Similarity of Classification Tasks

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent advances in meta-learning has led to remarkable performances on several benchmarks. Such success depends on not only the meta-learning algorithms, but also the similarity between training and testing tasks. However, such task similarity observation is often ignored when evaluating meta-learning methods, potentially biasing the classification results of the testing tasks. For instance, recent studies have found a large variance of classification results among testing tasks, suggesting that not all testing tasks are equally related to training tasks. This motivates the need to analyse task similarity to optimise and better understand the performance of meta-learning. Despite some successes in investigating task similarity, most studies in the literature rely on task-specific models or the need of external models pre-trained on some large data sets. We, therefore, propose a generative approach based on a variant of Latent Dirichlet Allocation to model classification tasks without depending on any particular models nor external pre-trained networks. The proposed modelling approach allows to represent any classification task in the latent "topic" space, so that we can analyse task similarity, or select the most similar tasks to facilitate the meta-learning of a novel task. We demonstrate that the proposed method can provide an insightful evaluation for meta-learning algorithms on two few-shot classification benchmarks. We also show that the proposed task-selection strategy for meta-learning produces more accurate classification results on a new testing task than a method that randomly selects the training tasks.

1 Introduction

Within the last decade, machine learning has been deployed 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 development of a novel learning approach, known as transfer learning, that exploits past experience (in the form of models learned from other training tasks) to quickly learn new tasks. Recent developments in transfer-learning, particularly in meta-learning, have achieved state-of-the-art results in several 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 the transfer learning algorithm, but also on the similarity between training and testing classification 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 Latent Dirichlet Co-Clustering to model classification tasks. The resulting model allows to quantify the similarity between training and testing classification tasks in a latent topic space. The proposed similarity measure enables the possibility of selecting the most related tasks from the training set for the meta-learning of a novel testing task. We

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

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

2 Related work

The main inspiration of our work is the desire to understand the mechanism (W.-Y. Chen et al., 2019; Dhillon et al., 2020) and improve further the performance of meta-learning algorithms. 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. To better justify the performance of meta-learning methods, the authors proposed to use "task hardness" which is based on the cosine similarity between the embedding of labelled and unlabelled data. 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 learn 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, extensive experiments to construct a taxonomy for 26 computer-vision tasks, known as "taskonomy", was carried out to correlate tasks (Zamir et al., 2018). 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 contrast, our proposal follows a generative approach which does not depend on any task-specific classifier. 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), or Latent Dirichlet Allocation (LDA) (Blei et al., 2003) in topic modelling that analyses text data. LDA assumes that each document within a given corpus can be presented as a finite mixture model, where its components are the latent topics shared across all documents. Training LDA and its variants on a large text corpus is challenging, and hence, 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 efficiency for large corpora (Canini et al., 2009; Hoffman et al., 2010; Foulds et al., 2013). Our work is slightly different from conventional LDA models, where we perform online learning with "word" as continuous data, instead of the discrete data represented by a bag-of-word vector used in topic modelling.

3 Method

To **relate classification to topic modelling**, we consider **a task as a document**, **an image as a word**, and **a class as a paragraph**. 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. Our modelling approach is slightly different from the original LDCC at the distribution of word-topics. Since the data in classification is assumed to be continuous, the categorical word-topic distribution in the original LDCC model is replaced by a Gaussian "image-topic" distribution as shown in Figure 1. Under this approach, each classification task is modelled as a mixture of "Gaussian topics". We can, therefore, utilise this representation to measure 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

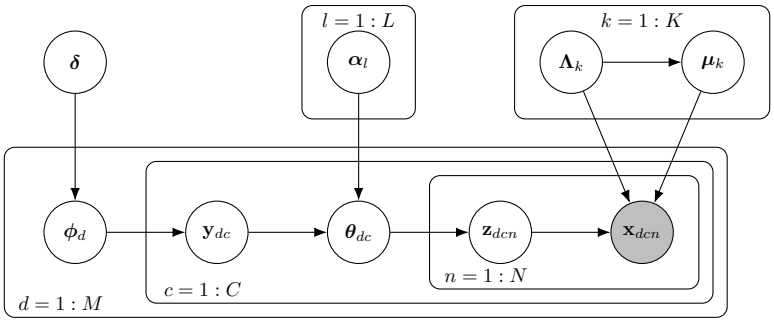


Figure 1: The graphical model LDCC used to model classification tasks.

