Similarity of Classification Tasks

Cuong Nguyen
Australian Institute for Machine Learning
University of Adelaide, Australia
cuong.nguyen@adelaide.edu.au

Thanh-Toan Do*
Faculty of Information Technology
Monash University, Australia
toan.do@monash.edu

Gustavo Carneiro
Australian Institute for Machine Learning
University of Adelaide, Australia
gustavo.carneiro@adelaide.edu.au

Abstract

Recent advances in meta-learning has led to remarkable performances on several few-shot learning benchmarks. However, such success often ignores the similarity between training and testing tasks, resulting in a potential bias evaluation. We, therefore, propose a generative approach based on a variant of Latent Dirichlet Allocation to analyse task similarity to optimise and better understand the performance of meta-learning. We demonstrate that the proposed method can provide an insightful evaluation for meta-learning algorithms on two few-shot classification benchmarks that matches common intuition: the more similar the higher performance. Based on this similarity measure, we propose a task-selection strategy for meta-learning and show that it can produce more accurate classification results than methods that randomly select training tasks.

1 Introduction

The vast development in machine learning has enabled possibilities to solve increasingly complex applications. Such complexity require high capacity models, which in turn need a massive amount of annotated data for training, resulting in an arduous, costly and even infeasible annotation process. This has, therefore, motivated the research of novel learning approaches, generally known as transfer learning, that exploit past experience (in the form of models learned from other training tasks) to quickly learn a new task using relatively small training sets.

Transfer-learning, and in particular, meta-learning, has recently achieved state-of-the-art results in several few-shot learning benchmarks (Vinyals et al., 2016; Snell et al., 2017; Finn et al., 2017; Yoon et al., 2018; Rusu et al., 2019). Such success depends not only on the effectiveness of transfer learning algorithms, but also on the similarity between training and testing tasks (Y. Chen et al., 2020). More specifically, the larger the subset of training tasks that are similar to the testing tasks, the higher the classification accuracy on those testing tasks. However, meta-learning methods are assessed without taking into account such observation, which can bias the meta-learning classification results depending on the policy for selecting training and testing tasks.

In this paper, we propose a generative approach based on a "continuous" version of Latent Dirichlet Co-Clustering to model classification tasks. The resulting model represents tasks in a latent "task-theme" simplex, and hence, allows to quantitatively measure their similarity. The proposed similarity measure enables the possibility of selecting the most related tasks from the training set for the meta-

---

*Work done while at the University of Liverpool, UK

learning of a novel testing task. We empirically demonstrate that the proposed task selection strategy outperforms the one that randomly selects training tasks across several meta-learning methods.

2 Related work

With this paper, we target an improved understanding of meta-learning algorithms (W.-Y. Chen et al., 2019; Dhillon et al., 2020), which can allow us to improve their current performance. Although meta-learning has progressed steadily with many remarkable achievements, it has been reported that there is a large variance of performance among testing tasks (Dhillon et al., 2020). This observation suggests that not all testing tasks are equally related to training tasks. "Task hardness" which is based on the cosine similarity between the embedding of labelled and unlabelled data is, therefore, proposed to better justify the performance of meta-learning methods (Dhillon et al., 2020). This, however, quantifies only the similarity between samples within a task without investigating the similarity between tasks.

Task similarity has been intensively studied in the field of multi-task learning. Some remarkable works include task-clustering using k-nearest neighbours (Thrun and O'Sullivan, 1996), modelling common prior between tasks as a mixture of distributions (Bakker and Heskes, 2003) with the extension using Dirichlet Process (Xue et al., 2007), applying a convex formulation to either cluster (Jacob et al., 2009) or learning task relationship through task covariance matrices (Zhang and Yeung, 2012). Other approaches try to provide theoretical guarantees when learning the similarity or relationship between tasks (Shui et al., 2019). Following a similar approach, an extensive experiment was carried out on 26 computer-vision tasks to determine the correlation between those tasks, also known as "taskonomy" (Zamir et al., 2018). Some recent works (Tran et al., 2019; C. V. Nguyen et al., 2020) take a slightly different approach by investigating the correlation of the label distribution between those tasks of interest. One commonality of those studies is their reliance on a discriminative approach, where the similarity of task-specific classifiers are used to quantify task relatedness. In addition, most of those works focus more on the conventional machine learning setting, which requires a sufficient number of labelled data on the novel tasks to perform transfer learning. In contrast, our proposal follows a generative approach which does not depend on any task-specific classifier. Our approach can also work in the few-shot setting, where only a few labelled data points from the targeted tasks are available. Another work that is slightly related to task similarity is Task2Vec (Achille et al., 2019), which employs Fisher information matrix of an external network, known as "probe" network, to model visual tasks as fixed vectors in an embedding space, allowing to analyse and calculate task similarity. However, its application is still limited due to the need of an external network pre-trained to perform specific tasks on some standard visual data sets.

