A Unified Perspective on Multi-Domain and Multi-Task Learning

A UNIFIED PERSPECTIVE ON MULTI-DOMAIN AND MULTI-TASK LEARNING

Yongxin Yang & Timothy M. Hospedales
Electronic Engineering and Computer Science
Queen Mary, University of London
{yongxin.yang, t.hospedales}@qmul.ac.uk

ABSTRACT

In this paper, we provide a new neural-network based perspective on multi-task learning (MTL) and multi-domain learning (MDL). By introducing the concept of a semantic descriptor, this framework unifies MDL and MTL as well as encompassing various classic and recent MTL/MDL algorithms by interpreting them as different ways of constructing semantic descriptors. Our interpretation provides an alternative pipeline for zero-shot learning (ZSL), where a model for a novel class can be constructed without training data. Moreover, it leads to a new and practically relevant problem setting of zero-shot domain adaptation (ZSDA), which is the analogous to ZSL but for novel domains: A model for an unseen domain can be generated by its semantic descriptor. Experiments across this range of problems demonstrate that our framework outperforms a variety of alternatives.

1 INTRODUCTION

Multi-task and multi-domain learning are established strategies to improve learning by sharing knowledge across different but related tasks or domains. Multi-domain learning refers to sharing information about the same problem across different contextual domains, while multi-task learning addresses sharing information about different problems in the same domain. Because the domain/task distinction is sometimes subtle, and some methods proposed for MTL can also address MDL and vice-versa, the two settings are sometimes loosely used interchangeably. However, it is useful to distinguish them clearly: Domain relates to some covariate, such as the bias implicitly captured in a particular dataset (Torralba & Efros, 2011), or the specific data capture device. For example the Office Dataset (Saenko et al., 2010) contains three domains related to image source: Amazon, webcam, and DSLR. A multi-domain learning problem can then be posed by training a particular object recogniser across these three domains (same task, different domains). In contrast, a multi-task problem would be to share information across the recognisers for individual object categories (same domain, different tasks). The issue of simultaneously addressing multiple tasks and multiple domains seems to be un-addressed in the literature to our knowledge.

In this paper, we propose a neural network framework that addresses both multi-domain and multi-task learning, and can perform simultaneous multi-domain multi-task learning. A key concept in our framework is the idea of a multivariate "*semantic descriptor*" for tasks and domains. Such a descriptor is often available as metadata, and can be exploited to improve information sharing for MTL and MDL. We show that various classic and recent MTL/MDL methods are special cases of our framework that make particular assumptions about this descriptor: Existing algorithms typically implicitly assume categorical domains/tasks, which is less effective for information sharing when more detailed task/domain metadata is available. For example, the classic "school dataset" poses a task of predicting students' grades, and is typically interpreted as containing a domain corresponding to each school. However, since each school has three year groups, representing domains by a semantic descriptor tuple (school-id, year-group) is better for information sharing. Our framework exploits such multi-variate semantic descriptors effectively, while existing MTL/MDL algorithms would struggle to do so, as they implicitly consider tasks/domains to be atomic.

Going beyond information sharing for known tasks, an exciting related paradigm for task-transfer is "zero-shot" learning (ZSL) (Larochelle et al., 2008; Lampert et al., 2009; Fu et al., 2014). This

setting addresses automatically constructing a test-time classifier for categories which are unseen at training time. Our neural-network framework provides an alternative pipeline for ZSL. More interestingly, it leads to the novel problem setting of zero-shot domain adaptation (ZSDA): Synthesising a model appropriate for a new unseen domain given only its semantic descriptor. For example, suppose we have an audio recogniser trained for a variety of acoustic playback environments, and for a variety of microphone types: Can we synthesise a recogniser for an arbitrary environment-microphone combination? To our knowledge, this is the first time that zero-shot domain adaptation has been addressed specifically.

2 Related Work

2.1 Multi-Task Learning

Multi-Task Learning (MTL) aims to jointly learn a set of tasks by discovering and exploiting task similarity. Various assumptions have been made to achieve this. An early study (Evgeniou & Pontil, 2004) assumed a linear model for ith task can be written as $w_i := w_0 + v_i$ where w_0 can be considered as the *shared knowledge* which benefits all tasks and v_i is the *task-specific knowledge*.

