WASP Software Engineering Assignment

Jayadev Naram

1 Introduction

In the supervised classification problem of machine learning, given a labeled dataset that is generated from an unknown data distribution the goal is to learn a classifier that performs well on a new datapoint sampled from the same data distribution. However, in practice the test datapoint might be sampled from a different but a related data distribution. For example, the training data could be synthetically generated images and test data could be real-world images depicting the same scenario. Under such data shift, a trained classifier does not perform well on test data. This gave rise to the domain adaptation (DA) problem where in additional to labeled data from the first domain generally called source and unlabeled data from the second domain called target is used to train a classifier that performs well on a new datapoint from target domain. A key strategy is to align features from source domain and target domain while ensuring that these features are predictive on source domain.

Another interesting extension to this problem is when the classes present in target domain are an unknown subset of that in the source domain. This leads to partial domain adaptation (PDA) problem where it is known that a DA classifier that gives equal weightage to all the source datapoints during training is shown to perform worse than performing supervised learning approach that uses no target datapoints during the training. The key strategy is only align partially across the domain features and to down weight the influence of outlier source datapoints that can be deduced from partial alignment process. However, there is no theoretical understanding for any alignment methods and the weighting strategies used in the literature.

A recent topic of my research addressed these gaps in the literature. We derived an upper bound on test performance containing a weighted version of classification error, a partial alignment cost of our interest and an uncomputable term that dictates the difficulty of the PDA problem. This result motivated the use of an alignment approach and also inspired a PDA algorithm that is comparable to those in the literature.

While implementing our algorithm, we heavily relied on an existing code base in the literature. Given the vast amount of existing code, our algorithm implementation only lead to minor tweaks to crucial steps in the code.

However, this exposed the issues in the existing code base. Specifically, the tools that computes partial alignment is highly sensitive to hyperparameters. Through vast grid search and intuitive case studies, we found a range for each hyperparameter that leads to stable computation during training.

A future direction we are exploring is the multi-source PDA (MSPDA) problem where we given labeled data from multiple source domains which have domain shift and differing label spaces. The goal is to perform well on target domain using no label information under the constraint that no single source domain contains all the target classes (which is the case in PDA), but combining all of source label across domains contains target label space.

2 Lecture principles

Verification: As mentioned in the introduction, developing on top of an existing code base exposed a numerical stability issue in the code. We used several combination of hyperparameter settings to understand the behavior of our algorithm which can be seen as black-box testing. This resulted in identifying reasonable range of values for hyperparameters.

In addition to these dynamic testing, we also performed static testing via code review. This helped us resolve many bugs in the code in the early development process. However, the lack of writing unit tests led to a subtle bug that passively affected the performance. The SE principle of "test early, test often, test throughout, test automatically" is of significant importance in this scenario.

Validation: Despite addressing the research gap found in the area, my project only tackles the simplest of the learning problems: classification. A real-world application to DA problem is semantic segmentation where the goal is to predict a class label to each pixel of the image. It is a crucial component of autonomous driving systems. Extending our project such real-world application would be the desirable outcome of DA.

Another aspect of the project that is questionable is the applicability in LLM regime. The models and datasets used in the current project are conventional and small compared to the ones used in the context of LLMs. Our current algorithm cannot scale to very large datasets. Given the capabilities of LLMs, it is desirable to consider this extension as well.

3 Guest-Lecture Principles

Refine system goals into measurable, specific requirements: Some of the research goals in my project did not lead to trivial requirements. One such goal was to derive a theoretical result that motivates the strategy used in literature. Due to lack of prior PDA works addressing this goal, we had to

rely on existing literature on DA bounds and understand the type of results and proving techniques used. This led to refine the goal into a desired bound containing quantities of our interest. This helped in eventually deriving a bound matching the requirements.

In the same way we refined an empirical goal which was to propose a weighting strategy that is theortically motivated and removes the influence of outlier source datapoints in training. The corresponding requirement turned out to be an experiment where we investigated the weights distributions of outlier and inlier source datapoints showing there is very little weights to the former distribution. In this way, refining the goals into tangible requirements helped in my research project.