Our work is also related to finite mixture models (Pritchard et al., 2000), such as the Latent Dirichlet Allocation (LDA) (Blei et al., 2003), in topic modelling which analyses and summarises text data, or in computer vision (Li and Perona, 2005). LDA assumes that each document within a given corpus can be represented as a finite mixture model, where its components are the latent topics shared across all documents. Training an LDA model or its variants on a large text corpus is challenging, so several approximate inference techniques have been proposed, ranging from mean-field variational inference (VI) (Blei et al., 2003), collapsed Gibbs' sampling (Griffiths and Steyvers, 2004) and collapsed VI (Teh et al., 2007). Furthermore, several online inference methods have been developed to increase the training efficiency for large corpora (Canini et al., 2009; Hoffman et al., 2010; Foulds et al., 2013). Our work is slightly different from the inference for conventional LDA models, where we perform online learning for Latent Dirichlet Co-clustering (Shafiei and Milios, 2006) – a variant of LDA – that includes the information of paragraphs into the model. In addition, our approach considers "word" as continuous data, instead of the discrete data represented by a bag-of-word vector generally used in topic modelling.

3 Method

To relate image **classification** to **topic modelling**, we consider **a task** as **a document**, **a class** as **a paragraph**, and **an image** as **a word**. Given these analogies, we employ the Latent Dirichlet Co-clustering (LDCC) (Shafiei and Milios, 2006) – a variant of LDA – to model classification tasks. The LDCC extends the conventional LDA to a hierarchical structure by including the information about paragraphs, or in our case, data classes, into the model. Since the data in classification is
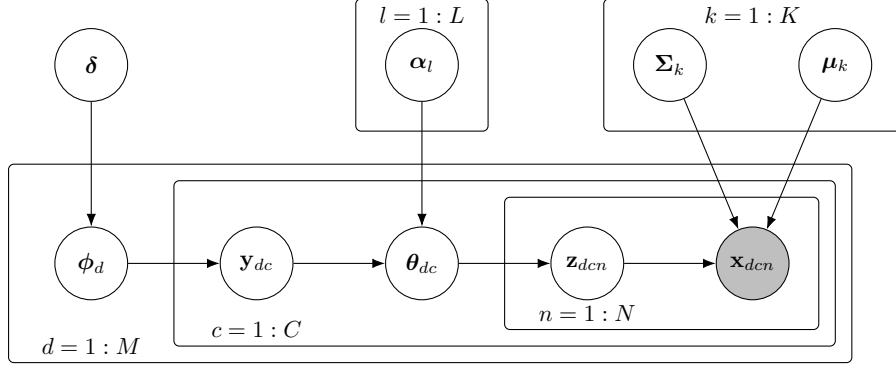
Figure 1: Directed acyclic graph represents the continuous LDCC that models classification tasks as a finite mixture of Gaussian distributions.

assumed to be continuous, the categorical word-topic distribution in the original LDCC model is replaced by a Gaussian **image-theme** distribution. Each classification task can be modelled as a mixture of L **task-themes** (corresponding to document topic in LDCC), where each task-theme is a "summary" of many finite mixtures of K **image-themes**. We can, therefore, utilise this representation, and in particular the *task-theme mixture* parameter to quantify the similarity between tasks.

We assume that there are M classification tasks, where each task consists of C classes, and each class has N images (i.e., using meta-learning nomenclature, this represents M C-way N-shot classification tasks). For simplicity, C and N are assumed to be fixed across all tasks, but the extension of varying C and N is trivial and can be implemented straightforwardly. The process to generate classification tasks from an "L-task-K-image theme" model shown in Figure 1 can be presented as follows:

- Initialise means and precision matrices of K Gaussian image-theme $\{\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k\}_{k=1}^{K}$, where $\boldsymbol{\mu}_k \in \mathbb{R}^D$, and $\boldsymbol{\Lambda}_k \in \mathbb{R}^{D \times D}$ is positive definite matrix
- For task d-th in the collection of M tasks:
  - Choose a task-theme mixture: $\boldsymbol{\phi}_d \sim \text{Dirichlet}_L(\boldsymbol{\phi}; \boldsymbol{\delta})$
  - For the c-th class in the d-th task:
    * Choose a task-theme assignment: $\mathbf{y}_{dc} \sim \text{Categorical}(\mathbf{y}; \boldsymbol{\phi}_d)$
    * Choose an image-theme mixture: $\boldsymbol{\theta}_{dc} \sim \text{Dirichlet}_K(\boldsymbol{\theta}; \boldsymbol{\alpha}_l)$, where $y_{dcl} = 1$
    * For image n-th in class c-th of task d-th:
      · Choose an image-theme assignment: $\mathbf{z}_{dcn} \sim \text{Categorical}(\mathbf{z}; \boldsymbol{\theta}_{dc})$
      · Choose an image: $\mathbf{x}_{dcn} \sim \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})$, where: $z_{dcnk} = 1$.

