Classifying Heterogeneous Sequential Data by Cyclic Domain Adaptation: An Application in Land Cover Detection

Xiaowei Jia*, Guruprasad Nayak*, Ankush Khandelwal*, Anuj Karpatne†, Vipin Kumar*

Abstract

Recent advances in processing remote sensing data have provided unprecedented potential for monitoring land covers. However, it is extremely challenging to deploy an automated monitoring system for different regions and across different years given the involved data heterogeneity over space and over time. The heterogeneity exists on two aspects. First, for many land covers, the distinguishing temporal patterns are only visible in certain discriminative period. Due to the change of weather conditions, the discriminative period can shift across space and time, which causes heterogeneity to the sequential data. Second, the collected remote sensing data are affected by acquisition devices and natural variables, e.g., precipitation and sunlight. In this paper, we introduce a novel framework to effectively detect land covers using the sequential remote sensing data. At the same time, we propose new learning strategies based on attention networks and domain adaptation to addresses the aforementioned challenges. The evaluation on two real-world applications - cropland mapping and burned area detection, demonstrate that the proposed method can effectively detect land covers under different weather conditions.

1 Introduction

Land Use and Land Cover (LULC) changes have drawn great attention from governments, companies and non-governmental organizations (NGOs) since they can provide promising insights for management of natural resources. For example, growth in the worlds population and the acceleration of industrialization are straining already scarce natural resources and food supplies. The land cover study can help monitor whether crop production is being scaled up (to keep pace with growing demand) at places that are most suitable from environmental perspective. Monitoring cropland varieties and planting area can also help analyze the consumption of water and energy in tillage, irrigation and harvesting, as well as the contaminants caused by fertilizers.

Advances in earth observation technologies have led to the acquisition of vast amounts of timely and reliable remote sensing data that can be used for monitoring changes on a large scale. Many existing land cover products are manually created through visual interpretation, which takes advantage of human expertise in the labeling process [5]. The limitations of this approach are manifold. First, manual labeling may result in both false positives and false negatives due to observational mistakes. Second, this approach usually requires multiple observers to delineate land covers. Their own subjective biases can result in inconsistent results. Most importantly, the required substantial human resources make it infeasible for large regions at yearly scale.

Hence, researchers are pursuing data-driven approaches to build automated monitoring system [8, 15–17, 26]. Although these methods have shown success for local regions or a specific year given sufficient training data, they greatly suffer from the involved data heterogeneity over space and time. For example, crops can be planted under different soil types, precipitation and other weather conditions for different places and different years. Therefore, a classification model learned from a specific region or a specific year cannot be generalized to other regions and time periods.

An intuitive solution to address data heterogeneity is to use domain adaptation techniques [12], which aim to train a classification model given that the joint distribution $P(X, Y)$ differs between training data and testing data. However, most existing domain adaptation approaches focus on static data while the complexity of earth system makes many land covers not able to be identified on a single date. Instead, the successful detection requires the discovery of distinctive temporal patterns from a sequence of collected data. More critically, many land covers only show distinctive temporal patterns during certain period of a year, which is also referred to as the discriminative period [18]. For example, croplands can be identified by analyzing their growing patterns in certain part of growing season, but they look similar to barren land after they are harvested. Likewise, when identifying burned area, we need to focus on fire seasons and burning scars.

Given these data characteristics, the heterogeneity for sequential data can be summarized from two aspects.

*University of Minnesota. {jiaxx221,nayak013,khand035, kumar001}@umn.edu

†Virginia Tech. karpatne@vt.edu

First, the discriminative period can shift across years and across regions. For example, farmers plant and harvest crops in different time across years due to weather conditions. Second, even after we locate the discriminative period from a long sequence, the obtained features in the discriminative period can still vary across regions and years. This is because the data collected by the optical sensors in satellites are affected by climate variables, such as precipitation and sunlight.

In this paper, we propose a novel learning framework, **D**omain **A**daptation for **S**equential data (DAS), which combines Long-Short Term Memory (LSTM) and attention model [25] to discover temporal patterns from the discriminative period. We first apply the Long-Short Term Memory (LSTM) model to capture long-term temporal dependencies in sequential remote sensing data, which are critical for land cover changes due to long-term climate impacts [16]. After we embed the raw input using LSTM, we utilize the attention model to capture the discriminative period in the entire sequence.

To overcome the challenges brought by data heterogeneity, we utilize the adversarial learning technique to learn a mapping between the data collected under different weather conditions. Combining with the attention model, we develop a new domain adaptation method which pays more attention to the discriminative period. This is essential for transferring discriminative knowledge across domains because the non-disriminative periods commonly involve much variability. Consider the burned area mapping as an example. Since fires can occur on a variety of land cover types (e.g., forest, savannas, crops), adaptation on the discriminative period is especially helpful for identifying burned areas because the model will not have to adapt the high variability in different land cover types before the fire period. Furthermore, to ensure the robustness of the attention model against different weather conditions, we propose a new learning strategy by advancing cyclic GAN model [20, 35].

Our evaluation on two real-world applications, cropland mapping and burned area detection, shows the superiority of DAS in classification performance over multiple baselines. Besides, we demonstrate that the proposed domain adaptation technique can reduce the impact of domain variation on the effectiveness of attention model. Finally, we discuss some interpretation for the variation across domains.

2 Related Work

Domain adaptation (DA) aims to leverage abundant labeled data from a source domain to learn a discriminative classifier and then generalize it to a target domain despite the data distribution discrepancy between the source and target domains. This situation is most common in earth observation data where same datasets are available at a global scale, but the data distribution varies across regions and across years due to spatial and temporal heterogeneity.

Existing DA techniques can be divided into two categories - semi-supervised DA [9] and unsupervised DA [12]. In this work, we focus on the unsupervised DA where we assume no labeled data are available in the target domain. The proposed method can be easily generalized to the semi-supervised case.

