For example, as this dataset was created for facial expression diversity, finding images of the same identity with/without beards was challenging. Therefore, images were manually sourced from the Internet to create the required datasets for tests in which the attributes of beards and glasses were examined. Furthermore, as creating these datasets can be a time intensive task, simple data augmentations are applied to images used in the intra-class tests in order to increase the sample size to that of the the extra-class tests. These augmentations are discussed in greater detail here 4.2.

Images used for the extra-class tests were manually chosen from the *facial expressions* dataset. Images for the test datasets are manually chosen based on the desired attribute for discrimination. For example, in a test where the attribute gender is investigated, the test dataset would contain 100 images, of which 50 are male and 50 are female. Each image is manually labeled based on the perceived gender of the person within the image.

Each image within the dataset is then passed through the model to create an embedding. The model applies the following steps prior to creating the embedding representation. The frontal face detector detects the face within the image and places a bounding box around it. This verifies that a face is present and that the image does not contain more than one face. The landmarks for the detected face are then identified using the *shape predictor 68 face landmarks* model [King, 2009]; these landmarks are used to precisely localize the face. The images and their respective landmarks are then passed to the *dlib face recognition ResNet model v1*, which converts the images into their respective 128-dimensional embeddings.

Once the embeddings have been created, clustering is performed on the embeddings using Scikit-learns K-Means algorithm [Pedregosa et al., 2011], this experimental process can be seen in Figure 1. A bottom up (divisive) hierarchical clustering approach, we feel, is not suitable for this work as the image embeddings require a top-down (agglomerative) approach for us to search for the implicit unseen embedding features, such as old or young. We approach the search in a top-down manner using K-Means. The Chinese whispers clustering algorithm from the Dlib library was used initially. However, this algorithm did not perform well and was substituted for the K-Means algorithm. Several tests were carried out to compare the initialization of the K-Means algorithm using random seeds against manual seeds. The random seeds provided more consistent results; therefore, the rest of the tests were carried out using random seed initialization. This form of K-Means initialization can create slightly different clusters on occasion. Each dataset was run through the algorithm five times to obtain a more reliable average classification accuracy.

When the data is clustered, output labels are generated by K-Means, which are then compared with the manually created labels to assess the performance of the clustering through the use of confusion matrices.

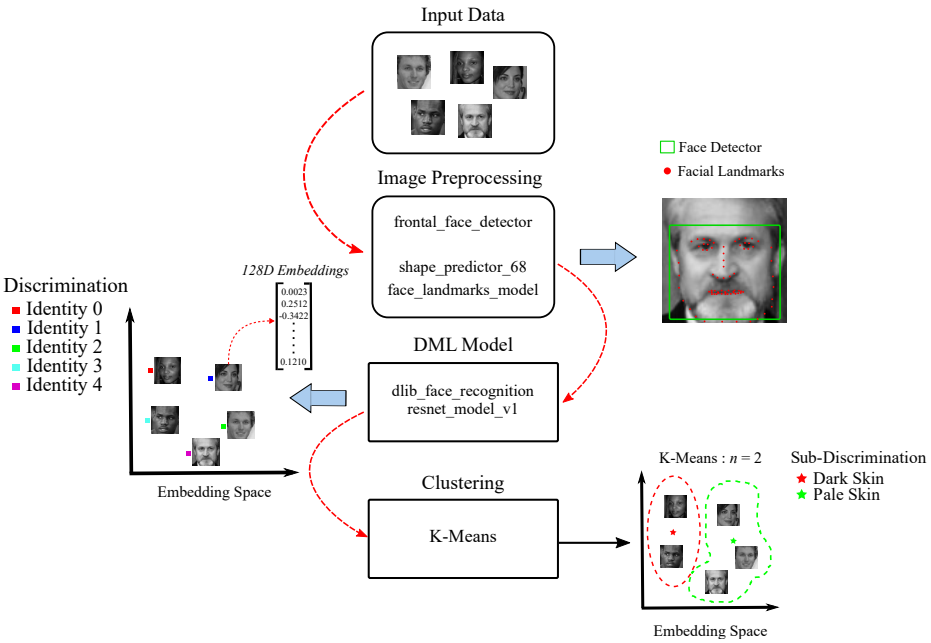

Figure 1: Flowchart of the experimental process.