If the K Gaussian image-themes $\{(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)\}_{k=1}^{K}$, and the Dirichlet concentration $\{\boldsymbol{\alpha}\}_{l=1}^{L}$ for each class are known, we can infer the mixture parameter $\boldsymbol{\phi}_d$ based on the observed images $\mathbf{x}_d$ of any arbitrary task d-th to represent that task in the latent task-theme simplex. This representation enables the possibility of performing further analysis, such as measuring distances between tasks. Hence, our objective is to learn these parameters from the M given classification tasks. In short, our objective is to maximise log-likelihood:

$$\max_{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}} \ln p(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}). \tag{1}$$

Due to the complexity of the graphical model with latent variables as shown in Figure 1, the inference for the likelihood in (1) is intractable, and therefore, the estimation must rely on approximate inference. Current methods to approximate the posterior of LDA-based models fall into two main categories: sampling (Griffiths and Steyvers, 2004; Canini et al., 2009) and optimisation (Blei et al., 2003; Teh et al., 2007). Each approach has strengths and weaknesses, where the choice mostly depends on the application of interest. For the problem of task similarity where M is very large, the optimisation approach, and in particular, the mean-field VI, is preferable due to its efficiency and scalability to large data sets. In this paper, VI is used to infer the parameters of interest.

The log-likelihood of interest can be lower-bounded by Jensen's inequality. The lower-bound is often known as evidence lower-bound (ELBO) and can be expressed as:

$$\mathsf{L} = \mathbb{E}_q\left[\ln p(\mathbf{x}, \boldsymbol{\phi}, \mathbf{y}, \boldsymbol{\theta}, \mathbf{z}|\boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\right] - \mathbb{E}_q\left[q(\boldsymbol{\phi}, \mathbf{y}, \boldsymbol{\theta}, \mathbf{z})\right]. \tag{2}$$

Following the conventional variational inference for LDA (Blei et al., 2003), we choose a fully factorised variational distribution q as our variational posterior:

$$q(\boldsymbol{\phi}, \mathbf{y}, \boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{d=1}^{M} q(\boldsymbol{\phi}_d; \boldsymbol{\lambda}_d) \prod_{c=1}^{C} q(\mathbf{y}_{dc}; \boldsymbol{\eta}_{dc})\, q(\boldsymbol{\theta}_{dc}; \boldsymbol{\gamma}_{dc}) \prod_{n=1}^{N} q(\mathbf{z}_{dcn}; \mathbf{r}_{dcn}), \tag{3}$$

where:

$$q(\boldsymbol{\phi}_d; \boldsymbol{\lambda}_d) = \mathrm{Dirichlet}_L\left(\boldsymbol{\phi}_d; \boldsymbol{\lambda}_d\right) \qquad q(\mathbf{y}_{dc}; \boldsymbol{\eta}_{dc}) = \mathrm{Categorical}\left(\mathbf{y}_{dc}; \boldsymbol{\eta}_{dc}\right)$$
$$q(\boldsymbol{\theta}_{dc}; \boldsymbol{\gamma}_{dc}) = \mathrm{Dirichlet}_K\left(\boldsymbol{\theta}_{dc}; \boldsymbol{\gamma}_{dc}\right) \qquad q(\mathbf{z}_{dcn}; \mathbf{r}_{dcn}) = \mathrm{Categorical}\left(\mathbf{z}_{dcn}; \mathbf{r}_{dcn}\right).$$

Given the variational distribution q defined in Eq. (3), we can rewrite the ELBO as:

$$\begin{aligned} \mathsf{L} = \mathbb{E}_q\,[&\ln p(\mathbf{x}|\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \ln p(\mathbf{z}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\alpha}) + \ln p(\mathbf{y}|\boldsymbol{\phi}) + \ln p(\boldsymbol{\phi}|\boldsymbol{\delta}) \\ &- \ln q(\mathbf{z}) - \ln q(\boldsymbol{\theta}) - \ln q(\mathbf{y}) - \ln q(\boldsymbol{\phi})]. \end{aligned} \tag{4}$$

Comparing to the conventional LDA (Blei et al., 2003, Eq. (14)), the ELBO in Eq. (4) contains 4 extra terms highlighted in violet. The presence of those terms are due to the hierarchical structure of LDCC that takes the factor of classes (analogous to paragraphs) into the model.