DA techniques have shown success in a variety of applications [19, 20, 24]. Many existing works [14, 23] learn transferable features by minimizing Maximum Mean Discrepancies (MMD) between source domain and target domain, which measures the difference in both marginal and conditional distributions. With the recent advances in deep adversarial learning, researchers have proposed another approach to learn task specific features which are also consistent features across two domains. These features are learned by minimizing the classification accuracy of a well-trained domain classifier [11, 22, 33]. However, some of these approaches learn consistent features by training a single high-capacity classifier for data from different domains [11]. These approaches have limited applicability to a broad class of problems without sufficient labeled data to train a high-capacity model. An alternative approach is to learn a separate mapping function from the target domain to the source domain for adaptation [35].

Due to the data heterogeneity and paucity of labels, researchers have also applied DA techniques in remote sensing [31]. For example, Elshamli et al. [10] utilize an end-to-end neural network model to extract invariant features across domains, which can further assist in classifying remote sensing images. However, these approaches mostly focus on static data, i.e., individual image snapshots. In contrast, several approaches have been proposed for health-care data that explore the information transfer between multi-temporal data from different patient groups using RNN and its variants [28, 29, 34]. However, the approaches introduced in [28, 29, 34] simply use the extracted features from RNN-based models and apply them to transfer the knowledge. They treat all the time steps equally in recurrent models to connect different domains, and thus lack the ability to avoid the transfer of non-informative time steps. Consequently, these approaches can be adversely affected by the variability in the non-informative period.

3 Problem Definition

In this problem, we are provided with the data points from the source domain $\mathcal{S}$ and the target domain $\mathcal{T}$. We

represent these data points as $X_{\mathcal{S}}$ and $X_{\mathcal{T}}$, respectively. Here each domain can be instantiated as a specific year, a region or a scenario with certain weather conditions.

Each data point from $X_{\mathcal{S}}$ or $X_{\mathcal{T}}$ contains multivariate spectral features for T time steps, e.g., a sample $x_{\mathcal{S},i}$ from the $\mathcal{S}$ can be expressed as $\{x^1_{\mathcal{S},i}, ..., x^T_{\mathcal{S},i}\}$, where $x^t_{\mathcal{S},i} \in \mathbb{R}^D$. Moreover, we have the label of each data point in the source domain $\mathcal{S}$, $Y_{\mathcal{S}} = \{y_{\mathcal{S},1}, ..., y_{\mathcal{S},N}\}$. Each label $y_{\mathcal{S},i}$ belongs to one of K land cover classes. We have no labeled data for the target domain. Hereinafter we omit the subscript i (sample index) when it causes no ambiguity.

Our objective is to predict the labels $Y_{\mathcal{T}}$ for data points in the target domain. To achieve this, we aim to train a classification model using the provided sequential data $X_{\mathcal{S}}, X_{\mathcal{T}}$ and labels $Y_{\mathcal{S}}$. Due to the shift of joint distribution across domains, i.e., $P(X_{\mathcal{S}}, Y_{\mathcal{S}}) \neq P(X_{\mathcal{T}}, Y_{\mathcal{T}})$, the obtained classifier from the source domain via traditional learning approaches cannot be directly applied to the target domain.

Besides the classification, we also wish to find the most discriminative time period for each sample in the source domain and the target domain. This provides interpretability to the classification result.

4 Method

In this section, we start with the Long-Short Term Memory (LSTM) networks and the attention model, which jointly model long-term land cover patterns and capture the discriminative period for classification. Then we propose a domain adaptation technique that handles the shift of both the data distribution and the discriminative period across domains.

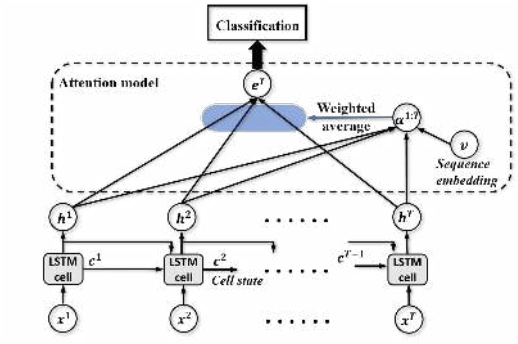


Figure 1: The structure of LSTM-Attention networks.

4.1 LSTM-Attention networks In this work, we utilize an LSTM-Attention networks model to detect the discriminative period from a sequence, which subsequently contributes to the classification. The structure of LSTM-Attention networks is shown in Fig. 1. The LSTM model generates hidden representation/embeddings h^t at every time step. These embeddings are then combined by the attention model via weighted summation for the final classification. We now briefly introduce the LSTM model and the attention model. For simplicity, we ignore the domain notation $\mathcal{S}$ when we describe the LSTM-Attention networks.

As an extension of standard recurrent neural networks (RNN), the LSTM model defines a transition relationship through an LSTM cell, which takes the input of features x^t at current time step and the inherited information from previous time steps.

Each LSTM cell contains a cell state c^t, which serves as a memory and allows the hidden units h^t to reserve information from the past. The cell state c^t is generated by combining c^{t-1}, h^{t-1} and the input features at t. Hence the transition of cell state over time forms a memory flow, which enables the modeling of long-term dependencies. Specifically, we first generate a new candidate cell state $\tilde{c}^t$ by combining x^t and h^{t-1} into a $\tanh(\cdot)$ function, as:

$$\tilde{c}^t = \tanh(W^c_h h^{t-1} + W^c_x x^t), \tag{4.1}$$

where $W^c_h \in \mathbb{R}^{H\times H}$ and $W^c_x \in \mathbb{R}^{H\times M}$ denote the weight parameters used to generate candidate cell state. Hereinafter we omit the bias terms as they can be absorbed into weight matrices. Then we generate a forget gate layer $f^t \in \mathbb{R}^H$, an input gate layer $g^t \in \mathbb{R}^H$ and an output gate layer o^t using the sigmoid function:

$$\begin{aligned} f^t &= \sigma(W^f_h h^{t-1} + W^f_x x^t), \\ g^t &= \sigma(W^g_h h^{t-1} + W^g_x x^t), \\ o^t &= \sigma(W^o_h h^{t-1} + W^o_x x^t), \end{aligned} \tag{4.2}$$

where $\{W^f_h \in \mathbb{R}^{H\times H}, W^f_x \in \mathbb{R}^{H\times M}\}$ and $\{W^g_h \in \mathbb{R}^{H\times H}, W^g_x \in \mathbb{R}^{H\times M}\}$ denote two sets of weight parameters for generating forget gate layer f^t and input gate layer g^t, respectively. The forget gate layer is used to filter the information inherited from c^{t-1}, and the input gate layer is used to filter the candidate cell state at time t. In this way we obtain the new cell state c^t and the hidden representation as follows:

$$\begin{aligned} c^t &= f^t \otimes c^{t-1} + g^t \otimes \tilde{c}^t, \\ h^t &= o^t \otimes \tanh(c^t), \end{aligned} \tag{4.3}$$

where $\otimes$ denotes entry-wise product.