Another common assumption of MTL is that the predictors (task parameters) lie in a low dimensional subspace (Argyriou et al., 2008). Imposing the (2,1)-norm on the predictor matrix W, where each column is a task, results in a low-rank W, which implicitly encourages parameter sharing. However, this assumes that all tasks are related, which is likely violated in practice. Forcing predictors to be shared across unrelated tasks can significantly degrade the performance – a phenomenon called negative transfer (Rosenstein et al., 2005). A task grouping framework is thus proposed by Kang et al. (2011) that partitions all tasks into disjoint groups where each group shares a low dimensional structure. This partially alleviates the unrelated task problem, but misses any fundamental information shared by all tasks, as there is no overlap between the subspaces of each group.

As a middle ground, the GO-MTL algorithm (Kumar & Daumé III, 2012) allows information to be shared between different groups, by representing the model of each task as a linear combination of latent predictors. Thus the concept of grouping is no longer explicit, but determined by the coefficients of the linear combination. Intuitively, model construction can be thought of as: $W = LS$ where L is the matrix of which each column is a latent predictor (*shared knowledge*), and $S = [s_1, s_2, \ldots, s_M]$ where s_i is a coefficient vector that cues how to construct the model for the ith task (*task-specific knowledge*). It is worth noting that this kind of predictor matrix factorisation approach – $W = LS$ – can explain several models: Kumar & Daumé III (2012) is L1/L2 regularised decomposition, Passos et al. (2012) is linear Gaussian model with IBP prior and an earlier study (Xue et al., 2007) assumes s_i are unit vectors generated by a Dirichlet Process (DP).

Most MTL methods in literature assume that each task is an atomic entity indexed by a single categorical variable. Some recent studies (Romera-paredes et al., 2013; Kishan Wimalawarne & Tomioka, 2014) noticed a drawback – this strategy can not represent a task with more structured metadata, e.g., (school-id, year-group) for school dataset. Thus they replace predictor matrix W with a tensor so that the linear models across more than one categorical variable can be placed in the tensor. Then they follow the line of Argyriou et al. (2008) to impose a variety of regularisations on the mentioned tensor, such as sum of the ranks of the matriciations of the tensors (Romera-paredes et al., 2013) and scaled latent trace norm (Kishan Wimalawarne & Tomioka, 2014). However, this again suffers from the strong assumption that all tasks are related.

2.2 Multi-Domain Learning

Domain Adaptation There has been extensive work on domain adaptation (DA) (Beijbom, 2012). A variety of studies have proposed both supervised (Saenko et al., 2010; Duan et al., 2012) and unsupervised (Gong et al., 2012; Sun & Saenko, 2014) methods. As we have mentioned, the typical assumption is that domains are indexed by a single categorical variable: For example a data source such as Amazon/DSLR/Webcam (Saenko et al., 2010), a benchmark dataset such as PASCAL/ImageNet/Caltech (Gong et al., 2012), or a modality such as image/video (Duan et al., 2012).

Despite the majority of research with the categorical assumption on domains, it has recently been generalised by studies considering domains with a (single) continuous parameter such as time (Hoff-

man et al., 2014) or viewing angle (Qiu et al., 2012). In this paper, we take an alternative approach to generalising the conventional categorical formulation of domains, and instead investigate information sharing with domains described by a *vector* of discrete parameters.

Multi-Domain Learning Multi-Domain Learning (MDL) (Dredze et al., 2010; Joshi et al., 2012) shares properties of both domain adaptation and multi-task learning. In conventional domain adaptation, there is an explicit pair of source and target domain, and the knowledge transfer is one way Source→Target. In contrast, MDL encourages knowledge sharing in both directions. Although some existing MTL algorithms reviewed in previous section tackle MDL as well, we distinguish them by the key difference during testing time: MDL makes prediction for same problem (binary classification like "is laptop") across multiple domains (e.g., datasets or camera type), but MTL handles different problems (such as "is laptop" versus "is mouse").