Instead of maximising likelihood, we maximise its lower-bound, resulting in an alternative objective function:

$$\max_{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}} \max_{\mathbf{r}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\lambda}} \mathsf{L}. \tag{5}$$

Given the usage of prior conjugate, all of the terms in the ELBO can be evaluated straightforwardly (please refer to Supplementary Material A). The optimisation is based on gradient, and performed in two steps, resulting in a process analogous to the expectation-maximisation (EM) algorithm. In the E-step, the task-specific variational-parameters $\mathbf{r}, \boldsymbol{\gamma}, \boldsymbol{\eta}$ and $\boldsymbol{\lambda}$ are iteratively updated, while holding the meta-parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}$ fixed. In the M-step, the meta-parameters are updated using the values of the task-specific variational-parameters obtained in the E-step. The inference for the meta image-themes are similar to the estimation of Gaussian mixture model (Bishop, 2006, Chapter 9). Please refer to Supplementary Material B for more details.

Conventionally, the iterative updates in the E-step and M-step require a full pass through the entire collection of tasks. This is, however, very slow and even infeasible since M is often in the magnitude of millions. We, therefore, propose an online VI inspired by the online learning for LDA (Hoffman et al., 2010) to infer the image-themes. When the d-th task is observed, we perform EM to obtain the "task-specific" image-themes (denoted by a tilde on top of variables) that are locally optimal for that task. The "meta" image-themes of interest are then updated as a weighted average between their previous values and the "task-specific" values:

$$\boldsymbol{\mu} \leftarrow (1 - \rho_d)\boldsymbol{\mu} + \rho_d\tilde{\boldsymbol{\mu}}, \quad \boldsymbol{\Sigma} \leftarrow (1 - \rho_d)\boldsymbol{\Sigma} + \rho_d\tilde{\boldsymbol{\Sigma}}, \quad \boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} - \rho_d\mathbf{H}^{-1}\mathbf{g}, \tag{6}$$

where $\rho_d = (\tau_0 + d)^{-\tau_1}$ with $\tau_0 \geq 0$ and $\tau_1 \in (0.5, 1]$ (Hoffman et al., 2010), and $\mathbf{g}$ is the gradient of L w.r.t. $\boldsymbol{\alpha}$, and $\mathbf{H}$ is the Hessian matrix. Please refer to Supplementary Material C for the details of the online learning algorithm.

Also, instead of updating the image-themes when observing a single task, we use multiple or a mini-batch of tasks to reduce noise. The mini-batch version requires a slight modification, where we calculate the average of all "task-specific" image-themes for the tasks in the same mini-batch, and use that as the "task-specific" value to update the corresponding "meta" image-theme.

Given the image-themes $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$ and the Dirichlet parameter $\{\boldsymbol{\alpha}_l\}_{l=1}^{L}$, we can represent a task by its variational Dirichlet posterior of the task-topic mixing coefficients $q(\boldsymbol{\phi}_d; \boldsymbol{\lambda}_d)$ in the latent task-theme simplex. This new representation of classification tasks has two advantages comparing to the recently proposed task representation Task2Vec (Achille et al., 2019): (i) it does not need any pre-trained networks, and (ii) the use of probability distribution, instead of a single value

vector as in Task2Vec, allowing to include modelling uncertainty when representing tasks. In addition, we can utilise this representation to quantitatively analyse the similarity between two tasks through a divergence between $q(\boldsymbol{\phi}_d; \boldsymbol{\lambda}_d)$. Commonly, symmetric distances, such as Jensen-Shannon divergence, Hellinger distance, or earth's mover distance are employed to calculate the divergence between distributions. However, it is argued that similarity should be represented as an asymmetric measure (Tversky, 1977). This is reasonable in the context of transfer learning, since knowledge gained from learning a difficult task might significantly facilitate the learning of an easy task, but the reverse might not always have the same level of effectiveness. In light of asymmetric distance, we decide to use Kullback-Leibler (KL) divergence, denoted as $D_{\mathrm{KL}}[.\|.]$. As $D_{\mathrm{KL}}[P\|Q]$ is defined as the information lost when using a code optimised for Q to encode the samples of P, we, therefore, calculate $D_{\mathrm{KL}}[q(\boldsymbol{\phi}_d; \boldsymbol{\lambda}_{M+1})\|q(\boldsymbol{\phi}_d; \boldsymbol{\lambda}_d)]$, where $d \in \{1, \ldots, M\}$, to assess how the d-th training task differs from the learning of the novel $(M+1)$-th task.

Correlation Diagram   We define a correlation diagram as a qualitative measure that represents visually the "performance effectiveness" for meta-learning algorithms. The diagram plots the expected classification accuracy as a function of KL divergence between testing and training tasks. Intuitively, the closer a testing task is from the training tasks, the higher the performance. Hence, we can use our proposed correlation diagram to qualitatively compare different meta-learning methods.

