Transductive Zero-Shot and Few-Shot CLIP

Ségolène Martin*, Yunshi Huang†, Fereshteh Shakeri†, Jean-Christophe Pesquet*, Ismail Ben Ayed†

* Université Paris-Saclay, Inria, CentraleSupélec, CVN
† ÉTS Montréal

Abstract

Transductive inference has been widely investigated in few-shot image classification, but completely overlooked in the recent, fast growing literature on adapting vision-langage models like CLIP. This paper addresses the transductive zero-shot and few-shot CLIP classification challenge, in which inference is performed jointly across a mini-batch of unlabeled query samples, rather than treating each instance independently. We initially construct informative vision-text probability features, leading to a classification problem on the unit simplex set. Inspired by Expectation-Maximization (EM), our optimization-based classification objective models the data probability distribution for each class using a Dirichlet law. The minimization problem is then tackled with a novel block Majorization-Minimization algorithm, which simultaneously estimates the distribution parameters and class assignments. Extensive numerical experiments on 11 datasets underscore the benefits and efficacy of our batch inference approach. On zero-shot tasks with test batches of 75 samples, our approach yields near 20% improvement in ImageNet accuracy over CLIP's zero-shot performance. Additionally, we outperform state-of-the-art methods in the few-shot setting. The code is available at: `https://github.com/SegoleneMartin/transductive-CLIP`.

1. Introduction

The emergence of large-scale vision-language models like CLIP [40] has marked a significant turning point in representation learning [24, 29, 54]. By integrating both visual and textual modalities, these models have shown remarkable potential in crafting generic and richly informative concepts. Unlike traditional vision models, often constrained by task specificity, the representations gleaned from vision-language models are versatile, setting the stage for a breadth of downstream vision tasks and expanding the horizons of what is achievable in the domain.

Among the vision tasks that can be addressed with vision-language models, zero-shot and few-shot classification have particularly attracted attention. Notably, CLIP has demonstrated strong performance in zero-shot classification [40]. Several subsequent works have leveraged few-shot data, a few labeled samples in the target downstream task, to further improve CLIP's classification accuracy. Following on from the research on prompt learning in the NLP community, CoOp and CoCoOp [57, 58] fine-tuned the pre-trained CLIP model using learnable textual tokens. Another type of approaches, like CLIP-Adapter [17] and TIP-Adapter [56] provided CLIP with a parametric feature transformation, which generates adapted features and combines them with the original CLIP-encoded features. Despite their efficacy on few-shot classification benchmarks, these methods predominantly operate within the so-called *inductive* setting, where inference is conducted independently for each query (i.e., test) sample.

In contrast, in the *transductive* paradigm, one makes joint predictions for a batch of query samples, taking advantage of the query set statistics. The transductive setting for few-shot classification with vision-only models was pioneered in [31], and have since become prominent research subject, triggering an abundant, very recent literature on the subject, e.g., [30, 34, 47, 48, 50, 52, 59, 60], to list a few. These transductive few-shot classifiers were shown to significantly outperform their inductive counterparts, with benchmarks indicating up to a 10% increase in classification accuracy [6]. In fact, this is in line with well-established theoretical facts in the classical literature on transductive learning [25, 51], which points to transductive prediction as a way to alleviate the scarcity of labeled data. Importantly, and beyond theoretical justification, the transductive setting is highly relevant in a breadth of practical computer vision scenarios, in which test data may come in mini-batches. This is the case, for instance, of online video streams and various types of time-series imaging, of portable-device photos, or of pixel-level tasks such as segmentation.

In this study, we take a close look at the transductive zero-shot and few-shot inference problems for the popular vision-language pre-trained CLIP model. We first make the surprising observation that standard clustering models, in the zero-shot case, and recent transductive methods, in the few-shot setting, do not bring improvements comparable to those observed with vision-only models, scoring even below their inductive counterparts; see Tables 1 and 2. This might explain why the transductive setting, despite its popularity, has not been explored so far for vision-language models. Potential questions that may fill this gap are (i) How to build informative text-image features for transductive inference, leveraging the textual knowledge in vision-language models? and (ii) Are the statistical assumptions underlying standard clustering and transductive inference methods appropriate for text-image features? In light of these challenges, this paper brings the following contributions:

1. We propose a methodology to compute text-vision probability feature vectors, setting the stage for transductive few-shot classification specifically tailored for CLIP.
2. We reformulate the transductive zero-shot and few-shot classification challenge as an optimization problem on the unit simplex set by modeling the data with Dirichlet probability distributions. Crucially, the non-trivial deployment of the Dirichlet distributions brings substantial improvements in comparison to the common statistical models underlying standard clustering and transductive few-shot methods (e.g. Gaussian).
3. We propose a novel block Majorization-Minimization algorithm that addresses our problem efficiently and effectively, removing the need for cumbersome inner iterations in estimating the Dirichlet parameters.
4. We report comprehensive evaluations, comparisons and ablations over 11 datasets, which point to the benefits of our mini-batch inference approach. On zero-shot ImageNet tasks with batches of 75 samples, the proposed method scores near 20% higher than inductive zero-shot CLIP in classification accuracy. Additionally, we outperform state-of-the-art methods in the few-shot setting.

2. Related works

2.1. Vision-language models

Vision-Language models, like CLIP, integrate visual and textual data to improve accuracy over various vision tasks. CLIP uses a dual-encoder structure, with one deep network dedicated for image encoding and another one specilaized for text. This structure, along with proper projections at its bottleneck, yield image and text embeddings lying in the same low-dimensional vector space. Trained on a large dataset of 400 million text-image pairs, CLIP maximizes the cosine similarity between text and image embeddings using a contrastive loss. CLIP is pre-trained to match images with text descriptions, making it well-suited for zero-shot prediction. At inference time, to classify an image $\boldsymbol{x}$ among K classes, the model predicts the class by choosing the one with the highest cosine similarity:

$$\underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}} \cos\left(f_{\text{im}}(\boldsymbol{x}), f_{\text{text}}(\boldsymbol{t}_k)\right), \quad (1)$$

where f_{im} and f_{text} are, respectively, the image and text encoders, and each $\boldsymbol{t}_k$ is based on a text prompt, typically "a photo of a [name of class k]".