2.3 Zero-Shot Learning

Zero-Shot Learning (ZSL) aims to eliminate the need for training data for a particular task. It has been widely studied in different areas, such as character (Larochelle et al., 2008) and object recognition (Lampert et al., 2009; Socher et al., 2013; Fu et al., 2014). Typically for ZSL, the label space of training and test data are disjoint, so no data has been seen for test-time categories. Instead, test-time classifiers are constructed given some mid-level information. Although diverse in other ways, most existing ZSL methods follow the pipeline in Palatucci et al. (2009): $X \rightarrow Z \rightarrow Y$ where Z is some "semantic descriptor", which refers to attributes (Lampert et al., 2009) or semantic word vectors (Socher et al., 2013). Our work can be considered as an alternative pipeline, which is more similar to Larochelle et al. (2008) and Frome et al. (2013) in the light of the following illustration: $Z \overset{\frown}{\longrightarrow} X \longrightarrow Y$.

Going beyond conventional ZSL, we generalise the notion of zero-shot learning of tasks to zero-shot learning of domains. In this context, zero-shot means no training data has been seen for the target domain prior to testing. The challenge is to construct a good model for a novel test domain based solely on its semantic descriptor. The closest work to our zero-shot domain adaptation setting is Ding et al. (2014), which addresses the issue of a missing modality with the help of the partially overlapped modalities that have been previously seen. However they use a single fixed modality pair, rather than synergistically exploiting an arbitrary number of auxiliary domains in a multi-domain way as in our framework. Note that despite the title, Blitzer et al. (2009) actually considers unsupervised domain adaptation without target domain labels, but *with* target data.

3 Model

3.1 General Framework

Assume that we have M domains (tasks), and the ith domain has N_i instances. We denote the feature vector of the jth instance in the ith domain (task) and its associated semantic descriptor by the pair $\{\{x_j^{(i)}, z^{(i)}\}_{j=1,2,\cdots,N_i}\}_{i=1,2,\cdots,M}$ and the corresponding label as $\{\{y_j^{(i)}\}_{j=1,2,\cdots,N_i}\}_{i=1,2,\cdots,M}$. Note that, in multi-domain or multi-task learning, all the instances are effectively associated with a semantic descriptor indicating their domain (task). Without loss of generality, we propose an objective function that minimises the empirical risk for all domains (tasks),

$$\arg\min_{P,Q} \frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{N_i}\sum_{j=1}^{N_i}\mathcal{L}\left(\hat{y}_j^{(i)}, y_j^{(i)}\right)\right), \quad \text{where} \quad \hat{y}_j^{(i)} = f_P(x_j^{(i)}) \cdot g_Q(z^{(i)}) \tag{1}$$

This model can be understood as a two-sided neural network illustrated by Figure 1. One can see it contains two learning processes: the left-hand side is representation learning $f_P(\cdot)$, starting with the original feature vector x; and the right-hand side is model construction $g_Q(\cdot)$, starting with an associated semantic descriptor z. P and Q are the weights to train for each side. To train P and Q, standard back propagation can be performed by the loss $\mathcal{L}(\cdot)$ calculated between ground truth y and the prediction $\hat{y}$.

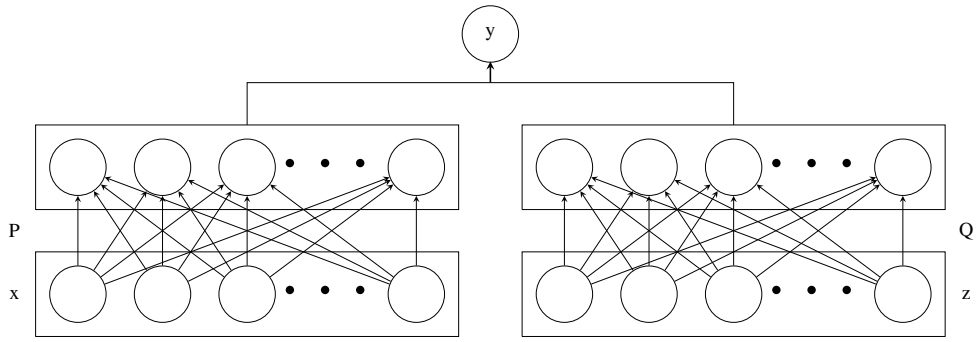

Figure 1: Two-sided Neural Network for Multi-Task/Multi-Domain Learning

With this neural network interpretation, the two sides can be arbitrarily complex but we find that one inner product layer for each is sufficient to unify some existing MDL/MTL algorithms and demonstrate the efficacy of the approach. In this case, P is a D-by-K matrix and Q is a B-by-K matrix, where K is the number of units in the middle layer; D and B is the length of feature vector x and semantic descriptor z respectively. The prediction is then based on $(x_j^{(i)}P)(z^{(i)}Q)'$.