A correlation diagram can be constructed by first calculating the average distance between each testing task, denoted as $M + i$ subscript with $i \in \mathbb{N}$, to all training tasks:

$$\overline{D}_{M+i} = \frac{1}{M} \sum_{d=1}^{M} D_{\mathrm{KL}}[q(\boldsymbol{\phi}; \boldsymbol{\lambda}_{M+i})\|q(\boldsymbol{\phi}; \boldsymbol{\lambda}_d)].$$

The obtained average distances are then grouped into J interval bins, each of size $\triangle_J = \max_i \overline{D}_{M+i}/J$. Let B_j with $j \in \{1, \ldots, J\}$ be the set of testing tasks that have their average KL distances falling within the interval $I_j = ((j-1)\triangle_J, j\triangle_J]$. The distance of bin B_j is defined as:

$$d(B_j) = \frac{1}{|B_j|} \sum_{i \in B_j} \overline{D}_{M+i}.$$

Next, a model trained on the training tasks is employed to evaluate the prediction accuracy $a_i^{(v)}$ on all the testing tasks to obtain the accuracy for bin B_j:

$$a(B_j) = \frac{1}{|B_j|} \sum_{i \in B_j} a_i^{(v)}.$$

Finally, plotting $d(B_j)$ against $a(B_j)$ gives the desired correlation diagram (e.g., Figure 2).

4   Experiments

We carry out two experiments – correlation diagram and task selection – to demonstrate the capability of the proposed approach. We evaluate the proposed approach on n-way classification tasks formed from two separated data sets: Omniglot (Lake et al., 2015) and mini-ImageNet (Vinyals et al., 2016). In this setting, a testing task is represented by a k-shot labelled data without the availability of unlabelled data following the transductive learning setting (Dhillon et al., 2020). We evaluate the performance on several meta-learning algorithms, such as MAML (Finn et al., 2017), Prototypical Networks (Snell et al., 2017), Amortised Meta-learner (ABML) (Ravi and Beatson, 2019), BMAML (Yoon et al., 2018) and VAMPIRE (C. Nguyen et al., 2020), to verify the distance-performance correlation using our proposed method.

For Omniglot, we follow the pre-processing steps as in few-shot image classification without any data augmentation, and use the standard train-test split in the original paper to prevent information leakage. For mini-ImageNet, we follow the common train-test split with 80 classes for training and 20 classes for testing (Ravi and Larochelle, 2017). Since the dimension of raw images in mini-ImageNet is large, we employ the 640-dimensional features extracted from a wide-residual-network (Rusu et al., 2019) to ease the calculation.

We follow Algorithm 1 in Supplementary Material C to obtain the posterior of the image-theme using tasks in training set. We use $L = 4$ task-themes and $K = 8$ image-themes for both data sets.

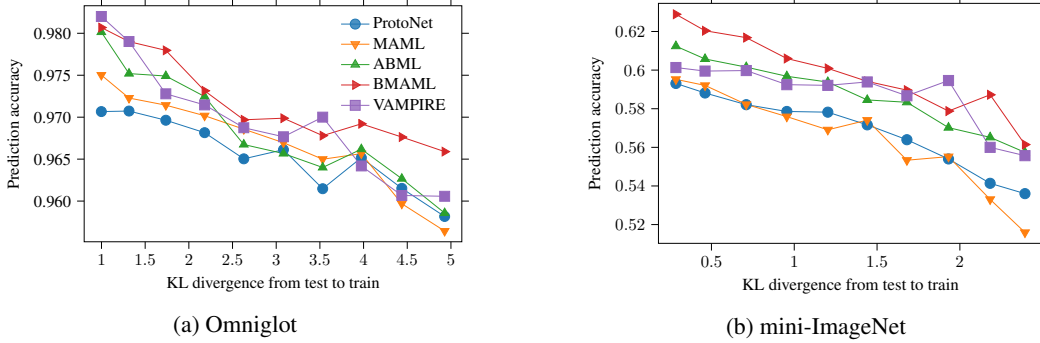(a) Omniglot   (b) mini-ImageNet

Figure 2: Correlation diagram plots the average accuracy predicted by meta-learning algorithms as a function of the average KL divergence of each task in the testing set to all tasks in the training set on the 5-way 1-shot setting.

The Dirichlet distribution for task-theme mixture, $\text{Dirchlet}_L(\boldsymbol{\phi}_d|\boldsymbol{\delta}$, is chosen to be symmetric with $\delta = 0.5$. The parameter inference, or training, is carried out with 16 images per class while varying the number of classes between 5 to 10 to fit into the memory of a Nvidia 1080 Ti GPU. The inference of the variational parameter $\boldsymbol{\lambda}$ is done on all available labelled data in each class (20 for Omniglot and 600 for mini-ImageNet). Note that this is used for the correlation diagram demonstration. For the task selection, this number matches the number of shots in the few-shot learning setting[2].