C and N is trivial and can be implemented straightforwardly. The process to generate classification tasks can be presented as follows (graphical model shown in Figure 1):

- Initialise K Gaussian components $\{\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k\}_{k=1}^K$
- For the d-th task in the collection of M tasks:
  - Choose a task-topic mixing vector: $\boldsymbol{\phi}_d \sim \text{Dirichlet}\,(\boldsymbol{\phi}; \boldsymbol{\delta}, L)$
  - For the c-th class in the d-th task:
    - Choose a task-topic assignment: $\mathbf{y}_{dc} \sim \text{Categorical}(\mathbf{y}; \boldsymbol{\phi}_d)$
    - Choose an image-topic mixing vector: $\boldsymbol{\theta}_{dc} \sim \text{Dirichlet}\,(\boldsymbol{\theta}; \boldsymbol{\alpha}_l, K)$, where $y_{dcl} = 1$
    - For the n-th image in the c-th class of d-th task :
      - Choose an image-topic assignment: $\mathbf{z}_{dcn} \sim \text{Categorical}\,(\mathbf{z}; \boldsymbol{\theta}_{dc})$
      - Choose an image: $\mathbf{x}_{dcn} \sim \mathcal{N}\left(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right)$, where: $z_{dcnk} = 1$.

If the K image-topics (or, Gaussian components) $\{(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)\}_{k=1}^K$ are known, we can infer the mixing parameter $\boldsymbol{\phi}$ for any arbitrary task to represent that task in the latent topic simplex, which allows to measure distances between tasks. Hence, our objective is to learn these image-topics from the M given classification tasks.

Due to the complexity of the graphical model shown in Figure 1, the posterior of the K image-topics 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). 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 scalability. In this paper, we will use VI to infer the image-topics for classification tasks.

In VI, the true posterior is approximated by a variational distribution $q(\boldsymbol{\phi}, \mathbf{y}, \boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ (Attias, 2000). The parameters of the variational distribution q are estimated by maximising a lower bound of the marginal log-likelihood, also known as evidence lower bound (ELBO). Similar to the inference for LDA (Blei et al., 2003), we choose a fully factorised distribution q as our variational posterior:

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

where:

$$q(\boldsymbol{\phi}_d; \boldsymbol{\lambda}_d) = \text{Dirichlet}_L\,(\boldsymbol{\phi}_d; \boldsymbol{\lambda}_d) \qquad q(\mathbf{y}_{dc}; \boldsymbol{\eta}_{dc}) = \text{Categorical}\,(\mathbf{y}_{dc}; \boldsymbol{\eta}_{dc})$$
$$q(\boldsymbol{\theta}_{dc}; \boldsymbol{\gamma}_{dc}) = \text{Dirichlet}_K\,(\boldsymbol{\theta}_{dc}; \boldsymbol{\gamma}_{dc}) \qquad q(\mathbf{z}_{dcn}; \mathbf{r}_{dcn}) = \text{Categorical}\,(\mathbf{z}_{dcn}; \mathbf{r}_{dcn})$$
$$q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k; \mathbf{m}_k, \kappa_k, \mathbf{W}_k, \nu_k) = \mathcal{N}\left(\boldsymbol{\mu}_k; \mathbf{m}_k, (\kappa_k \boldsymbol{\Lambda}_k)^{-1}\right) \mathcal{W}\,(\boldsymbol{\Lambda}_k; \mathbf{W}_k, \nu_k)\,.$$

The ELBO can, therefore, be written as:

$$\mathsf{L} = \mathbb{E}_q\,[\ln p(\mathbf{z}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\alpha}) + \ln p(\mathbf{y}|\boldsymbol{\phi}) + \ln p(\boldsymbol{\phi}|\boldsymbol{\delta}) + \ln p(\mathbf{x}|\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) + \ln p(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\mathbf{m}_0, \kappa_0, \mathbf{W}_0, \nu_0) - \ln q(\mathbf{z}) - \ln q(\boldsymbol{\theta}) - \ln q(\mathbf{y}) - \ln q(\boldsymbol{\phi}) - \ln q(\boldsymbol{\mu}, \boldsymbol{\Lambda})]\,. \tag{2}$$

Comparing to the conventional LDA (Blei et al., 2003), the ELBO in Eq. (2) contains 4 extra terms highlighted in purple. 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. Please refer to the Supplementary Materials A, B and C for detailed derivation, optimisation, and the training and testing procedures.

Given the image-topics, 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 topic 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 $D_{\mathrm{KL}}\left[P\|Q\right]$ is defined as the information lost when using distribution P to approximate distribution Q, we, therefore, calculate $D_{\mathrm{KL}}\left[q(\boldsymbol{\phi}_d; \boldsymbol{\lambda}_{M+1})\|q(\boldsymbol{\phi}_d; \boldsymbol{\lambda}_d)\right]$, where $d \in \{1, \ldots, M\}$, to assess how a training task d-th differs from the learning of the novel task $(M+1)$-th.

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 tasks and training tasks. Intuitively, the further a testing task is from the training tasks, the lower the performance. Hence, we can use our proposed correlation diagram to qualitatively compare different meta-learning methods for different values of training and testing tasks similarity.