After obtaining the hidden representation $\{h^1, ..., h^T\}$ from LSTM, we use an attention model to determine the discriminative period from the sequential data. The attention model aims to enforce the classifier to attend to different time steps based on different relevance scores. The higher relevance score at a time step indicates more expressed discriminative knowledge at this time step. In land cover problem, the time steps with higher relevance scores usually correspond to growing seasons (for cropland monitoring) and fire seasons (for burned area detection).

Specifically, we measure the relevance score of each time step t according to the similarity between its

hidden representation h^t and the sequence embedding $v \in \mathbb{R}^H$. Here v represents an embedding of the entire sequence, which has the same dimensionality with the hidden representation, and is jointly learned during the training process [21]. In the simplest case, we can embed $x^{1:T}$ into v using another LSTM.

More formally, the relevance score of time step t is computed as the inner-product between v and h^t. To normalize the relevance scores over all the time steps, we also apply a softmax function on the inner-product:

$$\alpha^t = \text{softmax}(v^{\mathrm{T}} h^t). \tag{4.4}$$

Then we aggregate h^t from all the time steps based on α, and apply a softmax function for classification:

$$\hat{y} = \text{softmax}(W_y \sum_t \alpha^t h^t), \tag{4.5}$$

where $W_y \in \mathbb{R}^{K \times E}$ denotes the parameters to transform aggregated hidden representation to the classification output $\hat{y}$. Here we use $\hat{y}$ to distinguish with the provided ground-truth labels y.

We train the LSTM-Attention networks using the labeled data from the source domain, i.e., $X_{\mathcal{S}}$ and $Y_{\mathcal{S}}$. We adopt the cross-entropy loss to define the training objective function, as follows:

$$\mathcal{J}_{sup} = \frac{1}{N_{\mathcal{S}}} \sum_i \sum_k y_{\mathcal{S},i,k} \log \hat{y}_{\mathcal{S},i,k}, \tag{4.6}$$

where $N_{\mathcal{S}}$ denotes the number of samples in the source domain. The provided label $y_{\mathcal{S}}$ is expressed in a one-hot representation where $y_{\mathcal{S},i,k} = 1$ if the i^{th} sample from the source domain belongs to class k.

4.2 Domain Adaptation Having described the LSTM-Attention model for discriminative period detection. Now we propose a domain adaptation approach so that the learned model can be applied to other places and other years under different weather conditions.

As mentioned earlier, we train an LSTM-Attention model using the data $X_{\mathcal{S}}$ and $Y_{\mathcal{S}}$ from the source domain. However, this model cannot be directly applied to $X_{\mathcal{T}}$ due to the shift of data distribution across domains, i.e., $P_{\mathcal{S}}(X, Y) \neq P_{\mathcal{T}}(X, Y)$. To apply the model to the target domain, we wish to first learn a transformation function from the target domain to the source domain $g : \mathcal{T} \to \mathcal{S}$. Such transformation aims to map the data in target domain to the similar distribution with the source domain. Then we will use the transformed data $g(X_{\mathcal{T}})$ as input to the learned LSTM-Attention model.

An effective domain adaptation process for our problem requires two properties: 1) Since the non-informative period in sequential data contains much variability and it is not relevant to the classification, the domain adaptation process should focus on the discrminative period. 2) After applying the attention model to the data transformed from the target domain (e.g., $g(X_{\mathcal{T}})$), the attention model should still be able to precisely locate the discriminative period.

Table 1: Notation used for cyclic domain adaptation.

Symbol	Expression	Meaning
$\bar{x}_{\mathcal{S}}$	$f(x_{\mathcal{S}})$	transformed data from $\mathcal{S}$ to $\mathcal{T}$
$\bar{x}_{\mathcal{T}}$	$g(x_{\mathcal{T}})$	transformed data from $\mathcal{T}$ to $\mathcal{S}$
$\tilde{x}_{\mathcal{S}}$	$g \circ f(x_{\mathcal{S}})$	reconstructed data from $\mathcal{S}$
$\tilde{x}_{\mathcal{T}}$	$f \circ g(x_{\mathcal{T}})$	reconstructed data from $\mathcal{T}$

To meet these requirements, we develop a new adversarial learning model based on cyclic GAN [35]. In this model, besides the mapping function $g : \mathcal{T} \to \mathcal{S}$, we also introduce another function $f : \mathcal{S} \to \mathcal{T}$, which maps the data from the source domain to the target domain. Given the mapping function f, we can better adapt the knowledge learned from the source domain to the target domain. This is especially helpful for transferring the knowledge of discriminative periods learned by the attention model from the source domain. For the ease of presentation, we define several notations in Table 1.

Since the LSTM-Attention model is trained on the source domain, it can generate meaningful outputs only given inputs from the source domain or data transformed to the source domain (e.g., $x_{\mathcal{S}}$, $\bar{x}_{\mathcal{T}}$, and $\tilde{x}_{\mathcal{S}}$). Using the LSTM model, we generate three sets of hidden representation, as follows:

$$\begin{aligned} h_{\mathcal{S}}^{1:T} &= \text{LSTM}(x^1_{\mathcal{S},(k)}, x^2_{\mathcal{S},(k)}, ..., x^T_{\mathcal{S},(k)}) \\ \bar{h}_{\mathcal{T}}^{1:T} &= \text{LSTM}(\bar{x}^1_{\mathcal{T},(k)}, \bar{x}^2_{\mathcal{T},(k)}, ..., \bar{x}^T_{\mathcal{T},(k)}) \\ \tilde{h}_{S}^{1:T} &= \text{LSTM}(\tilde{x}^1_{\mathcal{S},(k)}, \tilde{x}^2_{\mathcal{S},(k)}, ..., \tilde{x}^T_{\mathcal{S},(k)}), \end{aligned} \tag{4.7}$$

where $h_{\mathcal{S}}$ denotes the hidden representation for $x_{\mathcal{S}}$, $\bar{h}_{\mathcal{T}}$ denotes the hidden representation for the transformed testing data $g(X_{\mathcal{T}})$, and $\tilde{h}_{\mathcal{S}}$ denotes the hidden representation for the reconstructed training data $g \circ f(X_{\mathcal{S}})$.