2.2. Few-shot classification

Inductive v.s. transductive setting Few-shot image classification with pre-trained vision models has been the subject of extensive research recently [11, 49, 60]. Within this area, the problem is tackled either in the *transductive* or *inductive* setting. The latter assumes that each instance in the testing batch is classified independently, omitting the correlations or shared information among instances [11, 22, 53]. In contrast, transductive inference is more comprehensive, as it makes joint predictions for the entire mini-batch of query samples, leveraging their statistics and shared information. Recent research has increasingly focused on transductive few-shot learning, including, for instance, methods based on constrained clustering [7, 34], label propagation [31, 59], optimal transport [28, 48], information maximization [6, 52], prototype rectification [30], among other recent approaches [47, 50]. It has been consistently observed in this body of literature that the gap in accuracy between transductive and inductive methods could be considerable.

Few-shot CLIP Beyond its zero-shot capabilities, the CLIP model has also been explored for few-shot image classification. In [40], the authors evaluated linear probe, which performs a simple fine-tuning of the visual encoder's final layer using a few-shot support set (i.e., a few labeled samples in the downstream task). This approach has proven to be relatively ineffective in few-shot scenarios. Since then, a recent body of works have explored CLIP's few-shot generalization. For instance, there is a noticeable emergence of *prompt learning* methods in computer vision, focusing on this specific problem [10, 57, 58]. Inspired by intensive recent prompt learning research in NLP [21, 44], these methods fine-tune learnable input text tokens using the few-shot support set. A different type of approaches, coined *adapters* [17, 56], fine-tune the encoded features rather than input text. For example, CLIP-Adapter [17] incorporates additional bottleneck layers to learn new features, while performing residual-style blending with the original pre-trained features. In a similar spirit, TIP-Adapter [56] balances two prediction terms, one summarizing adaptively the information from the support set and the other preserving the textual knowledge from CLIP. All of these recent methods belong to the inductive family. To the best of our knowledge, our work is the first to explore transduction for CLIP's

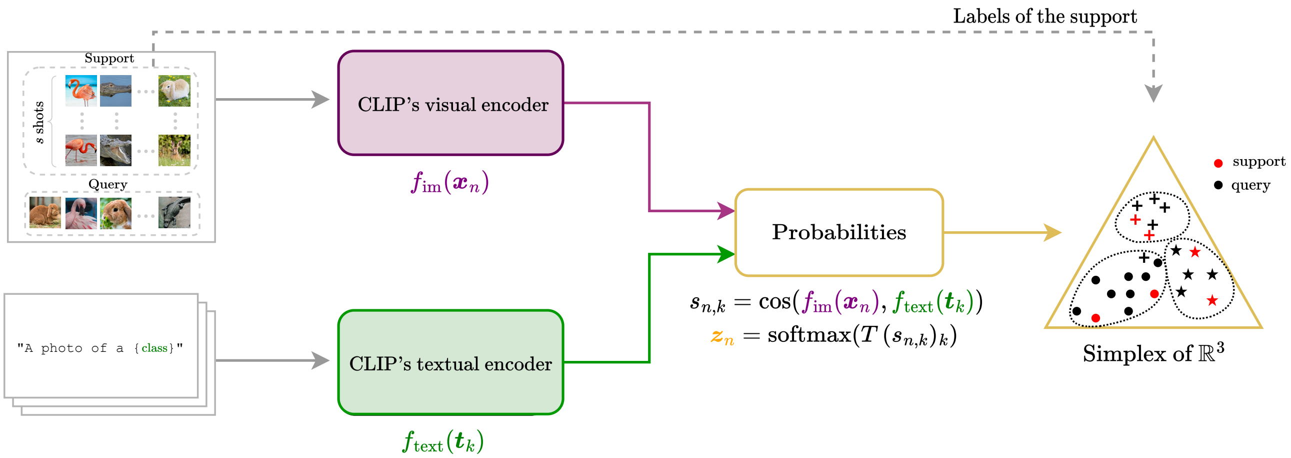

Figure 1. Given a transductive few-shot task, both visual and textual information are extracted from the images and class-wise prompts. The embeddings are next combined into vision-text probability vectors. Classification is carried out on the simplex set of $\mathbb{R}^K$ using the labels of the support set. An empty support set corresponds the the zero-shot scenario, which is akin to a clustering problem.

few-shot image classification.

3. Proposed method

Throughout the paper, we define and employ specific notations to describe a single, randomly sampled few-shot task, which, during transductive prediction, is treated independently of the other randomly sampled tasks:

- N is the number of images in each randomly sampled task, with $(\boldsymbol{x}_n)_{1\leq n\leq N}$ denoting the set of images.
- K is the total number of distinct classes in the whole data set, among which a much smaller set of randomly sampled classes might appear in each mini-batch task, and might differ from one batch to another. Hence, apart from knowing the set of K classes in the whole data, as in standard inductive inference [40, 56, 57], our transductive setting do not assume any additional knowledge about the particular set of classes that might appear randomly in each mini-batch.
- $\mathbb{S} \subset \{1, \ldots, N\}$ indicates the indices of samples within the support set in the few-shot setting. For all $n \in \mathbb{S}$, one has access to the one-hot-encoded labels $\boldsymbol{y}_n \in \{0,1\}^K$, such that for all $k \in \{1, \ldots, K\}$, $y_{n,k} = 1$ if $\boldsymbol{x}_n$ is an instance of class k, $y_{n,k} = 0$ otherwise.
- $\mathbb{Q} = \{1, \ldots, N\} \setminus \mathbb{S}$ represents the indices of samples in the query min-batch set. In our experiments, the mini-batch size $|\mathbb{Q}|$ is set to 75.

The goal is to predict the classes of the query samples leveraging the supervision available from the support set. Note that when the support set is empty ($\mathbb{S} = \varnothing$), we encounter a zero-shot scenario, which is akin to a clustering problem.

3.1. Computing informative feature vectors

A seemingly intuitive approach to tackle the transductive challenge might be to use the visual embeddings obtained from CLIP's visual encoder as the input features for the classifier. This is analogous to CLIP's linear probe when it operates inductively. We pinpoint two main difficulties raised by this approach:

1. **Overlooking textual information**: A significant limitation of this method is that it omits the model'stextual knowledge. This is problematic as textual information is one of CLIP's most powerful features.
2. **Normalization dilemma**: CLIP's pre-training maximizes the scalar product between normalized textual and visual embeddings. Using normalized embeddings can introduce complexities in data distribution modeling, which, if misjudged, can impact the method interpretability and accuracy.