3.2 Unification of Existing Algorithms

We next demonstrate how a variety of existing algorithms[1] are special cases of our general framework. For clarity we show this in an MDL/MTL setting with $M = 3$ domains/tasks. Observe that RMTL (Evgeniou & Pontil, 2004), FEDA[2] (Daumé III, 2007), MTFL (Argyriou et al., 2008) and GO-MTL (Kumar & Daumé III, 2012) each assume specific settings of Z, P and Q' as in Table 1.

Table 1: A Unifying Review of Some Existing MTL/MDL Algorithms

	Z	P	Norm on P	Q'	Norm on Q'
RMTL	$\begin{bmatrix}1&0&0&1\\0&1&0&1\\0&0&1&1\end{bmatrix}$	Identity	None	$\begin{bmatrix}\vert&\vert&\vert&\vert\\v_1&v_2&v_3&w_0\\\vert&\vert&\vert&\vert\end{bmatrix}$	None
FEDA[2]	$\begin{bmatrix}1&0&0&1\\0&1&0&1\\0&0&1&1\end{bmatrix}$	$a \otimes b$	None	$\begin{bmatrix}\underline{0}&\underline{0}&\underline{0}&w_0\\w_1&\underline{0}&\underline{0}&\underline{0}\\\underline{0}&w_2&\underline{0}&\underline{0}\\\underline{0}&\underline{0}&w_3&\underline{0}\end{bmatrix}$	None
MTFL	$\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}$	Identity	None	W	(2, 1)-Norm
GO-MTL	$\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}$	L	Frobenius	S	Entry-wise ℓ_1

The notion used is kept same with the original paper, e.g., P here is analogous to L in Kumar & Daumé III (2012). Each row of the matrices in the second (Z) column is the corresponding domain's semantic descriptor in different methods. These methods are implicitly assuming a single categorical domain/task index: with 1-of-N encoding as semantic descriptor (sometimes with a constant term).

We argue that more structured domain/task-metadata is often available, and with our framework it can be directly exploited to improve information sharing compared to simple categorical indices. For example, suppose two categorical variables (A,B) describe a domain, and each of them has two states (1,2), then four distinct domains can be encoded by Z in a distributed fashion (Table 2 left) in contrast to the 1-of-N form used by traditional multi-task learning methods (Table 2 right). The ability to exploit more structured domain/task descriptors Z where available, improves information sharing compared to existing MTL/MDL methods. In our experiments, we will demonstrate examples of problems with multivariate domain/task metadata, and its efficacy to improve learning.

[1] RMTL: **R**egularized **M**ulti–**T**ask **L**earning, FEDA: **F**rustratingly **E**asy **D**omain **A**daptation, MTFL: **M**ulti–**T**ask **F**eature **L**earning and GO-MTL: **G**rouping and **O**verlap for **M**ulti–**T**ask **L**earning

[2] a is an (M+1)-dimensional row vector with all ones, e.g., $a = [1, 1, 1, 1]$ when M=3, b is a D-by-D identity matrix, and $\otimes$ denotes Kronecker product. w_0, w_1, w_2, w_3 and $\underline{0}$ in Q' are D-dimensional column vectors.

Table 2: Illustration for Distributed Coding and 1-of-N Coding

	A-1	A-2	B-1	B-2
Domain-1	1	0	1	0
Domain-2	1	0	0	1
Domain-3	0	1	1	0
Domain-4	0	1	0	1

	A-1-B-1	A-1-B-2	A-2-B-1	A-2-B-2
Domain-1	1	0	0	0
Domain-2	0	1	0	0
Domain-3	0	0	1	0
Domain-4	0	0	0	1

3.3 LEARNING SETTINGS

Multi-domain multi-task (MDMT) Existing frameworks have focused on either MDL or MTL settings but not considered both together. Our interpretation provides a simple means to exploit them both simultaneously for better information sharing when multiple tasks in multiple domains are available. If $z^{(d)}$ and $z^{(t)}$ are the domain and task descriptors respectively, then MDMT learning can be performed by simply concatenating the descriptors $[z^{(d)}, z^{(t)}]$ corresponding to the domain and task of each individual instance during learning.