Tailoring - there is no one-size-fits-all solution: Adding to the first point mentioned under Validation in last section, let us consider the theoretical goal for DA semantic segmentation. Given the apparent difference in nature of classification and segmentation problems, the nature of the desired bound, i.e., the requirement changes.

Another interesting extension to DA problem which reinforces no one-size-fits-all solution argument is the MSPDA problem. In this scenario, the theoretical goal and corresponding requirement can be concretely, yet trivially framed based on the one for PDA problem. Yet, that is insufficient in understanding the problem. That would require to define a new requirement that is distinct from before and unique to this problem. Thus, requirements need to be tailored to the problem at hand.

4 Data Scientists versus Software Engineers

Opinion on Differences: Boringly, I agree to the key differences in the roles mentioned in the book. Data scientists tend to be research-oriented where the outcome is solving a research problem in a controlled environment. Whereas, software engineers product delivery-oriented where the real-world aspects such as uncertainty and budget constraints are factored in the development process. This indeed requires expertise in different backgrounds and focus on complementary metrics as mentioned in the book.

Future of these roles: To build a software product with a machine learning component, a common understanding is required between the two roles. Without such a common understanding, I believe it is impossible to successfully build a product together. As a result, I believe both sides will learn skills of the other side which will result in T-shaped interdisciplinary roles where, as mentioned in the book, each team member has a deep expertise in one topic and broad (but not deep) knowledge on other topics. I do believe that this will create new intermediate roles that is distinct from each side.

One such example is the dataset curator whose main job is to process the noisy user data into clean trainable data. This would involve the software engineer skills to combine different data pipelines to gather the data and maintain it, and also data scientist skills of cleaning the data and identify preliminary patterns in it that can help the model training process.

5 Paper analysis

Paper 1 [Barzamini et al., 2022]:

Core ideas and their SE importance: Due to the dependence of trained models on dataset variation, the paper suggests to augment dataset with missing variety by adopting the domain analysis task in SE. They proposed two approaches to acquire domain knowledge: i) textual-based extracted from existing human knowledge and ii) visual-based extracted from existing benchmark datasets. Extracted knowledge is used to design partial specifications of the domain which are semantically augmented to datasets that are used to train the model. When a model is deployed in the real-world, it is seldom true that the user data belongs to training data used. There will be domain shift in the data which is solved by the core idea of this paper.

Relation to my research: The paper also addressed domain shift problem which is at core my research project. My project tackles this problem by considering a particular form of data augmentation - labeled source data and unlabeled target data - which is used devise strategies that mitigate the domain shift during training. However, this paper tackles the domain shift problem by performing semantic data augmentation using SE task of domain analysis that is then used in training the model in a supervised manner. This way, the ideas in the paper can be seen as complemented that ones in my project.

Integration into a larger AI-intensive project: Consider an online semantic segmentation system that is integrated into a car's autonomous driving system which takes inputs from several different cameras. The domain analysis strategy can be used to introduce the variability in training the model. The feature alignment strategy used in my project can be integrated into the training data to leverage the unlabeled raw images captured by the cameras. Periodically updating the model by following the mentioned training strategy keeps it relevant.

Adaptation of my research: As mentioned earlier, the strategy mentioned in the paper is complementary to my strategy of domain adaptation. As such it can be seamlessly integrated into the PDA framework as follows.

Prior to the training, one can perform domain analysis on labeled source data which can be straightforward application of the ideas in the paper. The non-trivial task is to perform such analysis in the absence of labels in target dataset. This is because both the textual-based and visual-based approach of extracting domain knowledge rely on knowing the label of the data which violates the unsupervised constraints of the PDA problem. Applying the domain analysis on source data and using to in the PDA pipeline would help capture the real-world diversity in the data. This is a preliminary approach to extend my research project.

To further adapt the idea of the paper, one can address a potential in performing domain analysis only on source dataset in PDA problem. Suppose that as a result of analysis, we have augmented certain specification to the source dataset. However, as mentioned earlier, not all source classes are seen in the target dataset, specifically, only a subset of source classes are present in the target dataset. Therefore, a domain analysis that ignores this condition might incorrectly specify the knowledge of the outlier source classes to the dataset which clearly would not help to the PDA algorithm. In fact, it could worsen the performance when compared not performing any augmentation. As a result, one interesting venue to explore is to add the weighting strategy learned during the PDA training in the data augmentation process. This should reduce the negative impact of irrelevant domain knowledge and could lead to desired improved in combining the two strategies together. Thus, performing the domain adaptation iteratively during the PDA training could benefit from both the strategies.