While some works in the classification literature have explored spherical distributions like the Von Mises-Fisher [19, 41, 43] and the Fisher-Bingham [1, 18], our approach differs to address both issues mentioned above.

Our strategy consists in defining, for every $n \in \{1, \ldots, N\}$, the feature vector for the data sample $\boldsymbol{x}_n$ as CLIP's zero-shot probability. Precisely, we define

$$\boldsymbol{z}_n = \text{softmax}\left\{T\, \cos\left(f_{\text{im}}(\boldsymbol{x}_n), f_{\text{text}}(\boldsymbol{t}_k)\right)_{1\leq k\leq K}\right\}, \quad (2)$$

where $T > 0$ is a temperature parameter. Through this, both visual and textual information are incorporated into the feature vectors. Consequently, the task becomes a classification problem on the unit simplex of $\mathbb{R}^K$, defined as

$$\Delta_K = \left\{\boldsymbol{p} = (p_k)_{1\leq k\leq K} \in \mathbb{R}_+^K \,\middle|\, \sum_{k=1}^{K} p_k = 1\right\} \quad (3)$$

Observe that, for datasets with a modest number of classes, defining feature vectors according to (2) also acts as a dimensionality reduction, with embedding's dimension going from 1024 (from CLIP's ResNet50) down to K, the number of classes. A recap of our framework is given in Figure 1.

3.2. Data distribution

Given feature vectors lying within the unit simplex set of $\mathbb{R}^K$, we advocate modeling the data using Dirichlet distributions. The Dirichlet distribution extends the beta distribution into higher dimensions, serving as a natural choice for modeling probability vectors over the simplex. For each class k within the set $\{1,\ldots,K\}$, the data is assumed to follow a Dirichlet distribution, characterized by positive parameters $\boldsymbol{\alpha}_k = (\alpha_{k,i})_{1\le i\le K} \in (0,+\infty)^K$, which describes the distribution shape. An illustration in $\mathbb{R}^3$ is given is Appendix A. Mathematically, the density function is given by, for every $\boldsymbol{z} = (z_i)_{1\le i\le K} \in \mathbb{R}^K$,

$$\mathrm{p}\left(\boldsymbol{z} \mid \boldsymbol{\alpha}_k\right) = \frac{1}{\mathcal{B}(\boldsymbol{\alpha}_k)} \prod_{i=1}^{K} z_i^{\alpha_{k,i}-1} \mathbb{1}_{\boldsymbol{z}\in\Delta_K}, \tag{4}$$

where normalization factor $\mathcal{B}(\boldsymbol{\alpha}_k)$ is expressed as

$$\mathcal{B}(\boldsymbol{\alpha}_k) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_{k,i})}{\Gamma\left(\sum_{i=1}^{K} \alpha_{k,i}\right)}, \tag{5}$$

and Γ denoting the Gamma function.

3.3. Simplex-based classification criterion

The proposed method simultaneously determines: (*i*) the soft assignment vectors $\boldsymbol{u} = (\boldsymbol{u}_n)_{1\le n\le N}$ within the simplex $(\Delta_K)^N$, where the k-th component $u_{n,k}$ of vector $\boldsymbol{u}_n$ specifies the probability for the n-th sample belonging to class k; (*ii*) the Dirichlet distribution parameters $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_k)_{1\le k\le K}$ where each $\boldsymbol{\alpha}_k$ is a K-dimensional vector with nonnegative components. We achieve this through the following maximum-likelihood estimation

$$\begin{aligned} &\underset{\boldsymbol{u},\boldsymbol{\alpha}}{\text{minimize}} \ -\mathcal{L}(\boldsymbol{u},\boldsymbol{\alpha}) + \Phi(\boldsymbol{u}) + \lambda\Psi(\boldsymbol{u}), \\ &\text{subject to } \ \boldsymbol{u}_n \in \Delta_K \quad \forall n \in \mathbb{Q}, \\ &\qquad\qquad u_{n,k} = y_{n,k} \quad \forall n \in \mathbb{S}, \forall k \in \{1,\ldots,K\}. \end{aligned} \tag{6}$$

In (6), $\mathcal{L}$ is the log-likelihood model fitting objective for clustering:

$$\mathcal{L}(\boldsymbol{u},\boldsymbol{\alpha}) = \sum_{n=1}^{N}\sum_{k=1}^{K} u_{n,k} \ln(\mathrm{p}\left(\boldsymbol{z}_n \mid \boldsymbol{\alpha}_k\right)), \tag{7}$$

where the density functions are defined by the Dirichlet models in (4). When the support set is not empty, this term also includes the supervision derived from the labeled instances. Term Φ acts as a barrier imposing the nonnegativity constraints on assignment variables, as in the soft K-means objective [32, p.289], and is defined as

$$\Phi(\boldsymbol{u}) = \sum_{n=1}^{N}\sum_{k=1}^{K} u_{n,k} \ln u_{n,k}. \tag{8}$$

Finally, the penalty function Ψ, weighted by parameter $\lambda \in [0,+\infty)$, evaluates a partition complexity [8, 34], linked to the Minimum Description Length (MDL) concept in information theory:

$$\Psi(\boldsymbol{u}) = -\sum_{k=1}^{K} \pi_k \ln \pi_k, \tag{9}$$

where $\pi_k = \frac{1}{|\mathbb{Q}|}\sum_{n\in\mathbb{Q}} u_{n,k}$ is the proportion of query samples within class k. This MDL term penalizes the number of non-empty clusters, encouraging low-complexity partitions, i.e., with lower numbers of clusters.

4. Proposed algorithm

To tackle the minimization problem (6), our algorithm alternates minimization steps on the assignment variables and the Dirichlet parameters, producing sequences $(\boldsymbol{u}^{(\ell)})_{\ell\in\mathbb{N}}$ and $(\boldsymbol{\alpha}^{(\ell)})_{\ell\in\mathbb{N}}$ making the objective function decrease.