To learn the mapping relationships between $\mathcal{S}$ and $\mathcal{T}$, we wish to minimize the difference between $h_{\mathcal{S}}$ and $\bar{h}_{\mathcal{T}}$, and between $\bar{h}_{\mathcal{T}}$ and $\tilde{h}_{\mathcal{S}}$. For the first group (i.e., $h_{\mathcal{S}}$ and $\bar{h}_{\mathcal{T}}$), the intuition is to first transfer the data $X_{\mathcal{T}}$ to the source domain (i.e., $\bar{X}_{\mathcal{T}}$) and then reduce its divergence with the training data (i.e., $X_{\mathcal{S}}$). In contrast, for the second group (i.e., $\bar{h}_{\mathcal{T}}$ and $\tilde{h}_{\mathcal{S}}$), we first map $X_{\mathcal{S}}$ to the target domain via $f(\cdot)$ and then compare with $X_{\mathcal{T}}$ in the target domain. However, since the hidden representation can only be computed from the source domain, we apply another $g(\cdot)$ mapping for both $X_{\mathcal{T}}$ and $f(X_{\mathcal{S}})$ to obtain $\bar{h}_{\mathcal{T}}$ and $\tilde{h}_{\mathcal{S}}$.

Since there exist no coupled correspondence between data in the source domain and the target domain, we adopt the adversarial regularization [11], which enforces that the data from different domains cannot be easily distinguished by an extra well-trained classifier.

Instead of directly conducting adversarial regularization on the hidden representation, which have been

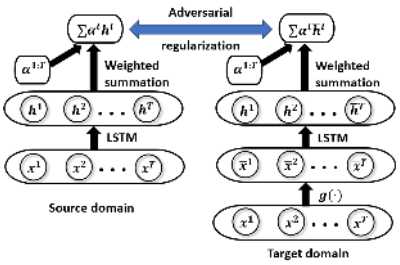


Figure 2: The adversarial regularization on weighted summation of hidden representation between $X_{\mathcal{S}}$ and $\bar{X}_{\mathcal{T}}$. We also include another regularization term between $\bar{X}_{\mathcal{T}}$ and $\tilde{X}_{\mathcal{S}}$ in DAS.

widely adopted in previous works [28,29], here we apply the adversarial training on the weighted summation of hidden representation at different time steps (Fig. 2). In this way, the adaptation process can assign higher weights for the periods with more discriminative information. Specifically, we utilize the relevance scores computed from attention model as the weight for each time step. We first compute the relevance scores for $x_{\mathcal{S}}$, $\bar{x}_{\mathcal{T}}$, and $\tilde{x}_{\mathcal{S}}$, using the attention model, as follows:

$$\begin{aligned}\alpha_{\mathcal{S}}^{1:T} &= \text{Attention}(x_{\mathcal{S},(k)}^1, x_{\mathcal{S},(k)}^2, ..., x_{\mathcal{S},(k)}^T)\\ \bar{\alpha}_{\mathcal{T}}^{1:T} &= \text{Attention}(\bar{x}_{\mathcal{T},(k)}^1, \bar{x}_{\mathcal{T},(k)}^2, ..., \bar{x}_{\mathcal{T},(k)}^T)\\ \tilde{\alpha}_{\mathcal{S}}^{1:T} &= \text{Attention}(\tilde{x}_{\mathcal{S},(k)}^1, \tilde{x}_{\mathcal{S},(k)}^2, ..., \tilde{x}_{\mathcal{S},(k)}^T)\end{aligned} \tag{4.8}$$

Then we define the adversarial loss between $X_{\mathcal{S}}$ and the transformed data $\bar{X}_{\mathcal{T}}$ as follows:

$$\begin{aligned}\mathcal{J}_{\mathcal{S}} = \sup_{D_{\mathcal{S}}} \sum_k \mathbb{E}_{h,\alpha|x\sim X_{\mathcal{S},(k)}} \log D_{\mathcal{S}}(\sum_t \alpha_{\mathcal{S},(k)}^t h_{\mathcal{S},(k)}^t)\\ + \mathbb{E}_{\bar{h},\bar{\alpha}|x\sim X_{\mathcal{T},(k)}} \log(1 - D_{\mathcal{S}}(\sum_t \bar{\alpha}_{\mathcal{T},(k)}^t \bar{h}_{\mathcal{T},(k)}^t)),\end{aligned} \tag{4.9}$$

where $k \in [1, 2, ..., K]$ is the index for different classes. Here we assume we have the provided labels for $X_{\mathcal{S}}$ and the pseudo-labels for $X_{\mathcal{T}}$. We will discuss how to generate these pseudo-labels in Section 4.3. $D_{\mathcal{S}}$ represents a domain classifier that maps R^H to [0,1]. The $\sup_{D_{\mathcal{S}}}$ operation aims to find the optimal classifier to distinguish between $h_{\mathcal{S}}$ and $\bar{h}_{\mathcal{T}}$. On the other hand, by minimizing $\mathcal{J}_{\mathcal{S}}$, we will reduce the classification performance by the optimal $D_{\mathcal{S}}$. This ensures that after transforming $X_{\mathcal{T}}$ to the source domain through the mapping function $g(\cdot)$, the transformed data cannot be easily distinguished with the original training data $X_{\mathcal{S}}$ even by a well-trained classifier.