Zero-shot learning (ZSL) As mentioned, the dominant zero-shot (task) learning pipeline is $X \rightarrow Z \rightarrow Y$. At train time, the $X \rightarrow Z$ mapping is learned by classifier/regressor, where Z is a task descriptor, such as a binary attribute vector (Lampert et al., 2009; Fu et al., 2014), or a continuous word-vector describing the task name (Socher et al., 2013; Fu et al., 2014). At testing time, the "prototype" semantic vector for a novel class z is presented, and zero-shot recognition is performed by matching the $X \rightarrow Z$ estimate and prototype z, e.g., by nearest neighbour (Fu et al., 2014).

In our framework, ZSL is achieved by presenting each novel semantic vector z_j^* (each testing category is indexed by j) in turn along with novel category instances x^*. Zero-shot recognition then is given by: $j^* = \arg\max_j f_P(x^*) \cdot f_Q(z_j^*)$.

Zero-shot domain adaptation (ZSDA) The zero-shot domain adaptation task can also be addressed by our framework. With a distributed rather than 1-of-N encoded domain descriptor, only a subset of domains is necessary to effectively learn Q. Thus a model suitable for data from a *novel held-out domain* can be constructed by applying its semantic descriptor z^* along with data x^*.

4 EXPERIMENTS

We demonstrate our framework on five experimental settings: MDL, ZSDA, MTL, ZSL and MDMT.

Implementation: We implement the model with the help of Caffe framework (Jia et al., 2014). Though we don't place regularisation terms on P or Q, a non-linear function $\sigma(x) = max(0, x)$ (i.e., ReLU activation function) is placed to encourage sparse models $g_Q(z^{(i)}) = \sigma(z^{(i)}Q)$. The choice of loss function for regression and classification is Euclidean loss and Hinge loss respectively. Preliminary experiments show $K = \frac{D}{\log(D)}$ leads to satisfactory solutions for all datasets.

MTL/MDL Baselines: We compare the proposed method with a single task learning baseline – linear or logistic regression with ℓ_2 regularisation (LR), and four multi-task learning methods: (i) RMTL (Evgeniou & Pontil, 2004), (ii) FEDA (Daumé III, 2007), (iii) MTFL (Argyriou et al., 2008) and (iv) GO-MTL (Kumar & Daumé III, 2012). Note that these methods are re-implemented within the proposed framework. We have verified our implementations with the original ones and found that the performance difference is not significant. Baseline methods use traditional 1-of-N encoding, while we use a distributed descriptor encoding based on metadata for each problem.

Zero-Shot Domain Adaptation: We follow the MDL setting to learn P and Q except that one domain is held out each time. We construct test-time models for held out domains using their semantic descriptor. We evaluate against two baselines: (i) Blind-transfer (LR): learning a single linear/logistic regression model on aggregated data from all seen domains. To ensure fair comparison, distributed semantic descriptors are concatenated with the feature vectors for baselines, i.e., they are included as a plain feature. (ii) Tensor-completion (TC): we use a tensor $W \in \mathcal{R}^{D,p_1,p_2,\cdots,p_N}$ to store all the linear models trained by SVM where N is the number of categorical variables and p_i is the number of states in the ith categorical variable ($p_1 + p_2 + \cdots + p_N = B$ in our context and $p_1 * p_2 * \cdots * p_N = M$ if there is always a domain for each of possible combinations). ZSDA can be formalised by setting the model parameters for the held-out domain to missing values, and recovering them by a low-rank tensor completion algorithm (Kressner et al., 2014). This low-rank strategy corresponds to our implementation of Romera-paredes et al. (2013).

4.1 SCHOOL DATASET - MDL AND ZSDA

Data This classic dataset[3] collects exam grades of 15,362 students from 139 schools. Given the 23 features[4], a regression problem is to predict each student's exam grade. There are 139 schools and three year groups. School IDs and year groups naturally form multivariate domains. Note that 64 of 139 schools have the data of students for all three year groups, and we also choose the school of which each year group has more than 50 students so that each domain has sufficient training data. Finally there are $23 \times 3 = 69$ distinct domains given these two categorical variables.

Settings and Results For MDL we learn all domains together, and for ZSDA we use a leave-one-domain-out strategy, constructing the test-time model based on the held-out domain's descriptor with P and Q learned from the training domains. In each case the training/test split is 50%/50%. Note that the test sets for MDL and ZSDA are the same. The results in Table 3 are averages over the test set for all domains (MDL), and averages over the held-out domain performance when holding out each of the 69 domains in turn (ZSDA). Our method outperforms the alternatives in each case.