4 Evaluation

4.1 Extra-class

A set of experiments was undertaken to examine the extra-class attributes representing gender, skin tone, and age, examples of images used in these tests can be seen in Figure 2. K-Means was initialised with random seeds and two clusters for each test. While it is appreciated that these attributes exist on a continuum, the attributes are classified in a binary fashion to understand how the representation within the embedding is structured. Therefore, dichotomic clusters are used as gender is classified as male/female, skin tone as dark/pale and age as young/old, where young is defined as individuals under the age of 50 and old as over 50.

The sample sizes for each test were 100 images, each test was run five times, and the average accuracy for each test, across all test runs, was 99.3%, 99.3% and 94.1% for gender, skin tone and age, respectively. The average accuracy represents the percentage of images that were assigned to the correct cluster over the number of test runs (five for this experiment). It is evident from the high clustering accuracies achieved above that the extra-class discriminative properties of gender, skin tone and age are represented within the embeddings.

An important outcome of this experiment was that it demonstrated the hierarchy of attributes within the embeddings. Each time the K-Means algorithm was run with two clusters, the embeddings would always cluster based on skin tone. Therefore, to examine the attributes of gender and age, all images in the dataset needed to contain only one skin tone. This requirement demonstrates that a hierarchy of representation exists between

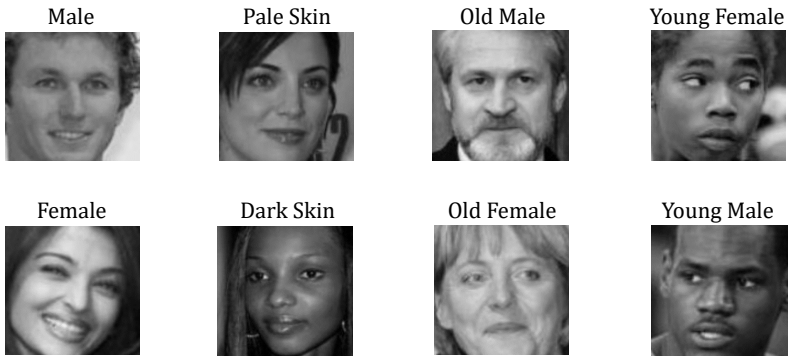

Figure 2: Examples of images used for the initial stage of extra-class testing, along with labels of how they were manually classified.

features within the embedding structure, such that specific attributes have a stronger representation within the embedding structure.

A new dataset of 200 images was created and manually labelled to explore this condition further, ensuring an equal representation of attributes within the dataset. In the first part of this experiment, K-Means was initialised with two clusters to determine which attribute held prominence above the other two. This process was repeated for K-Means, initialised with four clusters and eight clusters. The results of these experiments can be seen in Table 1.

When using two clusters, the data was classified based on skin tone, such that one cluster contained pale-skinned people and the other contained dark-skinned people. When four clusters were used, the clusters formed around the attributes of skin tone and gender. Of the four clusters formed, one contained males with dark skin tone, a second cluster contained females with dark skin tone, a third contained females with pale skin tone and a fourth contained males with pale skin tone. Finally, when eight clusters were used, each image was classified based on the skin tone, gender and age of the person in the image. For example, one cluster would contain old dark-skinned males, and another would contain young dark-skinned males. This hierarchy of extra-class attribute representation can be seen in Figure 3.

These results verify that not only are the extra-class attributes of age, gender and skin tone represented within the embedding structure, but they also follow a hierarchy of representation whereby certain attributes are represented better or deemed more critical to the facial recognition task during training. These attributes rank in order of skin tone, gender, and age, respectively. Subsequently, this new information regarding the *hierarchical coherence* within the embedding structure allows for the sub-discrimination of these attributes without retraining the original DML model. It is recognised that the model does not perform as strongly for the attribute age compared to the attributes of gender and skin tone. We speculate that the attributes of gender and skin tone can be represented more easily in a binary form and that this may not be the case for the attribute age. In future works, it may prove beneficial to represent age in more distinct age groups, such as 0-5 years, 6-10 years, etc., rather than over 50 years and under 50 years.