4.1. Minimization step w.r.t Dirichlet parameter

Suppose $\boldsymbol{u} \in \Delta_K^N$ is fixed. The estimation step with respect to the Dirichlet parameters consists in maximizing the log-likelihood in (7). Given the separability of the cost with respect to the variables $(\boldsymbol{\alpha}_k)_{1\le k\le K}$, we can, without loss of generality, consider the minimization of the function F_k defined as $(\forall \boldsymbol{\alpha}_k \in (0,+\infty)^K)$

$$\begin{aligned} F_k(\boldsymbol{\alpha}_k) = \sum_{n=1}^{N} u_{n,k} \Bigg(&\sum_{i=1}^{K} -(\alpha_{k,i}-1)\ln z_{n,i} \\ &+ \sum_{i=1}^{K} \ln\Gamma(\alpha_{k,i}) - \ln\Gamma\left(\sum_{i=1}^{K}\alpha_{k,i}\right)\Bigg). \end{aligned} \tag{10}$$

The minimization of the Dirichlet negative log-likelihood (10) has already been explored in the literature [23, 35, 36]. The main strategy consists in resorting to a Majorization-Minimization (MM) algorithm. Specifically, a generic MM procedure corresponds to finding a minimizer $\boldsymbol{\alpha}_k^*$ of F_k by iteratively producing a sequence $(\boldsymbol{\alpha}_k^{(m)})_{m\in\mathbb{N}}$ such that, for every $m \in \mathbb{N}$,

$$\boldsymbol{\alpha}_k^{(m+1)} = \underset{\boldsymbol{\alpha}_k\in(0,+\infty)^K}{\operatorname{argmin}} \ q(\boldsymbol{\alpha}_k; \boldsymbol{\alpha}_k^{(m)}), \tag{11}$$

where for all $\boldsymbol{\beta}_k \in (0,+\infty)^K$, $q(\,\cdot\,;\boldsymbol{\beta}_k)$ is a so-called *tangent majorant* of F_k, satisfying

$$\begin{cases} F(\boldsymbol{\alpha}_k) \le q(\boldsymbol{\alpha}_k;\boldsymbol{\beta}_k), \\ F(\boldsymbol{\beta}_k) = q(\boldsymbol{\beta}_k;\boldsymbol{\beta}_k). \end{cases} \tag{12}$$

The efficiency of the procedure (11) is highly dependent on the choice of the majorant. In [35], the author proposed a majorant function of (10) which consists in linearizing the concave term $\boldsymbol{\alpha}_k \mapsto -\ln\Gamma\left(\sum_{i=1}^{K}\alpha_{k,i}\right)$ at $\boldsymbol{\beta}_k$. The resulting MM algorithm was used for simplex clustering in [4, 38]. However, minimizing this majorant requires inverting the digamma function (i.e., the derivative of the log-Gamma function) with a Newton method, which can jeopardize the numerical convergence and slow down the overall algorithm.

In the following lemma, we introduce a novel tight majorant of F_k which yields closed-form updates, therefore avoiding sub-iterations within the MM algorithm.

Lemma 1 (Majorant of the negative log-likelihood)**.** *Let $\varphi = \ln\Gamma(\cdot+1)$. Then, for any $\boldsymbol{\beta}_k = (\beta_{k,i})_{1\le i\le K} \in (0,+\infty)^K$, the function $q(\,\cdot\,;\boldsymbol{\beta}_k)$ defined as, for every $\boldsymbol{\alpha}_k \in (0,+\infty)^K$,*

$$
\begin{aligned}
&q(\boldsymbol{\alpha}_k;\boldsymbol{\beta}_k)\\
&=\sum_{n=1}^{N}u_{n,k}\Bigg[\sum_{i=1}^{K}\Big(-(\alpha_{k,i}-1)\ln z_{n,i}-\ln\alpha_{k,i}\\
&+\varphi(\beta_{k,i})+\varphi'(\beta_{k,i})(\alpha_{k,i}-\beta_{k,i})+\frac{c(\beta_{k,i})}{2}(\alpha_{k,i}-\beta_{k,i})^2\Big)\\
&-\ln\Gamma\Big(\sum_{i=1}^{K}\beta_{k,i}\Big)-\left(\sum_{i=1}^{K}(\alpha_{k,i}-\beta_{k,i})\right)(\ln\Gamma)'\Big(\sum_{i=1}^{K}\beta_{k,i}\Big)\Bigg]
\end{aligned}
\tag{13}
$$

is a tangent majorant of F_k at $\boldsymbol{\beta}_k$, where the function c is defined by

$$
c\colon t\mapsto\begin{cases}\varphi''(0) & \text{if } t=0,\\ 2\dfrac{\varphi(0)-\varphi(t)+\varphi'(t)t}{t^2} & \text{otherwise.}\end{cases}\tag{14}
$$

A proof of Lemma 1 is provided in Appendix B. At each iteration of our MM procedure, the minimizer of the majorizing function (13) is the positive root of a quadratic polynomial equation, resulting in Algorithm 1. In Appendix B, we show that this majorant speeds up the MM scheme over Minka's [35].

4.2. Minimization step w.r.t assignment variable

Let the iteration number $\ell \in \mathbb{N}$ and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_k)_{1\le k\le K} \in ((0,+\infty)^K)^K$ be fixed. Because of the partition complexity term in (9), the direct minimization of the partial function with respect to $\boldsymbol{u}_n$, for every $n\in\mathbb{Q}$, is not closed form. Since the partition complexity penalty is concave, we propose to replace it by its linear upper-bound, leading us to minimize

$$
\begin{aligned}
\boldsymbol{u}_n \mapsto &-\sum_{k=1}^{K}u_{n,k}\ln(\mathrm{p}\,(\boldsymbol{z}_n\mid\boldsymbol{\alpha}_k))+\sum_{k=1}^{K}u_{n,k}\ln u_{n,k}\\
&-\frac{\lambda}{|\mathbb{Q}|}(\ln(\boldsymbol{\pi}^{(\ell+1)})+1)^\top(\boldsymbol{u}_n-\boldsymbol{u}_n^{(\ell)}),
\end{aligned}\tag{15}
$$

under the simplex and supervision constraints. In (15), $\boldsymbol{\pi}^{(\ell+1)} = (\pi_k^{(\ell+1)})_k$ is the vector whose k-th component is $\pi_k^{(\ell+1)} = \frac{1}{|\mathbb{Q}|}\sum_{n\in\mathbb{Q}}u_{n,k}^{(\ell)}$, and the log function operates componentwise.

Solving this minimization problem yields the updates, for every $n\in\mathbb{Q}$,

$$
\boldsymbol{u}_n^{(\ell+1)} = \mathrm{softmax}\left(\left(\ln\mathrm{p}\,(\boldsymbol{z}_n\mid\boldsymbol{\alpha}_k)+\frac{\lambda}{|\mathbb{Q}|}\ln(\pi_k^{(\ell+1)})\right)_k\right).
$$

Details for deriving this expression are given in Appendix C.