For the evaluation on meta-learning algorithms, we use a similar 4 convolutional module network to train on Omniglot (Vinyals et al., 2016; Finn et al., 2017; Snell et al., 2017), while using a fully connected network with 1 hidden layer consisting of 128 units to train on the extracted features of mini-ImageNet (C. Nguyen et al., 2020).

Note that the numbers of tasks formed from the two data sets are very large. For Omniglot, approximate 6.8×10^{12} and 10^{12} unique tasks can be generated from the training and testing sets, respectively. For mini-ImageNet, these numbers are slightly more manageable with about 2.4×10^6 unique tasks for training, and $15,504$ tasks for testing. To reduce the computation and facilitate the analysis, we randomly select 1 million Omniglot tasks for training, and $20,000$ tasks for testing. For mini-ImageNet, we select 2 million tasks for training and $15,504$ tasks for testing.

4.1 Correlation Diagram

To construct the correlation diagram, we train a continuous LDCC on n-way 16-shot setting (n varies from 5 to 10), and then infer the variational parameter $\boldsymbol{\lambda}$ of the task-theme mixture ϕ. The inferred λ is used to calculate the KL divergence distance between testing tasks to all training tasks. Note that the continuous LDCC is only trained on the training tasks. We then separately evaluate the performance of different meta-learning algorithms on the same 5-way 1-shot setting, and plot the correlation diagram in Figure 2. The results of the performance versus the task distance (or similarity) agree well with the common intuition: the testing tasks closer to the training tasks have higher prediction accuracy. Note that this observation is consistent across several meta-learning methods. It is also interesting to notice that some methods are more robust than others with respect to the dissimilarity between training and testing tasks.

4.2 Task Selection

We show that when there is a constraint on the number of training tasks, selecting tasks based on the proposed similarity outperforms the un-selective one that randomly selects training tasks. To demonstrate, we assume that one can pick a small number of mini-ImageNet tasks from the whole training set to train a meta-learning model, and evaluate on all tasks in the testing set. In the selective case, we use the LDCC model trained on all training tasks to infer the variational mixture parameters $\boldsymbol{\lambda}$ for all training and testing tasks. We then pick the training tasks that are closest to all

---

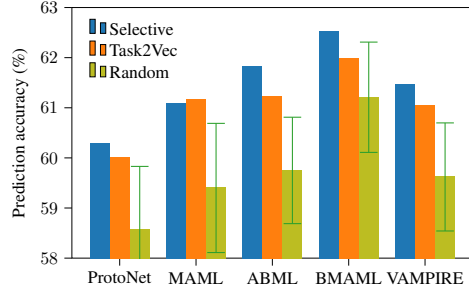[2]Implementation can be found at https://github.com/cnguyen10/similarity_classification_tasks

6

Figure 3: The prediction accuracy of several meta-learning methods on 5-way 5-shot mini-ImageNet testing tasks when training tasks are pro-actively selected outperforms the un-selective approaches, and slightly better than Task2Vec. The error bars on the un-selective cases represent the 95% confident intervals calculated on the 50 trials of random task selection.