Another potential area of improvement can be seen in combining multiple source domain as done in MSPDA problem. Leveraging the data augmentation and multi domain shifts in this scenario should further improve the performance compared to using one source domain at a time.

Given combining source domains is an easier task to implement although requiring higher cost of training. And data augmenting is difficult to implement due to inherent ambiguity in domain knowledge. However, once performed, the downstream training is straightforward. This bring forth an emergent research question: which strategy is better - combining source domain or augmenting source domains. My initial thoughts are that a combination of these two the best strategy.

To sum up, data augmentation exploits a previously unexplored resource in domain adaptation and it is definitely novel in leveraging the textual-based and visual-based domain knowledge in this literature.

Paper 2 [Vonderhaar et al., 2025]:

Core ideas and their SE importance: The goal is to generate synthetic datasets using generative model whose quality can be verified. A requirement engineering-based (RE) verification process is proposed for synthetically generated data. RE is also used in the data generating process via requirement specifications. As labeling a real-world data is costly, there is a increasing trend of using synthetic data to train the models. However, the performance and applicability of these models rely on the quality of the generated data, which emphasizes the need to control this process.

Relation to my research: In many semantic segmentation algorithms, the model is often trained on label-rich synthetic datasets. However, due to the sym-to-real shift in the real-world data, a domain adaptation step is necessary to leverage the learned knowledge from synthetic data. This is one example of how a machine learning problem can leverage both synthetic data addressed in this paper and domain adaptation addressed in my project to perform well in the real-world tasks.

Integration into a larger AI-intensive project: As mentioned above, one AI system that integrates the principles of the paper and my project is semantic segmentation model that works in real-world. An example product is a drone that uses this model to navigate in the environment without external controls.

Adaptation of my research: The requirements engineering is used to derive requirement specifications that are used in data generation process via prompts. The proposed verification strategy either approves the image or raises a fault based on these specifications. However, it is wasteful to discard these images that do not meet the specifications. One strategy that can be adopted from my research is to use a alignment process to measure the discrepancy in the generated images. A generated image that is close to the real image alignment well to the latter and an image that is farther from real image leads to a poor alignment. Our alignment strategy when used on a collection of generated data and real-world data gives rise to weights that are proportional to the alignment process. Attaching these weights to the generated images and using them in the downstream task to down weight the low quality images will lead to using all the generated data. As opposed to discarding the low quality images, this strategy is potentially useful when the computational cost in generation process is factored in. Also a hard thresholding discarding process might remove images that are almost good but were flagged due to minor issues. Despite the apparent uselessness of such images, they might still lead to meaning knowledge when used with the proposed weighting strategy.

6 Research Ethics & Synthesis Reflection

Search and screening process: I searched through the list of accepted papers based on relevance to the general area of data diversity and learning a model in its presence. This indeed helped in screening among several papers potential papers, the ones that are relevant and let me relate my research ideas to the ideas in the paper.

Pitfalls and mitigations: There were a few papers that had domain-specificity and knowledge distillation in the title. But reading the abstract gave a clear idea that they addressed an entirely different problem.

Ethical considerations: I used the direct approach of writing everything from own thoughts and not using LLMs at all. Perhaps, I could've refined my thoughts using an LLM, but I preferred not to.

References

[Barzamini et al., 2022] Barzamini, H., Rahimi, M., Shahzad, M., and Alhoori, H. (2022). Improving generalizability of ml-enabled software through domain specification. In *2022 IEEE/ACM 1st International Conference on AI Engineering–Software Engineering for AI (CAIN)*, pages 181–192.

[Vonderhaar et al., 2025] Vonderhaar, L., Elvira, T., and Ochoa, O. (2025). Generating and verifying synthetic datasets with requirements engineering. In *2025 IEEE/ACM 4th International Conference on AI Engineering–Software Engineering for AI (CAIN)*, pages 212–221.