In the next section of this paper, we investigate the presence of the intra-class attributes

No. Clusters	Cluster Content	Test 1	Test 2	Test 3	Test 4	Test 5	Average Accuracy
2	Dark Skin	99.0%	99.0%	99.0%	99.0%	99.0%	**99.0%**
	Pale Skin	100.0%	100.0%	100.0%	100.0%	100.0%	**100.0%**
4	Dark Skin	99.0%	99.0%	99.0%	99.0%	99.0%	99.0%
	Pale Skin	100.0%	99.0%	99.0%	99.0%	100.0%	99.4%
	Male	100.0%	100.0%	100.0%	100.0%	100.0%	**100.0%**
	Female	99.0%	99.0%	99.0%	99.0%	99.0%	**99.0%**
8	Male	99.0%	100.0%	99.0%	99.0%	100.0%	99.4%
	Female	99.0%	99.0%	99.0%	97.0%	99.0%	98.6%
	Dark Skin	99.0%	99.0%	98.0%	99.0%	98.0%	98.6%
	Pale Skin	99.0%	99.0%	99.0%	99.0%	99.0%	99.0%
	Young	94.0%	90.0%	91.0%	93.0%	87.0%	**91.0%**
	Old	88.0%	79.0%	89.0%	80.0%	80.0%	**83.2%**

Table 1: Results of the 2^{nd} extra-class test where the hierarchical coherence between features in the embedding structure is examined. Outcomes indicate the ability to correctly identify a person's gender, age and skin tone with high classification accuracy (shown in bold). It is also evident that a hierarchical coherence exists between features in the embedding structure. The features representing the extra-class attributes rank in order of skin tone, gender and age, respectively.

of beards, glasses and emotions within the embedding structure, along with the possibility that the features representing the intra-class attributes behave in the same manner as the extra-class results indicate.

4.2 Intra-class

As discussed in Section 3.2, the *facial expressions* dataset was compatible with the intra-class experiments as it contained 3–10 images for each unique identity and showed significant variety in terms of emotions for each unique person. The initial intra-class test mirrored the initial extra-class test in that we first examined whether specific intra-class attributes are represented within the embedding structure. Firstly, the four emotions happy, angry, sad, and neutral were investigated as these were the best represented emotions in the *facial expressions* dataset. However, although these emotions were represented the best, they were not represented sufficiently, resulting in a shortage of images and, subsequently, very small datasets. To overcome this, basic image augmentations [Buslaev et al., 2020] were applied to bulk the sample size up to 80 images, near the 100 sample size for the extra-class tests. The original datasets contained 10 images, 5 images/attribute, three augmentations are applied to the original images. Firstly, the original images are rotated at 5 degrees, resulting in 20 total images. Secondly, Gaussian

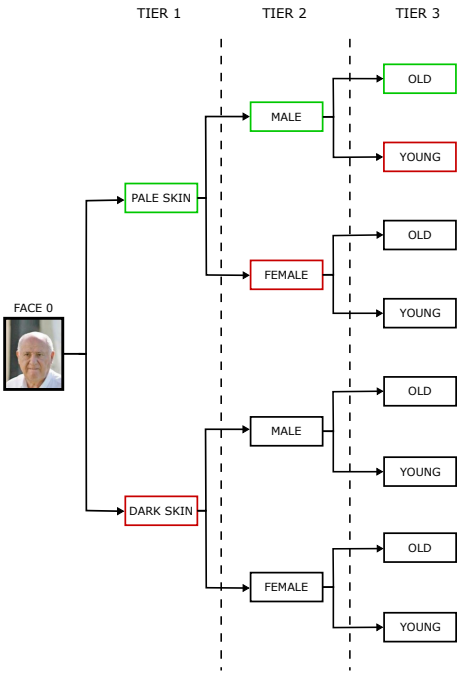

Figure 3: Hierarchy of extra-class attribute representation within the embedding structure. Tier 1 represents the attribute skin tone, which takes priority within the structure, while Tier 3 represents the attribute age, which ranks the lowest. The green boxes represent how face 0 was classified during the clustering process.

blur is applied to all images, resulting in 40 total images. Finally, image compression is applied to all images, resulting in a total of 80 images. Each attribute combination test was run 10 times, the average accuracy and max accuracy across all 10 test runs can be see in Table 2. High maximum accuracies for most emotion combinations suggest that emotions are represented within the embedding structure. However, low average accuracies suggest one-shot or few-shot learning techniques may be required to improve the sub-discriminative task. Additionally, the low average accuracies could be a consequence of applying an objective metric on a subjective attribute, such that the labels provided by the creators of the *facial_expressions* dataset can be subjective and may not be entirely representative of the emotion resulting in lower classification accuracies.