Similarly, we define the adversarial loss between $\bar{h}_{\mathcal{T}}$ and $\tilde{h}_{\mathcal{S}}$ with another classifier $D_{\mathcal{T}}$, as follows:

$$\begin{aligned}\mathcal{J}_{\mathcal{T}} = \sup_{D_{\mathcal{T}}} \sum_k \mathbb{E}_{\bar{h},\bar{\alpha}|x\sim X_{\mathcal{T},(k)}} \log D_{\mathcal{T}}(\sum_t \bar{\alpha}_{\mathcal{T},(k)}^t \bar{h}_{\mathcal{T},(k)}^t)\\ + \mathbb{E}_{\tilde{h},\tilde{\alpha}|x\sim X_{\mathcal{S},(k)}} \log(1 - D_{\mathcal{T}}(\sum_t \tilde{\alpha}_{\mathcal{S},(k)}^t \tilde{h}_{\mathcal{S},(k)}^t))\end{aligned} \tag{4.10}$$

By incorporating the cost of $\mathcal{J}_{\mathcal{S}}$ and $\mathcal{J}_{\mathcal{T}}$, the model can learn functions $f(\cdot)$ and $g(\cdot)$ to map data from one domain to the similar distribution with the other domain. However, in practice, the computation of relevance scores (using the learned model from $\mathcal{S}$) for transformed data from $\mathcal{T}$ can be adversely affected by the heterogeneity across domains. As we compute $\bar{\alpha}_{\mathcal{T}}$ and $\tilde{\alpha}_{\mathcal{S}}$ using the transformed data from the target domain (i.e., which involve $g(\cdot)$ mapping), it is critical to have a robust attention model so that it can produce accurate estimation for these relevance scores.

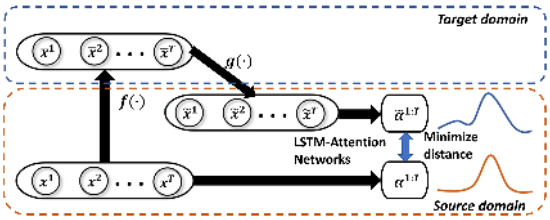


Figure 3: The loss of attention shift in domain adaptation process. Here we only illustrate the loss between $x_{\mathcal{S}}$ and $f(x_{\mathcal{S}})$. In DAS, we also include another term for the loss between $\bar{x}_{\mathcal{T}}$ and $f(\bar{x}_{\mathcal{T}})$.

To this end, we define another loss function which accounts for the attention shift across domains. Intuitively, after we map each instance from the source domain to the target domain, the discriminative period should stay the same. Here we consider two mapping directions. First, after we transform a sample $x_{\mathcal{S}}$ to the target domain via $f(\cdot)$, $f(x_{\mathcal{S}})$ should have the same relevance scores with $x_{\mathcal{S}}$ (Fig. 3). Since LSTM-Attention model only takes the input from the source domain, the relevance scores of $f(x_{\mathcal{S}})$ can be computed after applying another $g(\cdot)$ function, which is equivalent to $\tilde{\alpha}_{\mathcal{S}}$. Second, we consider the attention shift between $\bar{x}_{\mathcal{T}}$ and $f(\bar{x}_{\mathcal{T}})$ (equivalent to $\tilde{x}_{\mathcal{T}}$). We first compute the the hidden representation and relevance scores for $\tilde{x}_{\mathcal{T}}$, as:

$$\begin{aligned}h_{tc}^{1:T} &= \text{LSTM}(g(\tilde{x}_{\mathcal{T}}^1), g(\tilde{x}_{\mathcal{T}}^2), ..., g(\tilde{x}_{\mathcal{T}}^T)),\\ \alpha_{tc}^{1:T} &= \text{Attention}(g(\tilde{x}_{\mathcal{T}}^1), g(\tilde{x}_{\mathcal{T}}^2), ..., g(\tilde{x}_{\mathcal{T}}^T))\end{aligned} \tag{4.11}$$

Then we define the loss for attention shift as follows:

$$\mathcal{J}_{att} = \mathbb{E}_{\alpha,\tilde{\alpha}|x\sim X_{\mathcal{S}}} \|\alpha_{\mathcal{S}} - \tilde{\alpha}_{\mathcal{S}}\|^2 + \mathbb{E}_{\bar{\alpha},\alpha_{tc}|x\sim X_{\mathcal{T}}} \|\bar{\alpha}_{\mathcal{T}} - \alpha_{tc}\|^2 \tag{4.12}$$

On the other hand, even though we can learn $f(\cdot)$ and $g(\cdot)$ to map data from one domain to the other by reducing $\mathcal{J}_{\mathcal{S}}$ and $\mathcal{J}_{\mathcal{T}}$, these mapping functions cannot guarantee that an individual input sample x and the output (e.g., $f(x)$ or $g(x)$) are paired up in a meaningful way. Consider an example with two input training samples $\{x_{\mathcal{S},1}, x_{\mathcal{S},2}\}$ and the expected corresponding outputs in the target domain $\{x_{\mathcal{T},1}, x_{\mathcal{T},2}\}$. If the model mistakenly learns a mapping such that $f(x_{\mathcal{S},1}) = x_{\mathcal{T},2}$ and $f(x_{\mathcal{S},2}) = x_{\mathcal{T},1}$, the output distribution is still the same as the expected output distribution. Moreover, the optimization process in practice commonly leads to the well-known problem of mode collapse [13], where all input data map to the same output data.

To address these challenges, we introduce an addi-

tional cyclic self-reconstruction cost $\mathcal{J}_{cyc}$ so that each sample from the source domain can be recovered after a composite $g \circ f$ mapping and each sample from the target domain can be recovered after $f \circ g$ mapping. In this way, we can define the cyclic self-reconstruction loss, as:

$$\mathcal{J}_{cyc} = \mathbb{E}_{x \sim X_{\mathcal{S}}} \sum_t \left\| x_{\mathcal{S}}^t - \tilde{x}_{\mathcal{S}}^t \right\|^2 + \mathbb{E}_{x \sim X_{\mathcal{T}}} \sum_t \left\| x_{\mathcal{T}}^t - \tilde{x}_{\mathcal{T}}^t \right\|^2 \tag{4.13}$$

Combining the aforementioned loss objectives, the overall loss function can be expressed as:

$$\mathcal{J} = \mathcal{J}_{sup} + \lambda(\mathcal{J}_{\mathcal{S}} + \mathcal{J}_{\mathcal{T}}) + \mu \mathcal{J}_{cyc} + \gamma \mathcal{J}_{att}, \tag{4.14}$$

where λ, μ and γ are hyper-parameters to control the weight of each loss function.