(a) MAML 5-way

(b) ProtoNet 5-way

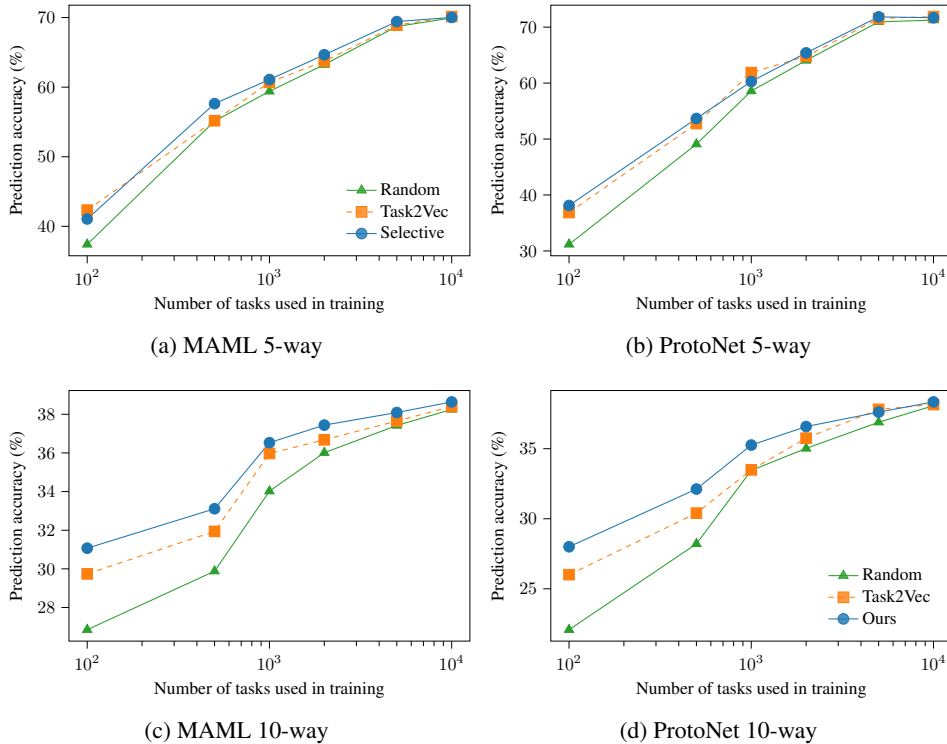(c) MAML 10-way

(d) ProtoNet 10-way

Figure 4: The proposed task-selective approach outperforms the randomly chosen training tasks, and shows slightly better results than Task2Vec when varying the number of classes within a classification task as well as the number of training tasks.

the testing tasks using the proposed KL divergence, and use them to train a meta-learning model. In the un-selective case, we randomly select the same number of training tasks without measuring any similarity. We also include Task2Vec as a baseline for the selective case to compare with our proposed approach. As the experiment is based on extracted features of mini-ImageNet, it is difficult to adapt to some common pre-trained networks, which is used as a "probe" network in Task2Vec. To work around, we use MAML to train a fully-connected network with three hidden layers on the training set under 5-way 5-shot setting, and use the feature extractor (excluding the last layer) of this network as the "probe" network for Task2Vec. This modelling approach results in a 3-D Task2Vec representation which is the same dimension as ϕ_d in the continuous LDCC, and hence, can be compared fairly. In addition, we directly calculate the diagonal of Fisher information matrix of the "probe" network without using the proposed approximation in Task2Vec to reduce the complexity of hyper-parameter tuning.

Figure 3 shows the accuracy results tested on $15,504$ mini-ImageNet testing tasks on the 5-way 5-shot setting for models trained on 1,000 training tasks. We also report the 95% confident interval for the case of random task selection. Statistically, meta-learning methods trained on tasks selected from our proposed solution outperform the un-selective cases, and slightly better than Task2Vec, especially for the probabilistic meta-learning methods such as BMAML, ABML and VAMPIRE.

To study the effects induced by the number of training tasks, and the number of ways within each task, we run an extensive experiment with a similar 5-shot setting, but varying the number of training tasks and ways, and plot the results in Figure 4. In general, the proposed approach out-performs the un-selective approach, and is slightly better than Task2Vec.

Despite promising results, there are some limitations of our proposed task selection. The proposed approach requires a sufficient number of labelled data in the testing tasks. More specifically, we need 5 labelled images per class, so that the trained LDCC model can correctly infer λ. Further reduction in the number of labelled data in the test set might result in a poor estimation of λ, hindering the task selection process. This is a well-known issue in LDA and its variations, which do not work well for short texts. Nevertheless, the assumption of 5-shot setting, which shows a promising result for task selection, is still reasonable in many few-shot learning applications.

5 Conclusion

We propose a generative approach based on the continuous LDCC adopted in topic modelling to model classification tasks. Under this modelling approach, a classification task can be expressed as a finite mixture model of Gaussian distributions, whose components are shared across all tasks. This new representation of classification tasks allows one to quantify the similarity between tasks through the asymmetric KL divergence. We also introduce a task selection strategy based on the proposed task similarity, and demonstrate its superiority in meta-learning comparing to the conventional approach where training tasks are randomly selected.

Acknowledgement

Broader impact

The proposed approach is helpful in transfer-learning tasks, especially when the amount of training data for the testing task is limited. By representing tasks in the topic space, the proposed approach allows to assess task similarity and provide insightful understanding when transfer-learning will be effective. This has the benefit of saving costs on data collection and annotation for the testing task. However, the trade-off is related to the computational cost involved in the training of the LDCC model for computing the task-to-task similarities.

References

Achille, Alessandro, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C Fowlkes, Stefano Soatto, and Pietro Perona (2019). "TASK2VEC: Task embedding for meta-learning". In: *IEEE International Conference on Computer Vision*, pp. 6430–6439.

Bakker, Bart and Tom Heskes (2003). "Task clustering and gating for Bayesian multitask learning". In: *Journal of Machine Learning Research* 4.May, pp. 83–99.

Bishop, Christopher M (2006). *Pattern Recognition and Machine Learning*. Springer.

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

Canini, Kevin, Lei Shi, and Thomas Griffiths (2009). "Online inference of topics with latent Dirichlet allocation". In: *Artificial Intelligence and Statistics*, pp. 65–72.

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

Chen, Yinbo, Xiaolong Wang, Zhuang Liu, Huijuan Xu, and Trevor Darrell (2020). "A new meta-baseline for few-shot learning". In: *arXiv preprint arXiv:2003.04390*.

Dhillon, Guneet S, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto (2020). "A baseline for few-shot image classification". In: *International Conference on Learning Representations*.

