Abstract
Large pre-trained models are usually fine-tuned on downstream task data, and tested on unseen data. When the train and test data come from different domains, the model is likely to struggle, as it is not adapted to the test domain. We propose a new approach for domain adaptation (DA), using neuron-level interventions. We modify the representation of each test example in specific neurons, resulting in a counterfactual example from the source domain, which the model is more familiar with. The modified example is then fed back into the model. While most other DA-methods are applied during training time, ours is applied during inference only, making it more efficient and applicable. Our experiments show that our method improves performance on unseen domains.

Method
Given:

- Model: \( M \) with a classifier: \( f \), fine-tuned on a source domain, \( D_s \) = \( \{ x_s, y_s \} \).
- Unlabeled data: \( D_t = \{ x_t \} \), from a target domain, only used for inference.

We make the representation of each test example in specific neurons, resulting in a counterfactual example from the source domain.

1. Process \( x_s \) and \( x_t \) through \( M \), producing representations \( R_s \), \( R_t \) \( \in \mathbb{R}^n \). Also compute \( s \) and \( t \), the neuron-sorted coefficients vector in the range \( \mathbb{R}^n \) such that \( s = \text{arg}\-\text{max} \sum \beta_i \). \( t \) only in the \( s \) domain. Then, calculate the element-wise difference between the mean vectors, \( r = \sum \beta_i \cdot (t_i - s_i) \).
2. and obtain a ranking by arg-sort \( r \), i.e., the first neuron in the ranking corresponds to the highest value in \( r \).

We showed that in some cases, IDANI can significantly help models to adapt to new domains.

Experiments

Datasets
- Binary sentiment analysis.
- Airline, Books, DVD, Electronics, Kitchen.
- Natural language inference (NLI: contradiction, entailment, neutral).
- Fiction, Government, Sports, Television, Travel.
- Aspect prediction — binary token classification.
- Device, Laptop, Restaurants, Service.

Setup — unsupervised domain adaptation (UDA)
- Train an algorithm on a single source domain.
- Test the algorithm on a different target domain.
- Low-resource scenario: 2000–3000 training examples from the source domain.

Model
- We experiment with different \( \beta \) (number of modified neurons) and \( \lambda \) (magnitude of the intervention) values.

Results

IDANI generally improves results with default hyperparameters.

- With oracle hyperparameters, IDANI improves performance in almost all experiments.
- IDANI performs better than Linear.

Qualitative Analysis
Importantly, IDANI is applied only during inference, unlike most other DA methods.

Acknowledgments
Research supported by the ISRAEL SCIENCE FOUNDATION (grant No. 448-20) and by an Ariel Foundation Early Career Faculty Fellowship. We thank the anonymous reviewers for their insightful comments and suggestions.