4.3 Learning process The proposed learning framework can be trained in a recursive EM-style fashion. In E-step, we aim to assign pseudo labels to data in the target domain. In M-step, we will update the model parameters and estimate mapping function $f(\cdot)$ and $g(\cdot)$. We now present the details in E-step and M-step.

E-step: For each data instance $x_{\mathcal{T}}$ in the target domain, we first transform it to the source domain through $g(\cdot)$. Then we apply the LSTM-Attention model (Section 4.1) to determine the posterior probability of $\mathrm{P}(y|g(x_{\mathcal{T}}))$ (Eq. 4.5). Then we sample the pseudo-labels $Y_{\mathcal{T}}$ according to this probabilistic distribution.

M-step: We implement M-step in two stages. First, we wish to update the domain classifiers $D_{\mathcal{S}}$ and $D_{\mathcal{T}}$. These classifiers can be trained by maximizing the objective described in Eqs. 4.9 and 4.10. The training of domain classifiers involves the provided data $X_{\mathcal{S}}$, $X_{\mathcal{T}}$ and the labels $Y_{\mathcal{S}}$, $Y_{\mathcal{T}}$.

Next, we update the parameters in LSTM-attention networks and in $f(\cdot)$ and $g(\cdot)$ by minimizing the objective function in Eq. 4.14. This can be implemented by standard back-propagation algorithm.

The time complexity for the learning process is $O((N_{\mathcal{T}} + N_{\mathcal{S}})T\eta)$, where η is a constant factor determined by the dimensionality of input features, hidden representation, and the number of classes. The detailed setting for hyper-parameters and network architecture will be discussed in Section 5.

5 Experimental Results

We evaluate the proposed algorithm in two applications, cropland mapping and burned area detection. We first describe the MODIS dataset used to populate the input sequential features for both applications.

We utilize MODIS MOD09A1 multi-spectral data product [2], collected by MODIS instruments onboard NASA's Terra satellites. This dataset provides global data for every 8 days at 500m spatial resolution. At each date, MODIS dataset provides reflectance values on 7 spectral bands (620-2155 nm) for every location.

We compare the proposed method against several baselines, including static methods: Artificial Neural Networks (three-layer ANN) and Random Forest (RF) (most popular in remote sensing) that are applied to the concatenation of data, sequential models: standard LSTM and LSTM-Attention networks (ATT), as well as advanced baselines:

ATT+MMD (ATMMD): We first learn the LSTM-Attention networks for the source domain, and estimate the mapping function $g(\cdot)$ by minimizing MMD between the source domain and the target domain [27].

ATT+ADV (ATADV): In this baseline, we utilize the standard adversarial learning method [13, 32] for estimating the mapping function $g(\cdot)$.

ATT+ADV2 (ATADV2): Rather than using the mapping function $g(\cdot)$, we train an LSTM-Attention model with adversarial learning using the data from both source and target domains such that their hidden representation cannot be easily distinguished [11].

For the baselines with domain adaptation process (ATMMD, ATADV, ATADV2), we also utilize the same EM learning strategy as discussed in Section 4.3. When deploying DAS, we set both $f(\cdot)$ and $g(\cdot)$ to be a two-layer neural networks with 30 hidden units. For the LSTM model, we utilize 50 hidden units, i.e., $H = 50$.

5.1 Learning tasks and dataset description *Cropland mapping:* We aim to distinguish between corn and soybean in southwestern Minnesota, US. The annual ground-truth information on these two classes is provided by USDA Crop Data Layer product [1]. Although this product also provides labels for other crop types, previous survey study shows that the labels are more reliable for major crops like corn and soybean.

This task is challenging in agricultural domain because corn and soybean frequently look similar in most single dates of a year but are more likely to be identified using the temporal profile at certain stage [30]. Also, the crops in different places and different years can look different due to the weather conditions.

We utilize the data from 2016 as the source domain. We select 1,000 balanced training data points from a region where farmers plant the same crop type in 2015 and in 2016. In this way, the residues left on the ground at the beginning of the 2016 are consistent with the crops planted in the growing season.

We conduct three different tests: 1) Group-test: We predict the crop types for a set of locations from a different region in Minnesota in 2016. These locations have different types of residues at the beginning of the year. 2) 2015-test: We apply the model to the same region in 2015. 3) 2011-test: We apply the model

Table 2: Classification performance of each method in cropland mapping: training data (source domain), Group-test, 2015-test and 2011-test, and in burned area detection: training data (source domain), Group-test, Year-test and Region-test.

	Cropland mapping								**Burned area detection**							
	Train		***Group-test***		***2015-test***		***2011-test***		***Train***		***Group-test***		***Year-test***		***Region-test***	
Method	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
ANN	0.863	0.797	0.763	0.731	0.665	0.679	0.664	0.602	0.940	0.905	0.897	0.708	0.782	0.504	0.768	0.435
RF	0.863	0.788	0.775	0.734	0.672	0.688	0.662	0.523	0.939	0.908	0.904	0.709	0.773	0.367	0.767	0.403
LSTM	0.865	0.807	0.786	0.734	0.767	0.718	0.762	0.704	0.943	0.916	0.901	0.730	0.801	0.576	0.772	0.517
ATT	0.909	0.811	0.828	0.767	0.786	0.731	0.775	0.712	0.965	0.935	0.925	0.743	0.835	0.637	0.796	0.629
ATMMD	0.909	0.811	0.836	0.785	0.799	0.753	0.779	0.721	0.965	0.935	0.931	0.815	0.859	0.677	0.846	0.679
ATADV	0.909	0.811	0.842	0.796	0.819	0.756	0.784	0.727	0.965	0.935	0.934	0.822	0.865	0.692	0.879	0.682
ATADV2	0.895	0.810	0.833	0.785	0.805	0.757	0.776	0.714	0.949	0.920	0.935	0.822	0.863	0.704	0.875	0.679
DAS	0.909	0.811	0.867	0.819	0.844	0.787	0.832	0.756	0.965	0.935	0.952	0.886	0.918	0.725	0.895	0.722

to the same region in 2011. Different environmental conditions across years results in the variation of multi-spectral features. The variation is even more obvious between 2016 and 2011 according to the weather history in Minnesota [6]. Each test is conducted on a balanced dataset with 2,000 selected data points.