A correlation diagram can be constructed by first calculating the average distance between each testing task to all training tasks:

$$\overline{D}_{\mathrm{KL}_i} = \frac{1}{M}\sum_{d=1}^{M} D_{\mathrm{KL}}\left[q(\boldsymbol{\phi}; \boldsymbol{\lambda}_i^{(v)})\|q(\boldsymbol{\phi}; \boldsymbol{\lambda}_d^{(t)})\right],$$

where the superscripts (t) and (v) denote tasks in training and testing sets, respectively, and the subscript denotes task index. The obtained average distances are then grouped into J interval bins, each of size $\triangle_J = \max_i \overline{D}_{\mathrm{KL}_i}/J$. 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:

$$d(B_j) = \frac{1}{|B_j|}\sum_{i \in B_j} \overline{D}_{\mathrm{KL}_i}.$$

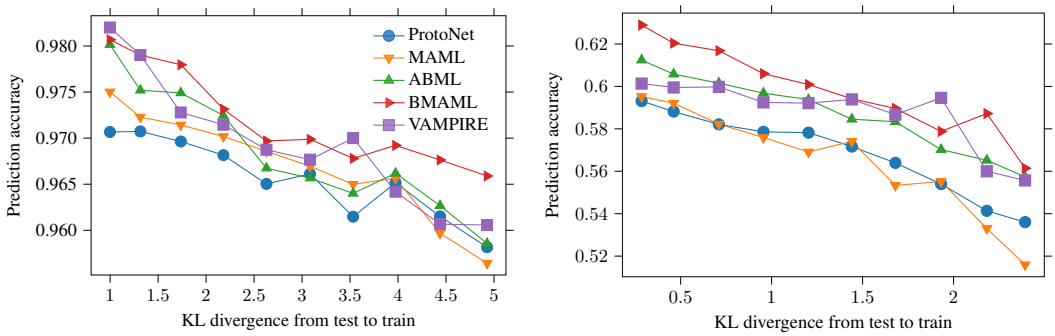
Next, a model trained on the training tasks is employed to evaluate the prediction accuracy $a_i^{(v)}$ on all the testing tasks to obtain the accuracy for bin B_j:

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

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),



(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.

Prototypical Networks (Snell et al., 2017), Amortised Meta-learner (ABML) (Ravi and Beatson, 2019), BMAML (Yoon et al., 2018) and VAMPIRE (Nguyen et al., 2020), to verify the distance-performance correlation using our proposed method. We use a similar 4 module CNN 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 (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, while keeping the original number of mini-ImageNet tasks. Please refer to Supplementary Material D for more details.

4.1 Correlation Diagram

We show the correlation diagrams for 5-way 1-shot testing tasks on the two data sets in Figure 2. The results 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-learing 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 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 overcome, we use MAML to train a 3-hidden layer fully-connected network 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 $\boldsymbol{\phi}_d$ in 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.

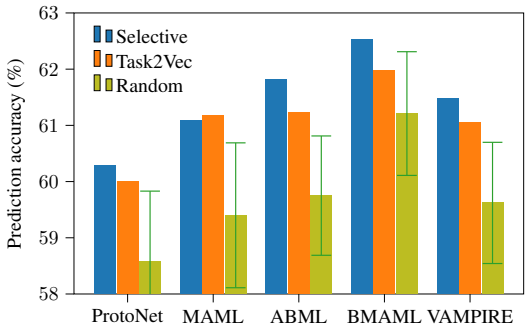
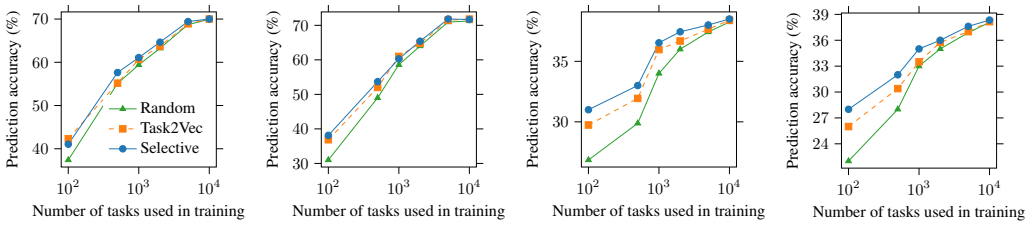


Figure 3: The prediction accuracy of several meta-learning methods on 5-way 5-shot mini-ImageNet testing tasks when training tasks are pro-actively selected outperforms the un-selective approaches, and slightly better than Task2Vec. The error bars on the un-selective cases represent the 95% confident 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: Additional experimental results on mini-ImageNet with Task2Vec baseline.

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.

To study the effect on the number of training tasks, and the number of ways within each task, we run experiments with the same 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 slightly better than Task2Vec. When the number of training tasks is large enough (about 1 million tasks), there is no difference between these three approaches.

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. In this case, we need 5 labelled images per class, so that the trained LDCC model can correctly infer $\boldsymbol{\lambda}$. Further reduction in the number of labelled data in the test set would result in a poor estimation of $\boldsymbol{\lambda}$, 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.

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.

Attias, Hagai (2000). "A variational Baysian framework for graphical models". In: *Advances in Neural Information Processing Systems*, pp. 209–215.

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: *19th International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD)*, pp. 446–454.

Griffiths, Thomas L and Mark Steyvers (2004). "Finding scientific topics". In: *Proceedings of the 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.

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.

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: *Sixth 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.

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*.