4.3. Global algorithm and class-assignment

Finally, given the estimation steps on the assignment variables $\boldsymbol{u} = (\boldsymbol{u}_n)_{n\in\mathbb{Q}}$ and on the Dirichlet parameters $(\boldsymbol{\alpha}_k)_{1\le k\le K}$ derived respectively in Sections 4.2 and 4.1, our complete procedure to tackle the minimization problem in (6) is detailed in Algorithm 2. We name it EM-Dirichlet as it shares close links to the EM algorithm, as it will be established in Proposition 1 in Section 5.

In the zero-shot scenario, the tasks at hand can be seen as a form of simplex clustering. There exists a straightforward method to map each cluster to a corresponding class label in an injective manner. Let $(\mathcal{C}_k)_{k\in\mathcal{K}}$ denote the set of non-empty clusters found with a clustering method, for instance ours, with $\mathcal{K}$ a subset of $k\in\{1,\ldots,K\}$ and for all $k\in\mathcal{K}$, $\mathcal{C}_k$ a subset of $\mathbb{Q}$. We proceed in the following way:

1. For each $k\in\mathcal{K}$, calculate the mean of cluster k, $\boldsymbol{m}_k = (m_{k,\ell})_{1\le\ell\le K} \in \Delta_K$, as $\boldsymbol{m}_k = \frac{1}{|\mathcal{C}_k|}\sum_{n\in\mathcal{C}_k}\boldsymbol{z}_n$. The element $m_{k,\ell}$ is interpreted as the probability that cluster k is associated with class ℓ. While it may seem intuitive to assign cluster k the class ℓ for which $m_{k,\ell}$ is maximal, this could lead to multiple clusters being assigned to the same class, which we wish to avoid.
2. Resolve the class-to-cluster assignments through a bipartite graph matching that maximizes $\sum_{k\in\mathcal{K}}\sum_{\ell=1}^{K}a_{k,\ell}m_{k,\ell}$ over all possible assignment matrices $\boldsymbol{A} = (a_{k,\ell}) \in \{0,1\}^{|\mathcal{K}|\times K}$ that satisfy $\boldsymbol{A}^\top\mathbf{1}_{|\mathcal{K}|} = \mathbf{1}_K$. This class assignment integer linear programming problem can be solved with algorithms such as [14]. An illustration of this process can be found in Appendix D.

Algorithm 1: MM-quadratic($\boldsymbol{u}_{\cdot,k}$, $\boldsymbol{\alpha}_k$)

Initialize $\boldsymbol{\alpha}_k^{(0)} = \boldsymbol{\alpha}_k$.
for $m = 0, 1, \ldots,$ **do**
  for $i \in \{1, \ldots, K\}$ **do**

$$b_{k,i} = \varphi'(\alpha_{k,i}^{(m)}) - (\ln \Gamma)' \left(\sum_{j=1}^{K} \alpha_{k,j}^{(m)} \right) - c(\alpha_{k,i}^{(m)})\alpha_{k,i}^{(m)} - \left(\sum_{n=1}^{N} u_{n,k} \right)^{-1} \sum_{n=1}^{N} u_{n,k} \ln(z_{n,i}).$$

$$\alpha_{k,i}^{(m+1)} = \left(-b_{k,i} + \sqrt{b_{k,i}^2 + 4c(\alpha_{k,i}^{(m)})} \right) \Big/ 2c(\alpha_{k,i}^{(m)}).$$

Algorithm 2: EM-Dirichlet

Initialize $\boldsymbol{u}^{(0)}$ as CLIP's probabilities and for all $k \in \{1, \ldots, N\}$, $\boldsymbol{\alpha}_k^{(0)} = \mathbf{1}_K$.
for $\ell = 0, 1, \ldots,$ **do**

```
// Update Dirichlet parameter for each class
```

$$\boldsymbol{\alpha}_k^{(\ell+1)} = \text{MM-quadratic}(\boldsymbol{u}_{\cdot,k}^{(\ell)}, \boldsymbol{\alpha}_k^{(\ell)}), \quad \forall k \in \{1, \ldots, K\},$$

```
// Update class proportions
```

$$\pi_k^{(\ell+1)} = \frac{1}{|\mathbb{Q}|} \sum_{n \in \mathbb{Q}} u_{n,k}^{(\ell)}, \quad \forall k \in \{1, \ldots, K\},$$

```
// Update assignment variable for all query samples
```

$$\boldsymbol{u}_n^{(\ell+1)} = \text{softmax} \left(\left(\ln \mathrm{p} \left(\boldsymbol{z}_n \mid \boldsymbol{\alpha}_k^{(\ell+1)} \right) + \frac{\lambda}{|\mathbb{Q}|} \ln(\pi_k^{(\ell+1)}) \right)_k \right), \quad \forall n \in \mathbb{Q}.$$

5. Links with other clustering and transductive few-shot objectives

The general log-likelihood model fitting objective in (7), also referred to as probabilistic K-means [8, 26], is well-established in the clustering literature. Indeed, it is a generalization of the ubiquitous K-means, which corresponds to the particular choice of the Gaussian distribution for the densities in (7), with covariance matrices fixed to the identity matrix. This general objective has a strong, inherent bias towards K-balanced partitions, a theoretically well-established fact in the clustering literature [8, 26]. To mitigate this bias and address realistic, potentially imbalanced few-shot query sets, the recent transductive few-shot method in [34] coupled the MDL term in (9) with the standard K-means objective. This corresponds to the general data-fitting function we tackle in (7), but with the likelihood densities assumed to be Gaussian. As we will see in our experiments (Table 2), the non-trivial deployment of the Dirichlet model is crucial, outperforming significantly [34] in CLIP's few-shot setting. Furthermore, we show in the following an interesting result, which connects the general unbiased clustering problem we propose in (6), to the well-known Expectation-Maximization (EM) algorithm for mixture models [3, p.438]. Indeed, optimizing the objective in (6) could be viewed as a generalization of EM, enabling to control the class-balance parameter λ.

Proposition 1. *Consider the unsupervised classification problem, i.e. $\mathbb{S} = \varnothing$. Suppose the value of λ in (6) is set to the size of the query set, i.e., $\lambda = |\mathbb{Q}|$. Then Algorithm 2 is equivalent to the EM algorithm when applied to a generic mixture model*

$$\mathrm{p}\left(\boldsymbol{z}_n \mid \boldsymbol{\pi}, \boldsymbol{\alpha}\right) = \sum_{k=1}^{K} \pi_k \mathrm{p}\left(\boldsymbol{z}_n \mid \boldsymbol{\alpha}_k\right), \tag{16}$$

where $\boldsymbol{\pi} = (\pi_k)_{1 \leq k \leq K} \in \Delta_K$ are the mixture coefficients.

