DOMA IN SEPARATION FOR EFFICIENT ADAPTIVE ACTIVE LEARNING

Giona Matasci\textsuperscript{1}, Devis Tuia\textsuperscript{2}, Mikhail Kanevski\textsuperscript{1}

\textsuperscript{1} Institute of Geomatics and Analysis of Risk, University of Lausanne, Lausanne, Switzerland.
\textsuperscript{2} Image Processing Laboratory (IPL), Universitat de València, València, Spain.

Email: giona.matasci@unil.ch
Motivation

- Collection of ground truth for remote sensing image classification is a demanding task.
- Label the least possible number of samples from the newly acquired images.
- Take advantage of the availability of already collected reference data on previously acquired images with similar properties.

⇒ Rapid and large scale mapping of the land cover based on minimal labeled datasets.
CURRENT SOLUTIONS

- **Domain Adaptation (DA)** to adapt an existing classifier already trained on a **source image** to classify the newly acquired **target image** [Bruzzone & Prieto, TGRS, 2001].
  - **Problem**: when available, the target domain **samples are passively obtained** at once.

- **Active Learning (AL)** to guide the **intelligent sampling** of ground truth data [Tuia *et al.*, JSTSP, 2011].
  - **Problem**: not designed to handle a **shift in class distributions**.
Combination of the DA and AL approaches:

- AL via **Breaking Ties** to identify the best pixels to label in the target image [Tuia et al., RSE, 2011].
- **TrAdaBoost** to differently re-weight the samples of the training set progressively extended by AL (following [Jun & Ghosh, IGARSS, 2008]).
PROPOSED SOLUTION

⇒ Combination of the DA and AL approaches:

- AL via Breaking Ties to identify the best pixels to label in the target image [Tuia et al., RSE, 2011].
- TrAdaBoost to differently re-weight the samples of the training set progressively extended by AL (following [Jun & Ghosh, IGARSS, 2008]).
Combination of the DA and AL approaches:

- **Domain Separation** [Rai et al., ALNLP, 2010] to pre-select target pixels worth the labeling.
- AL via **Breaking Ties** to identify the best pixels to label in the target image [Tuia et al., RSE, 2011].
- **TrAdaBoost** to differently re-weight the samples of the training set progressively extended by AL (following [Jun & Ghosh, IGARSS, 2008]).
**Domain separation steps**

- **Identify** the **informative regions of the target domain**.
- **Classify source VS. target domain** and predict **target domain probabilities** $p(\text{target}|x_j)$ for the unlabeled set of candidates $x_j$.
- Remove candidate examples $x_j$ with $p(\text{target}|x_j) < p_T$.
- **Avoid the re-sampling** of the input **space covered by source data**.
DOMAIN SEPARATION CONCEPT

Source domain

Source classifier

+1

-1
Domain separation concept

Source domain

Target domain

Source classifier

+1

-1

G. Matasci et al. IGARSS 2011, Vancouver, Canada
Domain separation concept

Source domain

Target domain

Source classifier

Domain separation

Source domain: +1, -1

Target domain: +1, -1
**Domain Separation Concept**

- **Source Domain**
- **Target Domain**
- Uninteresting target domain region

- Source classifier
- Domain separation

G. Matasci et al.  IGARSS 2011, Vancouver, Canada
DOMAIN SEPARATION CONCEPT

Source classifier

Source domain

Domain separation

Target domain

Interesting region for AL

Uninteresting target domain region
Main **AL loop** with a query of the candidate samples $x_j$ to label in the target domain by Breaking Ties:

$$\hat{x}^{BT} = \arg \min_{x_j \in U} \left( \max_{cl \in \Omega} p(y_j^* = cl|x_j) - \max_{cl \in \Omega \setminus cl^+} p(y_j^* = cl|x_j) \right)$$