Table 3: School Dataset (RMSE)

	LR	RMTL	FEDA	MTFL	GO-MTL	TC	Ours
MDL	9.51	9.46	10.75	10.22	10.00	-	**9.37**
ZSDA	10.35	-	-	-	-	12.41	**10.19**

4.2 AUDIO RECOGNITION - MDL AND ZSDA

Audio analysis tasks are affected by a variety of covariates, notably the playback device / environment (e.g., studio recording versus live concert hall), and the listening device (e.g., smartphone versus professional microphone). Directly applying a model trained in one condition/domain to another will result in poor performance. Moreover, as the covariates/domains are combinatorial: (i) models cannot be trained for all situations, and (ii) even applying conventional domain adaptation is not scalable. Zero-shot domain adaptation has potential to address this, because a model could be calibrated on the fly for a given environment.

Data We investigate recognition in a complex set of noise domains: covering both acoustic environment and microphone type. We consider a music-speech discrimination task introduced by Tzanetakis & Cook (2002), which includes 64 music and speech tracks. Two categorical variables are *smartphone microphone* and *live concert hall* environment, and each of them has two states: on or off. Then the four domains are generated as: (i) Original (ii) Live Recording (LR) (iii) Smartphone Recording (SR) and (iv) smartphone in a live hall (LRSR). The noises are synthesised by Audio Degradation Toolbox (Mauch & Ewert, 2013).

Settings and Results We use MFCC to extract audio features and K-means to build a K = 64 bag-of-words representation. We split the data 50%/50% for training and test and keep test sets same for MDL and ZSDA. The results in Table 4 break down the results by each domain and overall (MDL), and each domain when held-out (ZSDA). In each case our method is best or joint-best due to better exploiting the semantic descriptor (recall that it does not have any additional information; for fairness the descriptor is also given to the other methods as a regular feature). The only exception is the least practical case of ZSDA recognition in a noise free environment given prior training only in noisy environments. The ZSDA result here generally demonstrates that models can be synthesised to deal effectively with new multivariate domains / covariate combinations without needing to exhaustively see data and explicitly train models for all, as would be conventionally required.

4.3 ANIMAL WITH ATTRIBUTES - MTL AND ZSL

Animal with Attributes (Lampert et al., 2009) includes images from 50 animal categories, each with an 85-dimensional binary attribute vector. The attributes, such as "black", "furry", "stripes", describe an animal semantically, and provide a unique mapping from a combination of attributes to an animal. The original setting of ZSL with AwA is to split the 50 animals into 40 for training

[3] Available at http://multilevel.ioe.ac.uk/intro/datasets.html

[4] The original dataset has 26 features, but 3 that indicate student year group are used in semantic descriptors.

Table 4: Audio Recognition: Music versus Speech (Error Rate)

		Origin	LR	SR	LRSR	Avg
MDL	LR	**3.13**	18.75	6.25	17.19	11.33
	RMTL	6.25	18.75	6.25	17.19	12.11
	FEDA	7.81	18.75	9.38	18.75	13.67
	MTFL	6.25	21.88	9.38	**14.06**	12.89
	GO-MTL	**3.13**	**17.19**	6.25	18.75	11.33
	Ours	**3.13**	**17.19**	**4.69**	**14.06**	**9.77**
ZSDA	LR	**32.81**	28.13	14.06	23.44	24.61
	TC	46.88	21.88	26.56	59.38	38.67
	Ours	35.94	**9.38**	**12.50**	**18.75**	**19.14**

and hold out 10 for testing. We evaluate this condition to investigate: (i) if multi-task learning of attributes and classes improves over the STL approaches typically taken when analysing AwA, (ii) if it helps to use the attributes as an MTL semantic task descriptor against the traditional setting of MTL where semantic descriptor is a 1-of-N unit vector indexing tasks. For MTL training on AwA, we decompose the multi-class problem with C categories to C one-vs-rest binary classification tasks. Note that in this case the semantic descriptor reveals the label, so it is not given during testing. We run all one-vs-rest classifiers on each instance and rank the scores to produce the label.