In the remainder of this section, we investigate whether the intra-class attributes of beards and glasses are represented within the embedding structure.

The presence of beards and glasses were examined using three separate datasets per discriminative property. None of the datasets used in prior experiments contained enough images of the same identity with/without beards/glasses; therefore, images were manually sourced from the Internet; examples of these images can be seen in Figure 4.

Emotion 1	Emotion 2	Identity	Sample Size	Average	Max
Anger	Sad	0	80	60.0%	90.0%
Anger	Neutral	0	80	**86.0%**	100.0%
Anger	Happy	0	80	65.0%	90.0%
Happy	Neutral	1	80	50.0%	70.0%
Happy	Sad	2	80	64.0%	80.0%
Neutral	Sad	3	80	46.0%	60.0%

Table 2: Results of the initial intra-class test, where the presence of the attribute emotions are examined. Average refers to the average accuracy across 10 test runs, while max refers to the maximum accuracy achieved across 10 test runs. Four separate identities (0 - 3) are used across all tests. The highest average accuracy is in bold.

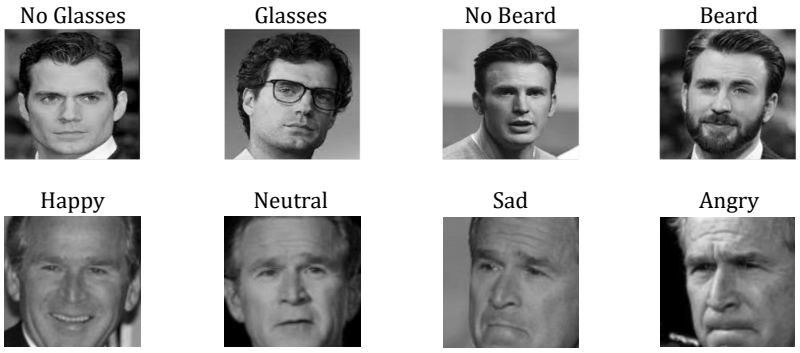

Figure 4: Examples of images sourced from the internet that were used to examine the presence of beards, glasses and emotions as intra-class discriminative properties.

Each dataset contains ten original images, augmented in the same manner as the initial intra-class test to increase the sample size to 80 images, results from these experiments are shown in Table 3.

High maximum and average accuracies for a majority of tests suggest that the attributes of beards and glasses are represented within the embedding structure. Similar to the extra-class results, the intra-class attributes of beards and glasses follow a hierarchy of representation within the embedding structure such that glasses rank above beards; this hierarchy can be seen in Figure 5.

The attributes of beards and glasses have a stronger representation within the embedding structure in comparison to the attribute emotions. This is highlighted by higher average and max classification accuracies in most cases. The attribute glasses display more stable results as there is not much variation between max and average accuracies for each dataset. Furthermore, although the attribute beards obtained the highest average accuracy of 90%, the low average max and average accuracies for identity 0 suggest the attribute glasses may have a stronger or more stable representation within facial

Attribute Test	Identity	Sample Size	Average	Max
Beard vs No Beard	0	80	52.0%	60.0%
	1	80	**90.0%**	100.0%
	2	80	85.0%	100.0%
Glasses vs No Glasses	2	80	70.0%	100.0%
	3	80	70.0%	100.0%
	4	80	**76.0%**	90.0%

Table 3: Results of the 2^{nd} intra-class test, where the presence of the attributes beards and glasses are examined. Average refers to the average accuracy across 10 test runs, while max refers to the maximum accuracy achieved across 10 test runs. Five separate identities (0 - 4) are used across all tests. The highest average accuracy for each attribute is in bold.

recognition embeddings.