The proof of Proposition 1 is given in Appendix E.

6. Experiments

We evaluated our method on 11 publicly accessible image classification datasets which were also utilized in CLIP [40]: ImageNet [42], Caltech101 [16], OxfordPets [39], StanfordCars [27], Flowers102 [37], Food101 [5], FGVCAircraft [33], SUN397 [55], DTD [12], EuroSAT [20] and UCF101 [46]. To ensure reproducibility, we adhere to the dataset splits provided by CoOp [57] and use the prompts employed in TIP-Adapter [56]. All experiments are conducted using CLIP's pre-trained ResNet50 visual encoder. The temperature in the probabilities (2) is fixed to $T = 30$.

6.1. Zero-shot

Tasks generation For generating query sets in our transductive zero-shot setting, we employ a practical approach that maintains manageable batch sizes. At each new task (mini-batch), we randomly select the classes that will be represented in the query set, with the actual number of distinct classes ranging from 3 to 10, also selected at random. It is important to note that the set of classes occurring in each batch remain undisclosed, and vary randomly from one batch to another, ensuring that the clustering task is still performed over all K potential classes present in the whole dataset. Subsequently, we randomly select $|\mathbb{Q}| = 75$ images in to the chosen classes to constitute the query set. During transductive inference, the query set of each task is treated independently of the other randomly sampled tasks.

Comparative methods We conduct a comparative evaluation of our clustering methodology, EM-Dirichlet, and its variant utilizing hard assignments, denoted as Hard EM-Dirichlet, against a range of clustering objective functions and algorithms: Hard and soft K-means [32, p.286], EM for Gaussian mixtures with identity covariance (EM-Gaussian (cov. Id)) and with diagonal covariance (EM-Gaussian (cov. diag)) [3, p.438], and Hard KL K-Means [9]. Furthermore, our comparison provides a full ablation study of the terms in general objective function (6):

1. The log-likelihood model fitting term (7), which varies across Gaussian (employed in Hard K-means, Soft K-means, EM-Gaussian), and Dirichlet (in our method).
2. The entropic barrier (8) featured in both Soft K-means and the EM-based approaches.
3. The MDL partition-complexity term (9), incorporated exclusively in the EM methods.

Initialization is uniform across different clustering techniques, utilizing CLIP's predictions from Equation (2). In all EM-based methods, the regularization parameter λ is set according to $\lambda = \frac{5}{K}|\mathbb{Q}|$, to maintain consistency across comparisons.

Results We assess the clustering methods on zero-shot tasks, using both the visual embeddings and the combined text-vision feature vectors. We also include the zero-shot classification results from CLIP. In Table 1, we report average accuracy over 1,000 tasks using the graph cluster-to-classes assignment described in Section 4.3. Table 1 conveys several crucial messages:

- Clustering visual embeddings alone does not suffice to surpass inductive CLIP's zero-shot performance. Incorporating textual information via probability features enhances the performance, even for methods initially designed for Gaussian distributions.
- Gaussian-based data-fitting approaches are sub-optimal for simplex clustering. Replacing the Gaussian metric with a Kullback-Leibler divergence is beneficial. Employing a Dirichlet data-fitting term within the EM framework significantly improves the results compared to EM-Gaussian methods, highlighting the necessity of accurate data distribution modeling.
- Introducing the partition complexity term (in the EM methods), which discourages overly balanced predictions, proves advantageous for the performance.
- Using an adapted transductive model like Hard EM-Dirichlet, accuracy improves considerably, showing a 9% rise across 11 datasets, and nearly 20% on ImageNet.

In Appendix F, we show that zero-shot performance improves with larger query set sizes, indicating enhanced transduction efficiency with increasing mini-batch size.

6.2. Few-shot

Task generation We follow the realistic transductive few-shot evaluation protocol proposed recently in [34]. Specifically, the query sets are constructed with a fixed number of effective classes $k_{\text{eff}} = 5$, from which $|\mathbb{Q}|$ samples are randomly selected. This approach aligns with established few-shot protocols in the literature [31, 45, 52]. These classes remain undisclosed during inference, ensuring the task is a K-way classification. The support set is created by uniformly selecting s images from each of the K classes. The ensuing results are derived performing few-shot tasks with 1, 2, 4, 8, and 16 shots. During inference on the test set, the size of the query set is set to $|\mathbb{Q}| = 75$, while for validation, the size is reduced to$|\mathbb{Q}| = 35$ due to data limitations.

Hyper-parameters Parameter λ in EM-Dirichlet is set to the fixed value $\lambda = \frac{k_{\text{eff}}}{K}|\mathbb{Q}|$. Methods with tunable hyper-parameters are fine-tuned using the validation split provided with each dataset. In line with [22], validation is performed on five s-shot tasks across all datasets and for every shot number. These tasks, crafted as previously detailed, use support and query instances drawn from the validation set. The hyper-parameters are then optimized through a grid search to maximize accuracy on the validation set.

Results We evaluate the accuracy of our proposed transductive methods, EM-Dirichlet and Hard EM-Dirichlet, against several recent transductive few-shot methods, including BD-CSPN [30], Laplacian Shot [60], α-TIM [52], and PADDLE [34]. Additionally, we benchmark against two inductive few-shot methods designed for CLIP: TIP-Adapter [56] and CoOp [58]. The results, averaged across $1,000$ tasks with 4 shots, are presented in Table 2 and for the other number of shots in Appendix G.