Multi-Task Learning We use the recently released DeCAF feature (Donahue et al., 2015) for AwA. For MTL, we pick five animals from the training set with moderately overlapped attributes, and use the first half of the images for training then test on the rest. The results in Table 5 show limited improvement by existing MTL approaches over the standard STL. However, our attribute-descriptor approach to encoding tasks for MTL improves the accuracy by about 2% over STL.

Table 5: AwA: MTL Multi-Class Accuracy

	antelope	killer whale	otter	walrus	blue whale	Avg
LR	92.31	87.08	89.26	75.60	82.44	85.34
RMTL	86.08	71.22	80.99	61.90	**96.18**	79.28
FEDA	92.31	83.39	88.15	79.17	89.31	86.47
MTFL	92.67	85.61	90.36	79.76	87.02	87.09
GO-MTL	91.21	84.87	89.81	**80.36**	84.73	86.20
Ours	**93.41**	**91.51**	**94.21**	79.76	79.39	**87.66**

Zero-Shot Learning For ZSL, we adopt the training/testing split in Lampert et al. (2009). The blind-transfer baseline is not meaningful because there are different binary classification problems, and aggregating does not lead to anything. Also, tensor-completion is not practical because of its exponential space ($D * 2^{85}$) against $D * 40$ observations. Our method achieves 43.79% multi-class accuracy, compared to 41.03% from direct-attribute prediction (DAP) approach (Lampert et al., 2009) using DeCAF features. A recent result using DeCAF feature is 44.20% in Deng et al. (2014), but this uses additional higher order attribute correlation information. Given that we did not design a solution for AwA specifically, or exploit this higher order correlation cue, the result is encouraging.

4.4 Restaurant & Consumer Dataset - MDMT

The restaurant & Consumer Dataset, introduced by Vargas-Govea et al. (2011) contains 1161 customer-to-restaurant scoring records, where each record has 43 features and three scores: food, service and overall. We build a multi-domain multi-task problem as follows: (i) a domain refers to a restaurant, (ii) a task is a regression problem to predict one of the three scores given an instance and (iii) an instance is a 43-dimensional feature vector based on customer's and restaurant's profile. The 1161 records cover 130 restaurants but most of them just have few scores, so we just pick 8 most frequently scored ones, and we split training and test sets equally. The semantic descriptor is constructed by concatenating 8-bit domain and 3-bit task indicator. Conventional MTL interpretations of this dataset consider $8 \times 3 = 24$ atomic tasks. Thus the task overlap across domain or

domain overlap across task is ignored. Results in Table 6 shows that our approach outperforms this traditional MTL setting by better representing it as a distributed MDMT problem.

Table 6: Restaurant & Consumer Dataset (RMSE)

LR	RMTL	FEDA	MTFL	GO-MTL	Ours
2.32	1.23	1.17	1.13	1.06	**0.78**

5 CONCLUSION

In this paper we proposed a unified framework for multi-domain and multi-task learning. The core concept is a semantic descriptor for tasks or domains. This can be used to unify and improve on a variety of existing multi-task learning algorithms. Moreover it naturally extends the use of a single categorical variable to index domains/tasks to the multivariate case, which enables better information sharing where additional metadata is available. Beyond multi task/domain learning, it enables the novel task of zero-shot domain adaptation and provides an alternative pipeline for zero-shot learning.

Neural networks have also been used to address MTL/MDL by learning shared invariant features (Donahue et al., 2015). Our contribution is complementary to this (as demonstrated e.g., with AwA) and the approaches are straightforward to combine by placing more complex structure on left-hand side $f_P(\cdot)$. Our future directions are: (i) The current semantic descriptor is formed by discrete variables. We want to extend this to continuous and periodic variable like the pose, brightness and time. (ii) We assume the semantic descriptor (task/domain) is always observed, an improvement for dealing with a missing descriptor is also of interest.

Acknowledgements We gratefully acknowledge the support of NVIDIA Corporation for the donation of the GPUs used for this research.

REFERENCES

Argyriou, A., Evgeniou, T., and Pontil, M. Convex multi-task feature learning. *Mach. Learn.*, 73 (3):243–272, December 2008.

Beijbom, O. Domain adaptations for computer vision applications. Technical report, UCSD, 2012.

Blitzer, J., Foster, D. P., and Kakade, S. M. Zero-shot domain adaptation: A multi-view approach. Technical report, 2009.

Daumé III, H. Frustratingly easy domain adaptation. In *ACL*, 2007.