4.3 Summary of Intra and extra-class tests

In summary, from initial inspection, the extra-class attributes of skin tone, gender and age are represented within the structure of facial recognition embeddings. Additionally, results suggest that a *hierarchical coherence* exists between attributes within the embedding structure whereby certain attributes are better represented; for example, when using dichotomic clusters, if a dataset contained more than one skin tone, the embeddings would always cluster based on skin tone. Furthermore, the presence of the intra-class attributes of beards and glasses have been identified to also exist within the embedding structure. Additionally, results indicate that the intra-class attributes of emotions are not represented as strongly as other attributes. Overall, most intra and extra-class attributes examined in this study held representation within the embedding structure. Subsequently, the presence of these attributes allows for the sub-discrimination between features not presented during the initial training of the DML model.

5 Application of Results

Once the coherence property is discovered, a myriad of applications can follow. In any realm where one-shot learning is possible, zero-shot learning in subclass discrimination through clustering of embeddings is a potential candidate. In the context of Natural Language Processing (NLP), for instance, if a one-shot learning model is trained to get the mood of the writer, a subdiscrimination on the embedding could give information about the topic of the written text. In the context of Automatic Speech Recognition (ASR) using one-shot learning to identify the speaker, a sub-discrimination on the embeddings could give information about the accent, the mood, or the pitch of the voice. In time series market analysis, if one-shot learning is trained to discriminate customers' typology, a zero-shot sub-discrimination on the embedding could give information about age, gender, and nationality.

From now on in the article, we give an application example in the domain of face recognition. While the results of this study suggest that additional techniques are required for the application of intra-class discriminative properties, a large avenue appears

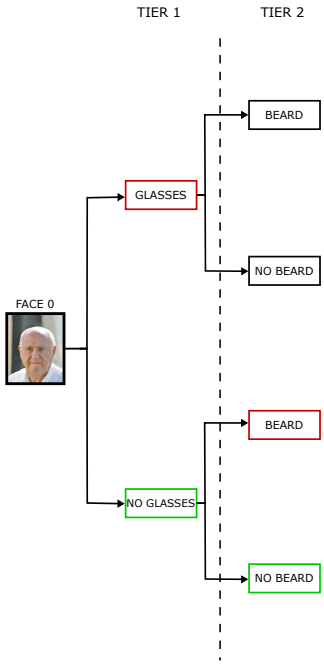

Figure 5: Hierarchy of intra-class attribute representation within the embedding structure. Tier 1 represents the attribute glasses which take priority within the structure, while Tier 2 represents the attribute beards, which ranks the lowest. The green boxes represent how face 0 was classified during the clustering process.

prevalent for the application of extra-class discriminative properties. One possible avenue includes training one of the many existing unsupervised learning algorithms to cluster data based on skin tone, gender and age. This can be accomplished by saving the resulting cluster centroids in training and clustering unseen data by taking the Euclidean distance of each unseen embedding and classifying its discriminative properties based on which cluster centroid it lies closest to. To highlight the possibility of this application, a dataset consisting of 1000 samples, where each sample represents a unique identity, was manually created from the CelebA dataset [Liu et al., 2014]. The dataset used for this experiment consisted of:

- 125 young pale-skinned males;
- 125 young dark-skinned males;
- 125 young pale-skinned females;
- 125 young dark-skinned females;
- 125 old pale-skinned males;
- 125 old pale-skinned females;
- 125 old dark-skinned males;

– 125 old dark-skinned females.

This dataset was broken down into 70% train data and 30% test data; a validation set was not used as we are training an unsupervised learning algorithm. In adherence with previous extra-class experiments, K-Means initialized with random seeds was chosen as the unsupervised learning algorithm. The training data was processed through K-Means several times using two clusters initially, then four clusters and finally eight clusters. The cluster centroids for each cluster are generated by the K-Means algorithm and saved. The centroids that produced the highest attribute accuracies in training are chosen to cluster the test data. Classification is achieved by classifying embeddings based on the cluster centroid they lie closest to. The results of this experiment can be seen in Figure 6 denoted as 'K-Means Train' and 'K-Means Test'.

