Generalizability of Deep Domain Adaptation in Case of Sample Selection Bias

Author: Emiel Witting (e.a.witting@student.tudelft.nl) Faculty EEMCS at TU Delft Thesis committee: Joana de Pinho Gonçalves, Yasin Tepeli, Julián Urbano Merino

1. Background

There is often a domain shift between the train data (source) and the test data (target) in machine learning. This can lead to unexpected performance decreases after deploying a model [1].

Deep domain adaptation (DDA) methods mitigate this, using labelled train (source) data and unlabelled target data.

They remove shift and align the domains with neural network encodings, then follow with a normal learning task:

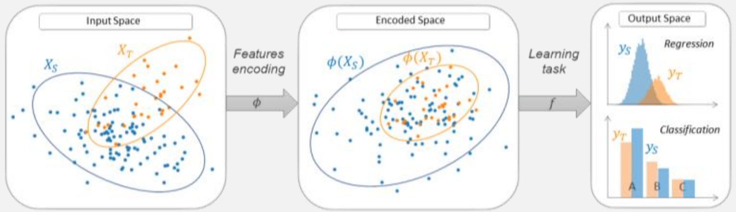


Fig. 1 Shift Mitigation by Encoding Features [2]

The shift between source and target can be caused by them both originating from a sample selection bias on a "global" domain. Causing cases such as these:

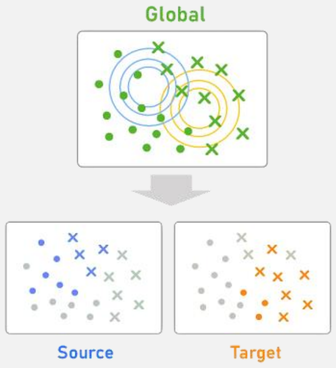


Fig. 2a Covariate Shift

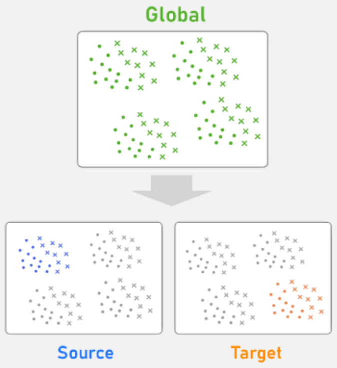


Fig. 2b Concept Shift

2. Research question

Adapting to the global domain has not been explored well.

We applied existing DDA methods to this new use case.

RQ: "How effective is DDA when adapting to the global domain, instead of the target domain, in case of sample selection bias?

3. Method

1. Randomly generate classification data sets

2. Induce bias
   2a. Covariate shift
   1. Partition: Global / Source *candidates* / Target *candidates*
   2. Sample source and target from candidates, with probabilities based on two normal distributions (fig. 2a)

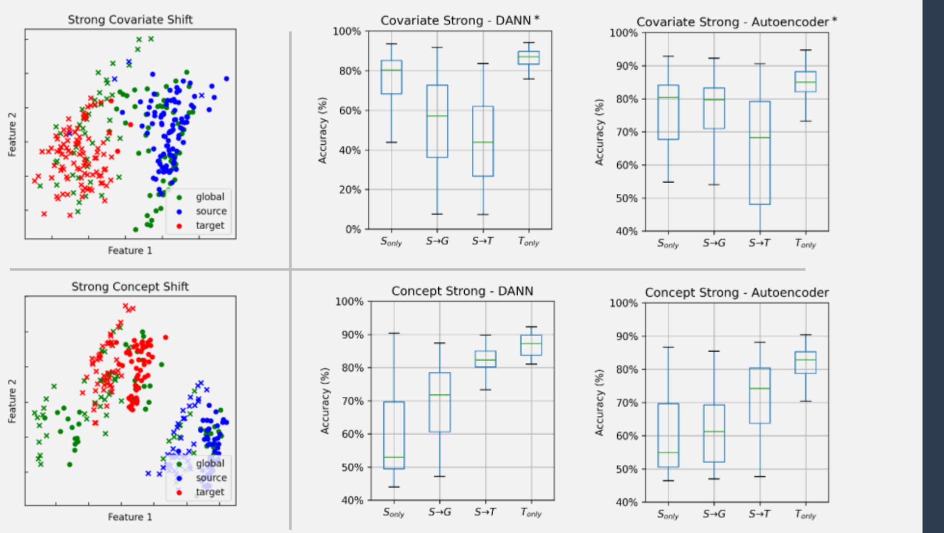
   2b Concept shift
   1. Partition into N pseudo-domains
   2. Translate each randomly through feature space
   3. Sample global data randomly from all pseudo-domains
   4. Assign specific ones to source and target (fig. 2b)

3. Implement two DDA methods
- Domain adversarial neural network (DANN) [3]
- Autoencoder with MMD loss [4]

4. Optimize, train and evaluate:
- 2x DDA methods
- 4x Bias types: weak/strong, covariate/concept
- 4x Train and adaptation combinations:
  - Source-only (supervised)
  - Source→Global adaptation (ours)
  - Source→Target adaptation
  - Target-only (supervised)

4. Results

Bias generation Target performance



(Weak variations omitted) *prevented parameter optimization from disabling adaptation

5. Conclusion

- DDA works with concept shift, not covariate shift.
- DANN worked better than the autoencoder.

With that in mind,

Global adaptation is...
significantly better than supervised learning but not as effective as target adaptation.

6. Future Work

- Repeat with real data sets, using larger models on i.e. image data.
- Investigate why the autoencoder approach did not work as well.
- Modify methods to use labels of pseudo-domains within global

7. References

[1] Antonio Torralba and Alexei A. Efros, Unbiased look at dataset bias, 2011
[2] https://adapt-python.github.io/adapt/contents.html
[3] Yaroslav Ganin et al., Domain-adversarial training of neural networks, 2015
[4] Xiang Li et al. Intelligent cross-machine fault diagnosis approach with deep auto-encoder and domain adaptation, 2019