Deng, J., Ding, N., Jia, Y., Frome, A., Murphy, K., Bengio, S., Li, Y., Neven, H., and Adam, H. Large-scale object classification using label relation graphs. In *ECCV*, pp. 48–64, 2014.

Ding, Z., Ming, S., and Fu, Y. Latent low-rank transfer subspace learning for missing modality recognition. In *AAAI*, pp. 1192–1198, 2014.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, 2015.

Dredze, M., Kulesza, A., and Crammer, K. Multi-domain learning by confidence-weighted parameter combination. *Machine Learning*, 79(1-2):123–149, 2010.

Duan, L., Xu, D., and Chang, S.-F. Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach. In *CVPR*, 2012.

Evgeniou, T. and Pontil, M. Regularized multi–task learning. In *KDD*, 2004.

Frome, A., Corrado, G., Shlens, J., Bengio, S., Dean, J., Ranzato, M., and Mikolov, T. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.

Fu, Y., Hospedales, T., Xiang, T., Fu, Z., and Gong, S. Transductive multi-view embedding for zero-shot recognition and annotation. In *ECCV*, 2014.

Gong, B., Shi, Y., Sha, F., and Grauman, K. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012.

Hoffman, J., Darrell, T., and Saenko, K. Continuous manifold based adaptation for evolving visual domains. In *CVPR*, 2014.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

Joshi, M., Dredze, M., Cohen, W. W., and Rosé, C. P. Multi-domain learning: When do domains matter? In *EMNLP*, 2012.

Kang, Z., Grauman, K., and Sha, F. Learning with whom to share in multi-task feature learning. In *ICML*, 2011. ISBN 978-1-4503-0619-5.

Kishan Wimalawarne, M. S. and Tomioka, R. Multitask learning meets tensor factorization: task imputation via convex optimization. In *NIPS*, 2014.

Kressner, D., Steinlechner, M., and Vandereycken, B. Low-rank tensor completion by riemannian optimization. Technical Report 2, 2014.

Kumar, A. and Daumé III, H. Learning task grouping and overlap in multi-task learning. In *ICML*, 2012.

Lampert, C. H., Nickisch, H., and Harmeling, S. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

Larochelle, H., Erhan, D., and Bengio, Y. Zero-data learning of new tasks. In *AAAI*, 2008.

Mauch, M. and Ewert, S. The audio degradation toolbox and its application to robustness evaluation. In *ISMIR*, 2013.

Palatucci, M., Pomerleau, D., Hinton, G., and Mitchell, T. Zero-shot learning with semantic output codes. In *Neural Information Processing Systems (NIPS)*, 2009.

Passos, A., Rai, P., Wainer, J., and Daumé III, H. Flexible modeling of latent task structures in multitask learning. In *ICML*, 2012.

Qiu, Q., Patel, V. M., Turaga, P., and Chellappa, R. Domain adaptive dictionary learning. In *ECCV*, 2012. ISBN 978-3-642-33764-2.

Romera-paredes, B., Aung, H., Bianchi-berthouze, N., and Pontil, M. Multilinear multitask learning. In *ICML*, 2013.

Rosenstein, M. T., Marx, Z., Kaelbling, L. P., and Dietterich, T. G. To transfer or not to transfer. In *In NIPS05 Workshop, Inductive Transfer: 10 Years Later*, 2005.

Saenko, K., Kulis, B., Fritz, M., and Darrell, T. Adapting visual category models to new domains. In *ECCV*, 2010.

Socher, R., Ganjoo, M., Manning, C. D., and Ng, A. Y. Zero-shot learning through cross-modal transfer. In *NIPS*, pp. 935–943, 2013.

Sun, B. and Saenko, K. From virtual to reality: Fast adaptation of virtual object detectors to real domains. In *BMVC*, 2014.

Torralba, A. and Efros, A. A. Unbiased look at dataset bias. In *CVPR*, 2011.

Tzanetakis, G. and Cook, P. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, July 2002.

Vargas-Govea, B., González-Serna, G., and Ponce-Medellın, R. Effects of relevant contextual features in the performance of a restaurant recommender system. *ACM RecSys*, 11, 2011.

Xue, Y., Liao, X., Carin, L., and Krishnapuram, B. Multi-task learning for classification with dirichlet process priors. *Journal of Machine Learning Research*, 8:35–63, 2007.