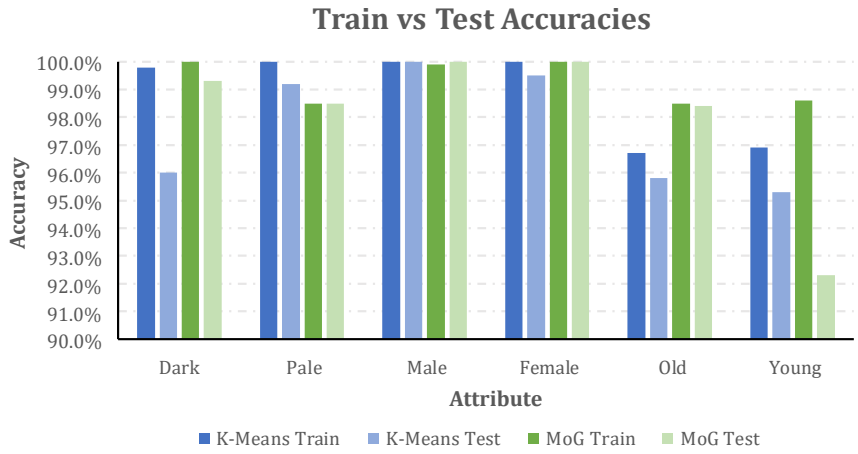

Figure 6: The train and test accuracies for both MoG and K-Means algorithms. Both algorithms display high accuracies for each extra-class attribute.

The results from this experiment indicate the possibility of using means generated by an unsupervised learning algorithm to cluster embeddings based on unseen extra-class discriminative properties. This yields an advantage in terms of computation speed for the sub-discrimination task. In contrast to comparing a new embedding to each known identity, it is more computationally efficient to compare a new embedding to each cluster centroid. This comparison reduction, especially with larger sample sizes, can significantly increase computation speed while maintaining high performance on the sub-discrimination task, as shown by the high attribute accuracies achieved for this experiment.

In the final stage of this experimental study, the substantial drop in accuracy for the dark skin attribute between the train and test sets in the prior experiment is examined. To investigate whether K-Means was accurate enough, a different unsupervised algorithm was trained and tested on the same datasets. The algorithm used for this experiment was

the MoG. The method used to initialize the weights, the means and the precision was left as K-Means, which is the default initial parameter option for the MoG algorithm.

Figure 6 shows the train and test accuracies for both MoG and K-Means. The results show that both algorithms can accurately identify extra-class discriminative properties. MoG does increase the accuracy of the dark skin attribute by 3.3%. However, it does not perform as well as K-Means for the pale skin attribute, dropping by 0.7%. Although both algorithms perform well, the results are inconclusive enough to state which algorithm performs best. A path for future work in the area is to examine the behaviour of several unsupervised algorithms to determine which performs best in terms of feature classification.

The decrease in accuracy for the dark skin attribute can be seen as a consequence of the content of the test and training datasets, additionally, due to the limitation of the experiment whereby, this study attempts to perform binary classification on continuous attributes. For the chosen dataset, dark skin people are defined as being of dark or mixed skin tone. It is noted that 19% of the images representing the dark class in the training set were mixed skin tones, while 32% of the images representing the black class in the test set were mixed skin tones. The content of the train and test sets for the dark skin tone attribute can be seen in Figure 7. This indicates that because of the low percentage of mixed skin tone images in the training set, the model is less likely to classify mixed skin tone data in the test dataset correctly; therefore, providing a more balanced training set in terms of skin tone should lead to improved results.