Burned area detection: In this application, we wish to detect burned area across regions and across years using limited manually labeled data. We randomly select 1,000 burned locations and 1,000 normal locations in California, US 2008 as training data (i.e., the source domain). The burned area samples are only taken from the locations which used to be the forests. We obtained fire validation data from government agencies responsible for monitoring and managing forests and wildfires [4]. For the land cover information before the fire period, we refer to NASA land cover dataset [3].

We apply each method to three tests: 1) Group-test: we test on a region in California 2008 which contains 1,200 burned locations and 5,800 normal locations. Here the fires occur on the woody savannas rather than forests. 2) Year-test: we test on a region with forest fire in California 2007. The testing data contains 300 burned locations and 1,200 normal locations. 3) Region-test: we apply each method to detect burned area in Montana, US 2007. The selected testing data contain 2,000 burned locations and 9,000 normal locations.

5.2 Classification performance In Table 2, we report the performance of each method in terms of Area Under Curve (AUC) and F-1 score. We can observe that DAS outperforms other baselines by a considerable margin for all the three tests. Compared with cropland mapping, the F1-scores in burned area detection are lower due to the skewness of the testing data.

The comparison between LSTM and static baselines (ANN and RF) shows that the modeling of temporal profile can help detect the crop type. The improvement from LSTM to ATT shows that the attention model assists in further improving the classification performance by explicitly modeling the discriminative period.

In general, the performance of each method is degraded in all other test domains compared with the training domain. The domain adaptation-based baseline approaches (ATMMD, ATADV, ATADV2) shows superior performance compared with ATT since they can potentially reduce the divergence between source and target domains. However, they are limited in their ability to properly adapt the information of discriminative period to the target domains, and thus their performance is inferior to that of DAS.

ATADV slightly outperforms ATADV2 in cropland mapping, but they are similar in burned area detection. This is because burned areas show more distinctive signatures than croplands. Hence, it is more likely to have a single classifier that can identify burned areas under different weather conditions.

5.3 Impact of data heterogeneity Now we inspect the impact on the attention model by the data heterogeneity. Figs. 4 (a) and (b) show the obtained relevance scores for the corn locations (in cropland mapping) in 2016 by LSTM-Attetion networks (ATT) and the obtained relevance scores by ATT and DAS in 2015 and 2011, respectively. Figs. 4 (c) and (d) depict the obtained relevance scores for the burned locations (by forest fires) in California, 2008 by ATT and the obtained relevance scores by ATT and DAS for Group-test (woody savannas fires) and Region-test (Montana, 2007), respectively. It is noteworthy that in Fig. 4 (c) the discriminative periods are expected to be the same between training and testing since fires occur at the same region and in the same year.

For both tests, we can observe that the LSTM-Attention networks cannot detect a period with larger relevance scores when directly applied to the testing

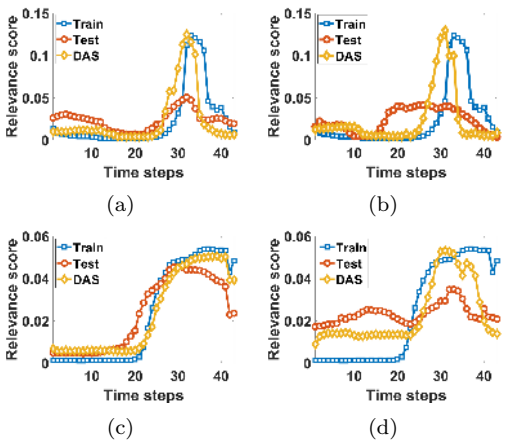


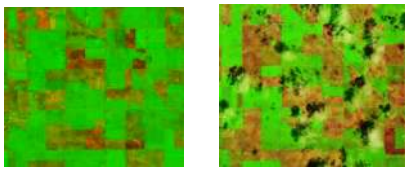
Figure 4: The impact of heterogeneity on the attention model in (a) 2015-test, and (b) 2011-test for cropland mapping, and (c) Group-test and (d) Region-test for burned area detection. Train: The relevance scores on training data. Test: the relevance scores on test data by directly applying LSTM-Attention networks. DAS: the relevance scores on test data by DAS.

scenario. Therefore it cannot precisely capture the discriminative period. In contrast, DAS is capable of mitigating the impact of variability across domains and thus producing meaningful relevance scores. Here the discriminative periods last longer than the periods detected in cropland mapping because fires commonly leave burning scars on the ground which also help identify burned locations.

We can also observe that the discriminative periods lasts longer in burned area detection than in cropland mapping because fires commonly leave burning scars on the ground which also help identify burned locations.

5.4 Discussion on domain variation We now discuss the interpretation of the variation across domains. As mentioned earlier, data heterogeneity exists on two aspects - the shift of discriminative period and the variation of multi-spectral data. For the shift of discriminative period, we can easily find the discriminative period for both domains $\mathcal{S}$ and $\mathcal{T}$ using the robust attention model learned by DAS. According to Fig. 4 (a), we can observe that the crops in 2015 are planted earlier than the crops in 2016. To verify this, we show high-resolution Landsat images around 25^{th} time step in 2015 (Fig. 5 (a)) and 2016 (Fig. 5 (b)). It can be seen that the selected region shows higher greenness level at this selected time in 2015 than in 2016.

More critically, we investigate the difference in multi-spectral data across domains using the learned mapping function $g(\cdot)$. This can help scientific researchers analyze the exact difference of land cover phenology across regions and years. As we standardized the



(a) (b)

Figure 5: The Landsat images (in RGB, 30 m resolution) for an example region in southwestern Minnesota in (a) 2016 and (b) 2015 at the beginning of July (corresponding to around 25^{th} time step in MODIS).

Table 3: The spectrum range of each spectral band, and the variation of each band in crop mapping (2011-test) and burned area detection (Region-test).