Our method surpasses competing approaches on the majority of datasets, with a more pronounced advantage observed on challenging datasets that have a large number of

		Food101	EuroSAT	DTD	OxfordPets	Flowers102	Caltech101	UCF101	FGVC Aircraft	Stanford Cars	SUN397	ImageNet	**Average**
	Zero-shot CLIP	77.1	36.5	42.9	85.1	66.1	84.4	61.7	17.1	55.8	58.6	58.3	58.5
Vis. embs.	Hard K-means	52.2	37.9	40.0	54.5	44.8	62.9	49.2	14.3	22.4	39.6	29.3	40.6
	Soft K-means	17.6	29.9	19.1	40.4	36.1	21.3	13.3	10.6	11.1	9.1	10.1	19.8
	EM-Gaussian (Id cov.)	14.0	14.5	9.4	6.9	5.3	30.3	7.4	1.9	2.5	5.3	3.9	8.3
	EM-Gaussian (diag cov.)	51.4	**40.6**	37.5	59.2	45.6	61.8	47.2	13.6	24.3	35.1	28.3	40.4
Probabilities	Hard K-means	49.5	35.2	38.7	62.4	44.5	52.2	46.6	14.5	29.6	41.4	31.0	40.5
	Soft K-means	41.8	21.5	18.3	56.6	34.3	50.5	30.2	7.2	34.8	18.8	19.1	30.3
	EM-Gaussian (Id cov.)	21.4	14.5	16.5	21.1	23.1	33.6	19.3	6.8	18.5	18.7	19.1	19.3
	EM-Gaussian (diag cov.)	63.3	33.1	38.7	71.1	51.1	66.6	56.0	16.5	46.9	54.8	48.5	49.7
	Hard KL K-means	72.2	34.9	40.8	73.0	61.1	72.0	60.6	17.7	56.2	61.8	61.0	55.6
	EM-Dirichlet	88.2	33.0	47.7	87.3	71.5	88.4	69.0	19.2	65.5	77.3	76.9	65.8
	Hard EM-Dirichlet	**90.2**	36.1	**49.3**	**90.9**	**73.1**	**89.7**	**70.3**	**20.4**	**67.7**	**78.5**	**77.6**	**67.6**

Table 1. Average accuracy of clustering methods over 1,000 zero-shot classification tasks. Inference is performed both on the visual embeddings and on the text-vision probability features.

		Food101	EuroSAT	DTD	OxfordPets	Flowers102	Caltech101	UCF101	FGVC Aircraft	Stanford Cars	SUN397	ImageNet	**Average**	**Time** (s)
Ind.	Tip-Adapter [56]	76.7	**72.5**	54.7	86.4	83.2	88.8	72.1	23.7	63.9	66.7	62.7	68.3	6.76×10^{0}
	CoOp [58]	76.3	63.2	52.2	86.2	81.0	87.7	67.0	22.2	61.3	63.4	59.9	65.5	3.35×10^{3}
Trans.	BDCSPN [30]	74.7	46.1	45.2	81.3	74.2	82.0	59.0	18.0	48.1	54.5	49.2	57.5	4.49×10^{-1}
	Laplacian Shot [60]	76.6	53.0	52.6	88.4	85.5	86.8	67.0	22.2	60.4	63.8	56.3	64.8	2.10×10^{-1}
	α-TIM [52]	66.1	46.1	45.3	87.1	79.1	83.3	59.4	20.4	53.4	53.4	42.7	57.8	1.65×10^{0}
	PADDLE [34]	71.8	45.9	50.0	84.7	82.3	81.9	63.7	21.3	56.1	60.6	52.1	60.9	4.04×10^{-1}
	EM-Dirichlet	**88.7**	50.8	**62.6**	**92.5**	**91.3**	**90.1**	**76.1**	**24.9**	**73.5**	**80.9**	**78.4**	**73.6**	1.04×10^{0}
	Hard EM-Dirichlet	87.9	50.8	60.5	91.7	90.5	89.8	75.3	24.2	72.6	80.2	78.3	72.9	6.97×10^{-1}

Table 2. Evaluation of our approach against two benchmarks – 1) inductive methods specifically designed for few-shot classification using CLIP, and 2) transductive few-shot methods applied to probability feature vector classification. The analysis encompasses 1,000 distinct 4 shots tasks. We also report average execution time for a single task, computed over 1,000 tasks, on the ImageNet dataset.

classes, such as SUN397 and ImageNet. The accuracy gap between our method and the inductive ones shows the benefits of transductive inference. On the other hand, the inferior perfomance of other transductive methods can be attributed to their lack of adaptability to simplex classification.

Interestingly, our results indicate that on some datasets such as Food101, our method perform better in the zero-shot than in the few-shot setting. This is consistent with Radford et al. [40], suggesting that few labeled examples can negatively impact classification, possibly due to outliers or ambiguous examples in the support set.

Lastly, we observe that inductive methods outperform ours on the EuroSAT dataset. This might be due to the inclusion of text information in the vision-text features. While typically advantageous, it is possible that the text information introduces a confounding effect specific to this dataset.

7. Conclusion

In conclusion, our study expands transductive inference to vision-language models like CLIP, previously unexplored in this domain. We demonstrate that the transductive methodology can boost image classification accuracy, including in zero-shot scenarios. Future work could apply our transductive CLIP approach to other tasks like segmentation and out-of-distribution detection.

References

[1] Muhammad Ali, Junbin Gao, and Michael Antolovich. Parametric classification of Bingham distributions based on Grassmann manifolds. *IEEE Transactions on Image Processing*, 28(12):5771–5784, 2019. 3

[2] Heinz H. Bauschke and Patrick Louis Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer New York, 2nd, corrected printing edition, 2019. 13

[3] Christopher M. Bishop and Nasser M. Nasrabadi. *Pattern Recognition and Machine Learning*. Springer, 2006. 6, 7

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003. 5

[5] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision*, pages 446–461, Zurich, Switzerland, 2014. Springer. 6

[6] Malik Boudiaf, Imtiaz Ziko, Jérôme Rony, José Dolz, Pablo Piantanida, and Ismail Ben Ayed. Information maximization for few-shot learning. *Advances in Neural Information Processing Systems*, 33:2445–2457, 2020. 1, 2

[7] Malik Boudiaf, Etienne Bennequin, Myriam Tami, Antoine Toubhans, Pablo Piantanida, Céline Hudelot, and Ismail Ben Ayed. Open-set likelihood maximization for few-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[8] Y. Boykov, H. N. Isack, C. Olsson, and I. Ben Ayed. Volumetric bias in segmentation and reconstruction: Secrets and solutions. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015. 4, 6

[9] Jie Cao, Zhiang Wu, Junjie Wu, and Hui Xiong. SAIL: Summation-bAsed Incremental Learning for information-theoretic text clustering. *IEEE Transactions on Cybernetics*, 43(2):570–584, 2013. 7

[10] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Prompt learning with optimal transport for vision-language models. In *International Conference on Learning Representations*, 2022. 2