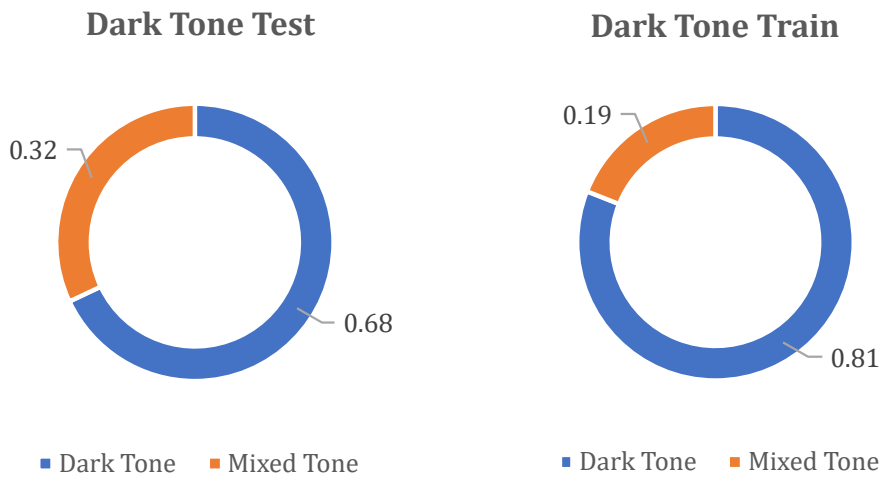

Figure 7: The contents of the train and test sets for the dark skin tone attribute. An uneven representation of dark and mixed tone images between the test and train sets could explain the decrease in in accuracy for the dark skin attribute.

6 Future Work

The most relevant area with respect to future work is applying the techniques developed in this study to embeddings created by other data. Embeddings created for any purpose may contain inherent information that could be used to perform sub-discriminative tasks through near zero-shot learning. Speech recognition is one of many possible areas where these principles can be applied. For example, spectrograms of voices can be represented as embeddings, and through the use of an unsupervised learning algorithm, it may be possible to identify disparate speakers accurately.

7 Conclusion

In this paper, the presence and application of additional information represented within the structure of embeddings generated for facial recognition were experimentally evaluated, and the results confirm the presence of intra and extra-class facial attributes within the embedding structure. In addition, it is shown that through the use of clustering, this inherent information has application in sub-discriminating additional features not presented during the initial training of the DML model. extra-class sub-discrimination can be achieved with high accuracy, notably with an average attribute accuracy of 99.3%, 99.3% and 94.1% for the attributes of skin tone, gender, and age, respectively. In addition, the intra-class attributes of beards and glasses yielded an average attribute accuracy of 90.0% and 76.0%, respectively. The main findings of this experimental study are summarized below:

- It is possible to perform extra-class sub-discriminative tasks with a high degree of accuracy through the use of unsupervised clustering algorithms. The discovery of inherent information within embeddings designed for facial recognition confirms the ability to perform extra-class sub-discrimination, namely for the attributes gender, skin tone and age;
- The results from the intra-class sub-discrimination experiments highlight the need for additional techniques to help increase the ability to extract/identify these intra-class attributes with higher accuracies. A one-shot or few-shot learning approach may substantially increase the attribute accuracy;
- A *hierarchical coherence* between attributes suggest that specific facial attributes take precedence during the training/creation of facial recognition embeddings; our results indicate that the extra-class attributes rank in order of skin tone, gender, and age, respectively, and the intra-class attributes rank in order of glasses and beards, respectively.
- Finally, the possibility of training unsupervised algorithms to perform extra-class sub-discrimination at extremely high accuracies by saving cluster centroids created during training is demonstrated.

References

[Buslaev et al., 2020] Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., and Kalinin, A. A. (2020). Albumentations: Fast and flexible image augmentations. *Information*, 11(2).

[Chen et al., 2020] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *37th International Conference on Machine Learning, ICML 2020*, PartF168147-3(Figure 1):1575–1585.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.*, 20(3):273–297.

[Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

[Fan et al., 2014] Fan, H., Yang, M., Cao, Z., Jiang, Y., and Yin, Q. (2014). Learning compact face representation: Packing a face into an int32. In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM '14, page 933–936, New York, NY, USA. Association for Computing Machinery.

[Górriz et al., 2020] Górriz, J. M., Ramírez, J., Ortíz, A., Martinez-Murcia, F. J., Segovia, F., Suckling, J., Leming, M., Zhang, Y.-D., Álvarez-Sánchez, J. R., Bologna, G., et al. (2020). Artificial intelligence within the interplay between natural and artificial computation: Advances in data science, trends and applications. *Neurocomputing*, 410:237–270.

[Hadsell and LeCun, 2006] Hadsell, C. and LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2:1735–1742.

[Hasrul et al., 2012] Hasrul, M., Hariharan, M., and Yaacob, S. (2012). Human affective (emotion) behaviour analysis using speech signals: A review. In *2012 International Conference on Biomedical Engineering (ICoBE)*, pages 217–222. IEEE.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

[Hong-Wei and Stefan, 2014] Hong-Wei, N. and Stefan, W. (2014). A data-driven approach to cleaning large face datasets hong-wei ng and stefan winkler advanced digital sciences center (adsc), university of illinois at urbana-champaign , singapore. *International Conference on Image Processing(ICIP)*, pages 343–347.