Finn, Chelsea, Pieter Abbeel, and Sergey Levine (2017). "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks". In: *International Conference on Machine Learning*, pp. 1126–1135.

Foulds, James, Levi Boyles, Christopher DuBois, Padhraic Smyth, and Max Welling (2013). "Stochastic collapsed variational Bayesian inference for latent Dirichlet allocation". In: *International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD)*, pp. 446–454.

Griffiths, Thomas L and Mark Steyvers (2004). "Finding scientific topics". In: *National Academy of Sciences* 101, pp. 5228–5235.

Hoffman, Matthew, Francis R Bach, and David M Blei (2010). "Online learning for latent Dirichlet allocation". In: *Advances in Neural Information Processing Systems*, pp. 856–864.

Jacob, Laurent, Jean-philippe Vert, and Francis R Bach (2009). "Clustered multi-task learning: A convex formulation". In: *Advances in Neural Information Processing Systems*, pp. 745–752.

Lake, Brenden M, Ruslan Salakhutdinov, and Joshua B Tenenbaum (2015). "Human-level concept learning through probabilistic program induction". In: *Science* 350.6266, pp. 1332–1338.

Li, Fei-Fei and Pietro Perona (2005). "A bayesian hierarchical model for learning natural scene categories". In: *International Conference on Computer Vision and Pattern Recognition*. Vol. 2. IEEE, pp. 524–531.

Minka, Thomas (2000). *Estimating a Dirichlet distribution*.

Nguyen, Cuong, Thanh-Toan Do, and Gustavo Carneiro (2020). "Uncertainty in model-agnostic meta-learning using variational inference". In: *IEEE Winter Conference on Applications of Computer Vision*, pp. 3090–3100.

Nguyen, Cuong V, Tal Hassner, Cedric Archambeau, and Matthias Seeger (2020). "LEEP: A New Measure to Evaluate Transferability of Learned Representations". In: *International Conference on Machine Learning*.

Pritchard, Jonathan K, Matthew Stephens, and Peter Donnelly (2000). "Inference of population structure using multilocus genotype data". In: *Genetics* 155.2, pp. 945–959.

Ravi, Sachin and Alex Beatson (2019). "Amortized Bayesian meta-learning". In: *International Conference on Learning Representations*.

Ravi, Sachin and Hugo Larochelle (2017). "Optimization as a model for few-shot learning". In: *International Conference on Learning Representations*.

Rusu, Andrei A, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell (2019). "Meta-learning with latent embedding optimization". In: *International Conference on Learning Representations*.

Shafiei, M Mahdi and Evangelos E Milios (2006). "Latent Dirichlet co-clustering". In: *International Conference on Data Mining*. IEEE, pp. 542–551.

Shui, Changjian, Mahdieh Abbasi, Louis-Emile Robitaille, Boyu Wang, and Christian Gagné (2019). "A Principled Approach for Learning Task Similarity in Multitask Learning". In: *International Joint Conference on Artificial Intelligence*, pp. 3446–3452.

Snell, Jake, Kevin Swersky, and Richard Zemel (2017). "Prototypical networks for few-shot learning". In: *Advances in Neural Information Processing Systems*, pp. 4077–4087.

Teh, Yee W, David Newman, and Max Welling (2007). "A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation". In: *Advances in Neural Information Processing Systems*, pp. 1353–1360.

Thrun, Sebastian and Joseph O'Sullivan (1996). "Discovering structure in multiple learning tasks: The TC algorithm". In: *International Conference on Machine Learning*. Vol. 96, pp. 489–497.

Tran, Anh T, Cuong V Nguyen, and Tal Hassner (2019). "Transferability and hardness of supervised classification tasks". In: *International Conference on Computer Vision*, pp. 1395–1405.

Tversky, Amos (1977). "Features of similarity." In: *Psychological review* 84.4, p. 327.

Vinyals, Oriol, Charles Blundell, Tim Lillicrap, Daan Wierstra, et al. (2016). "Matching networks for one shot learning". In: *Advances in Neural Information Processing Systems*, pp. 3630–3638.

Xue, Ya, Xuejun Liao, Lawrence Carin, and Balaji Krishnapuram (2007). "Multi-task learning for classification with Dirichlet process priors". In: *Journal of Machine Learning Research* 8.Jan, pp. 35–63.

Yoon, Jaesik, Taesup Kim, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn (2018). "Bayesian Model-Agnostic Meta-Learning". In: *Advances in Neural Information Processing Systems*, pp. 7343–7353.

Zamir, Amir R, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese (2018). "Taskonomy: Disentangling task transfer learning". In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3712–3722.

Zhang, Yu and Dit-Yan Yeung (2012). "A convex formulation for learning task relationships in multi-task learning". In: *Conference on Uncertainty in Artificial Intelligence*.