- **Append the target samples to the initial source training set.**
- At each AL cycle a nested **TrAdaBoost loop** is run to update misclassified training samples’ weights:
  - decrease weights of $x_i \in \textbf{source}$ domain
  - increase weights of $x_i \in \textbf{target}$ domain
Adaptive AL Concept

Source Domain

Target Domain

Source classifier

uncertain region identified by AL

+1

-1
Domain separation and adaptive active learning experiments

Adaptive AL concept

Source classifier

TrAdaBoost to adapt the classifier

Uncertain region identified by AL

Source domain

Target domain

G. Matasci et al.  IGARSS 2011, Vancouver, Canada
Adaptive AL Concept

- **Source Domain**
- **Target Domain**

The diagram illustrates the process of domain separation and adaptive active learning.

1. **Source Classifier**: The classifier is applied to the source domain.
2. **TrAdaBoost**: This algorithm is used to adapt the classifier by identifying an uncertain region.
3. **Uncertain Region**: The region identified by AL is marked for further analysis.
4. **Adaptation**: The classifier is adapted to better fit the target domain.
Adaptive AL Concept

Source classifier

Source domain

TARGET DOMAIN

Adapted target classifier

Uncertain region identified by AL

TrAdaBoost to adapt the classifier

+1

-1

-1

+1
**DATASETS: QUICKBIRD IMAGES OF ZURICH**

- **Source domain**: image taken in **autumn**.
- **Target domain**: image taken in **summer**.

⇒ **Dataset shift** due to:
  - **different look angle**: different shadowing of the objects.
  - **different season**: shifted vegetation signatures.
  - **different materials** for building roofs and roads.

- Spatial resolution: 0.6 m (after pansharpening).

- **Histogram matching** to reduce the shift.

⇒ **15 features** (4 MS bands, 1 PAN band, 4 textural f., 6 morphological f.)

- **5 classes**: buildings, grass, vegetation, roads, shadows.
**Source Image (Autumn):**
False Color

**Target Image (Summer):**
False Color
SOURCE IMAGE (AUTUMN): GROUND TRUTH

TARGET IMAGE (SUMMER): GROUND TRUTH
INTRODUCTION

DOMAIN SEPARATION AND ADAPTIVE ACTIVE LEARNING

EXPERIMENTS

CONCLUSIONS

DATASETS

SETUP

RESULTS

SOURCE TRAINING SET

TARGET TEST SET

<table>
<thead>
<tr>
<th>buildings</th>
<th>grass</th>
<th>vegetation</th>
<th>roads</th>
<th>shadows</th>
</tr>
</thead>
</table>

standardized R band

standardized NIR band
**Experimental setup**

- **Base classifier:** SVM with Gaussian kernel (integrating instance weights for TrAdaBoost).

- **10 experiments with random initialization** of the training sets with 1000 pixels.

- AL with addition of 10 target samples per iteration.

- **Domain Separation ruling out samples with** $p(\text{target}|x_j)$ lower than $p_T = 0.8$.

- Target test set: 26'797 pixels from spatially separate ROIs.
**Source Domain vs. Target Domain**

- Source pixels: 
- Target pixels: 

**Posterior Target Prob. for the Candidates**

- Unlabeled candidate pixels:

---

G. Matasci et al. IGARSS 2011, Vancouver, Canada

14/18
- Faster convergence to the optimal accuracy using Domain Separation.
- Target model performance is even exceed using adaptive AL approaches.
Significant superiority at $\alpha = 0.05$ of the Domain Separation approach over the two baselines in the first key iterations.

Superiority lasting until convergence (with 1130 pixels in the training set) when compared to the standard BT method.
**SUMMARY**

- **Domain Separation** is an efficient yet intuitive technique to **pre-select interesting examples** to label in a domain adaptation setting.

- Properly **reduces the space to be searched by AL**.

- **TrAdaBoost** proved potential in **meaningfully adapting samples’ weights** way according to their origin (source or target domain).

⇒ The **user** is **smartly guided** in the collection of ground truth samples when dealing with a newly acquired target image.
Thank you for your attention!

Any questions?