[11] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations*, 2019. 2

[12] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 6

[13] Patrick L. Combettes and Jean-Christophe Pesquet. Deep neural network structures solving variational inequalities. *Set-Valued and Variational Analysis*, 28(3):491–518, 2020. 13

[14] David F Crouse. On implementing 2D rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4):1679–1696, 2016. 5

[15] Hakan Erdogan and Jeffrey A. Fessler. Monotonic algorithms for transmission tomography. In *2002 5th IEEE EMBS International Summer School on Biomedical Imaging*, pages 14–pp. IEEE, 2002. 12

[16] Li Fei-Fei, Robert Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006. 6

[17] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, pages 1–15, 2023. 1, 2

[18] Onur C Hamsici and Aleix M Martinez. Spherical-homoscedastic distributions: The equivalency of spherical and normal distributions in classification. *Journal of Machine Learning Research*, 8(7), 2007. 3

[19] Md Abul Hasnat, Julien Bohné, Jonathan Milgram, Stéphane Gentric, and Liming Chen. Von Mises-Fisher mixture model-based deep learning: Application to face verification. *arXiv preprint arXiv:1706.04264*, 2017. 3

[20] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 6

[21] Z hong, D Friedman, and D Chen. Factual probing is [mask]: Learning vs. learning to recall. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2021. 2

[22] Shell Xu Hu, Da Li, Jan Stühmer, Minyoung Kim, and Timothy M Hospedales. Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9068–9077, 2022. 2, 7

[23] Jonathan Huang. Maximum likelihood estimation of Dirichlet distribution parameters. Technical report, Carnegie Mellon University, 2005. 4

[24] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 1

[25] Thorsten Joachims. Transductive inference for text classification using support vector machines. In *International Conference on Machine Learning*, pages 200–209, 1999. 1

[26] M. Kearns, Y. Mansour, and A. Ng. An information-theoretic analysis of hard and soft assignment methods for clustering. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 1997. 6

[27] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 6

[28] Michalis Lazarou, Tania Stathaki, and Yannis Avrithis. Iterative label cleaning for transductive and semi-supervised few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8751–8760, 2021. 2

[29] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations*, 2022. 1

[30] Jinlu Liu, Liang Song, and Yongqiang Qin. Prototype rectification for few-shot learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 741–756. Springer, 2020. 1, 2, 7, 8

[31] Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. Learning to propagate labels: Transductive propagation network for few-shot learning. In *International Conference on Learning Representations*, 2019. 1, 2, 7

[32] David J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003. 4, 7

[33] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, Oxford University, 2013. 6

[34] Ségolène Martin, Malik Boudiaf, Emilie Chouzenoux, Jean-Christophe Pesquet, and Ismail Ayed. Towards practical few-shot query sets: Transductive minimum description length inference. *Advances in Neural Information Processing Systems*, 35:34677–34688, 2022. 1, 2, 4, 6, 7, 8

[35] Thomas Minka. Estimating a Dirichlet distribution. Technical report, MIT, 2000. 4, 5, 12

[36] A. Narayanan. Algorithm AS 266: maximum likelihood estimation of the parameters of the dirichlet distribution. *Applied Statistics*, 40(2):365–374, 1991. 4

[37] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Fraphics & Image Processing*, pages 722–729. IEEE, 2008. 6

[38] Samyajoy Pal and Christian Heumann. Clustering compositional data using Dirichlet mixture model. *PLOS One*, 17(5): e0268438, 2022. 5

[39] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 6

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 6, 8

[41] Fabrice Rossi and Florian Barbaro. Mixture of von Mises-Fisher distribution with sparse prototypes. *Neurocomputing*, 501:41–74, 2022. 3

[42] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015. 6

[43] Tyler R Scott, Andrew C Gallagher, and Michael C Mozer. Von Mises-Fisher loss: An exploration of embedding geometries for supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10612–10622, 2021. 3

[44] T Shin, Logan R. L. IV Razeghi, Y, E Wallace, and S Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2020. 2

[45] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, 30, 2017. 7

[46] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. Technical report, Center for Research in Computer Vision, University of Central Florida, 2012. arXiv:1212.0402. 6

[47] Ran Tao, Hao Chen, and Marios Savvides. Boosting transductive few-shot fine-tuning with margin-based uncertainty weighting and probability regularization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15752–15761, 2023. 1, 2

[48] Long Tian, Jingyi Feng, Xiaoqiang Chai, Wenchao Chen, Liming Wang, Xiyang Liu, and Bo Chen. Prototypes-oriented transductive few-shot learning with conditional transport. In *IEEE International Conference on Computer Vision (ICCV)*, pages 16317–16326, 2023. 1, 2

[49] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B. Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *European Conference on Computer Vision (ECCV)*, 2020. 2

[50] Daniel J. Trosten, Rwiddhi Chakraborty, Sigurd Løkse, Kristoffer Knutsen Wickstrøm, Robert Jenssen, and Michael C. Kampffmeyer. Hubs and hyperspheres: Reducing hubness and improving transductive few-shot learning with hyperspherical embeddings. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7527–7536, 2023. 1, 2

[51] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999, 1999. 1

[52] Olivier Veilleux, Malik Boudiaf, Pablo Piantanida, and Ismail Ben Ayed. Realistic evaluation of transductive few-shot learning. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 2, 7, 8

[53] Yan Wang, Wei-Lun Chao, Kilian Q. Weinberger, and Laurens Van der Maaten. Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv preprint arXiv:1911.04623*, 2019. 2

[54] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. MedCLIP: Contrastive learning from unpaired medical images and text. *Conference on Empirical Methods in Natural Language Processing*, 2022. 1

[55] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492. IEEE, 2010. 6

[56] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-

adapter: Training-free adaption of clip for few-shot classification. In *European Conference on Computer Vision (ECCV)*, pages 493–510. Springer, 2022. 1, 2, 3, 6, 7, 8

[57] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 1, 2, 3, 6

[58] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. In *International Journal of Computer Vision*, pages 2337–2348. Springer, 2022. 1, 2, 7, 8

[59] Hao Zhu and Piotr Koniusz. Transductive few-shot learning with prototype-based label propagation by iterative graph refinement. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23996–24006, 2023. 1, 2

[60] Imtiaz Ziko, Jose Dolz, Eric Granger, and Ismail Ben Ayed. Laplacian regularized few-shot learning. In *International Conference on Machine Learning*, pages 11660–11670. PMLR, 2020. 1, 2, 7, 8