Band	Range(nm)	Crop	Burned area
Band 1	620-670	0.77±0.34	0.39±0.18
Band 2	841-876	1.71±0.55	0.40±0.15
Band 3	459-479	0.83±0.47	0.13±0.06
Band 4	545-565	0.96±0.34	0.59±0.28
Band 5	1230-1250	2.21±0.61	0.11±0.09
Band 6	1628-1652	1.15±0.34	0.45±0.22
Band 7	2105-2155	0.49±0.15	0.14±0.08

input in the training process, the features should fall in $\mathcal{N}(0, I)$. Then we randomly sample 20 data points from $\mathcal{N}(0, I)$ as input to $g(\cdot)$. For each spectral band (i.e., each feature), we measure the average and standard deviation of the absolute difference between the input value and the output value (Table 3).

For cropland mapping, the variation mostly occurs in Band 2, Band5 and Band 6. According to NASA's document on MODIS [7], Band 2 reflects "Vegetation Land Cover Transformation", Band 5 reflects "Leaf/Canopy Differences" and Band 6 reflects "Snow/Cloud Differences". The meaning of these bands conforms to our result since all these three factors play important roles in identifying crops in Minnesota.

For the burned area detection, the variation mostly occurs in Band 2, Band 4 and Band 6. Here Band 4 reflects "Green Vegetation". The meaning of these three bands also verify the correctness of our result since fires can directly impact the vegetation level.

6 Acknowledgement

This work was funded by the NSF Awards 1029711 and DTC seed grant. Access to computing facilities was provided by Minnesota Supercomputing Institute.

7 Conclusion

In this paper, we propose a framework DAS that utilizes the discriminative temporal information for domain adaptation. The results demonstrate the effectiveness of DAS in classifying land covers under different weather conditions and maintaining the robustness

of the attention model. Also, DAS can provide interpretations for the heterogeneity across domains.

Although the proposed advancements are motivated by land cover application, they are generally applicable to other applications as well. For example, the LSTM and attention model can be used to model discriminative patterns for disease progression in EHRs (Electronic Health Records). The proposed domain adaptation technique can be used for adapting the model across different patient groups.

References

[1] Global food security support analysis data - usda nass. https://geography.wr.usgs.gov/science/croplands.

[2] Modis mod09a1. https://lpdaac.usgs.gov/dataset_discovery/modis/modis_products_table/mod09a1.

[3] Modis web: Modis land cover type/dynamics. https://modis.gsfc.nasa.gov/data/dataprod/mod12.php.

[4] Monitoring trends in burn severity. https://www.mtbs.gov/.

[5] Mrlc nlcd 2011. https://www.mrlc.gov/nlcd2011.php.

[6] Weather forecast and report, weather underground. https://www.wunderground.com/.

[7] Modis overview. https://lpdaac.usgs.gov/dataset_discovery/modis, 2018.

[8] Gustavo Camps-Valls. Machine learning in remote sensing data processing. In *MLSP*. IEEE, 2009.

[9] Hal Daumé III. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*, 2009.

[10] Ahmed Elshamli, Graham W Taylor, Aaron Berg, and Shawki Areibi. Domain adaptation using representation learning for the classification of remote sensing images. *J-STARS*, 2017.

[11] Yaroslav Ganin et al. Domain-adversarial training of neural networks. *JMLR*, 2016.

[12] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*, 2014.

[13] Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *preprint arXiv:1701.00160*, 2016.

[14] Cheng-An Hou et al. Unsupervised domain adaptation with label and structural consistency. *TIP*, 2016.

[15] Xiaowei Jia, Ankush Khandelwal, James Gerber, Kimberly Carlson, Paul West, and Vipin Kumar. Learning large-scale plantation mapping from imperfect annotators. In *IEEE BigData*, 2016.

[16] Xiaowei Jia, Ankush Khandelwal, Guruprasad Nayak, James Gerber, Kimberly Carlson, Paul West, and Vipin Kumar. Incremental dual-memory lstm in land cover prediction. In *SIGKDD*. ACM, 2017.

[17] Xiaowei Jia, Ankush Khandelwal, Guruprasad Nayak, James Gerber, Kimberly Carlson, Paul West, and Vipin Kumar. Predict land covers with transition modeling and incremental learning. In *SDM*, 2017.

[18] Xiaowei Jia, Sheng Li, Ankush Khandelwal, Guruprasad Nayak, Anuj Karpatne, and Vipin Kumar. Spatial context-aware networks for mining temporal discriminative period in land cover detection. In *Proceedings of the 2019 SDM*, 2019.

[19] Jing Jiang and ChengXiang Zhai. Instance weighting for domain adaptation in nlp. In *ACL*, 2007.

[20] Guoliang Kang, Liang Zheng, Yan Yan, and Yi Yang. Deep adversarial attention alignment for unsupervised domain adaptation: the benefit of target expectation maximization. *arXiv preprint arXiv:1801.10068*, 2018.

[21] Huayu Li, Martin Renqiang Min, Yong Ge, and Asim Kadav. A context-aware attention network for interactive question answering. In *SIGKDD*. ACM, 2017.

[22] Yanghao Li et al. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 2018.

[23] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*, 2015.

[24] Lingkun Luo et al. Discriminative label consistent domain adaptation. *arXiv preprint:1802.08077*, 2018.

[25] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[26] Guruprasad Nayak, Varun Mithal, Xiaowei Jia, and Vipin Kumar. Classifying multivariate time series by learning sequence-level discriminative patterns. In *SDM*. SIAM, 2018.

[27] Sinno Jialin Pan et al. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 2011.

[28] Sanjay Purushotham, Wilka Carvalho, Tanachat Nilanon, and Yan Liu. Variational recurrent adversarial deep domain adaptation. 2016.

[29] Sanjay Purushotham, Wilka Carvalho, Tanachat Nilanon, and Yan Liu. Variational adversarial deep domain adaptation for health care time series analysis, 2017.

[30] Toshihiro Sakamoto et al. Detecting seasonal changes in crop community structure using day and night digital images. *PE&RS*, 2010.

[31] Devis Tuia, Claudio Persello, and Lorenzo Bruzzone. Domain adaptation for the classification of remote sensing data: An overview of recent advances. *GRSM*.

[32] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017.

[33] Jacob M Williams. Deep learning and transfer learning in the classification of eeg signals. 2017.

[34] Zhilin Yang, Ruslan Salakhutdinov, and William W Cohen. Transfer learning for sequence tagging with hierarchical recurrent networks. *arXiv preprint arXiv:1703.06345*, 2017.

[35] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017.