[Hu et al., 2016] Hu, Y., Liu, H., Pfeiffer, M., and Delbruck, T. (2016). Dvs benchmark datasets for object tracking, action recognition, and object recognition. *Frontiers in neuroscience*, 10:405.

[Jain et al., 2018] Jain, S., Farooque, M. U., and Sharma, V. (2018). Comparative Analysis of Clustering Algorithm for Facial Recognition System. *Proceedings of the 8th International Advance Computing Conference, IACC 2018*, pages 102–107.

[Karlinsky et al., 2019] Karlinsky, L., Shtok, J., Harary, S., Schwartz, E., Aides, A., Feris, R., Giryes, R., and Bronstein, A. M. (2019). Repmet: Representative-based metric learning for classification and few-shot object detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:5192–5201.

[King, 2009] King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758.

[Kulis, 2012] Kulis, B. (2012). Metric learning: A survey. *Foundations and Trends in Machine Learning*, 5(4):287–364.

[Lake et al., 2015] Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338.

[LeCun et al., 1998] LeCun, Y., Cortes, C., and Burges, C. J. (1998). The mnist database of handwritten digits, 1998. *URL http://yann. lecun. com/exdb/mnist*, 10:34.

[Lin et al., 2015] Lin, T. Y., Roychowdhury, A., and Maji, S. (2015). Bilinear CNN models for fine-grained visual recognition. *Proceedings of the IEEE International Conference on Computer*

Vision, 2015 Inter:1449–1457.

[Liu et al., 2015] Liu, J., Deng, Y., Bai, T., Wei, Z., and Huang, C. (2015). Targeting ultimate accuracy: Face recognition via deep embedding. *Arxiv*, pages 1–5.

[Liu et al., 2014] Liu, Z., Luo, P., Wang, X., and Tang, X. (2014). Deep learning face attributes in the wild. *CoRR*, abs/1411.7766.

[Nilsback and Zisserman, 2008] Nilsback, M. E. and Zisserman, A. (2008). Automated flower classification over a large number of classes. *Proceedings - 6th Indian Conference on Computer Vision, Graphics and Image Processing, ICVGIP 2008*, pages 722–729.

[Parkhi et al., 2015] Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). Deep face recognition. In *British Machine Vision Conference*.

[Parkhi et al., 2012] Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. V. (2012). Cats and dogs. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2:3498–3505.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., and Thirion, B. (2011). Scikit-learn: Machine learning in python.

[Radford et al., 2021] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *Arxiv*.

[Rezende et al., 2016] Rezende, D. J., Mohamed, S., Danihelka, I., Gregor, K., and Wierstra, D. (2016). One-shot generalization in deep generative models. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 1521–1529. JMLR.org.

[Rowe, 2016] Rowe, B. L. Y. (2016). $github : muxspace/facial_expressions : Asetofimagesforclassifyingfacialexpressions.$

[Russakovsky et al., 2015] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

[Russakovsky and Fei-Fei, 2012] Russakovsky, O. and Fei-Fei, L. (2012). Attribute learning in large-scale datasets. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6553 LNCS(PART 1):1–14.

[Santoro et al., 2016] Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., and Lillicrap, T. (2016). One-shot learning with memory-augmented neural networks. *arXiv preprint arXiv:1605.06065*.

[Tan et al., 2015] Tan, C., Lallee, S., and Orchard, G. (2015). Benchmarking neuromorphic vision: lessons learnt from computer vision. *Frontiers in neuroscience*, 9:374.

[Vanesa Sancho, 2011] Vanesa Sancho, E. (2011). TEMA 6 LA CLASIFICACIÓN DE LOS SERES VIVOS Contenidos. *2011 International Conference on Computer Vision*, pages 89–96.

[Wang et al., 2019] Wang, W., Zheng, V. W., Yu, H., and Miao, C. (2019). A survey of zero-shot learning: Settings, methods, and applications. *ACM Trans. Intell. Syst. Technol.*, 10(2).

[Wang et al., 2020] Wang, Y., Yao, Q., Kwok, J. T., and Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)*, 53(3):1–34.