AIL 862

Lecture 24

What are roles that ViT can play in DA

What are roles that ViT can play in DA

At bare minimum, it can provide a more effective feature extractor backbone in methods that use domain adversarial training

What are roles that ViT can play in DA

At bare minimum, it can provide a more effective feature extractor backbone in methods that use domain adversarial training

Let’s look into this aspect from point of view of multi-target adaptation

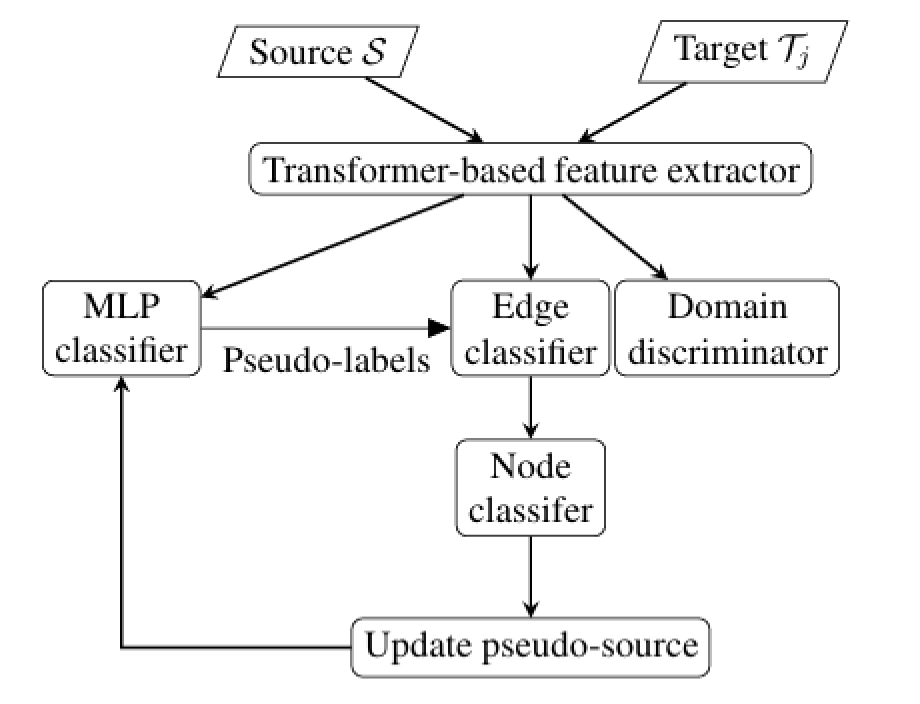
Multi-target adaptation

Single source, multiple targets

Assumption: We know which datapoint is coming from which target during the training process

Reiterative domain aware multi-target adaptation. In *German Conference on Pattern Recognition*

Adapting on single target



k^*	B_s	B_t	**Backbone**	**Acc.**
1	32	32	ResNet-50	70.5
10	32	32	ResNet-50	73.6
10	48	16	ResNet-50	73.7
10	48	16	Transformer	80.8

Table 7. Variation of 4 components of the proposed method (k^*, B_s, B_t, and backbone) on Office-Home dataset, source: Art, target: rest.

- Train on the source dataset

- Train on the source dataset
- Choose a target (from the collection of targets) to adapt on

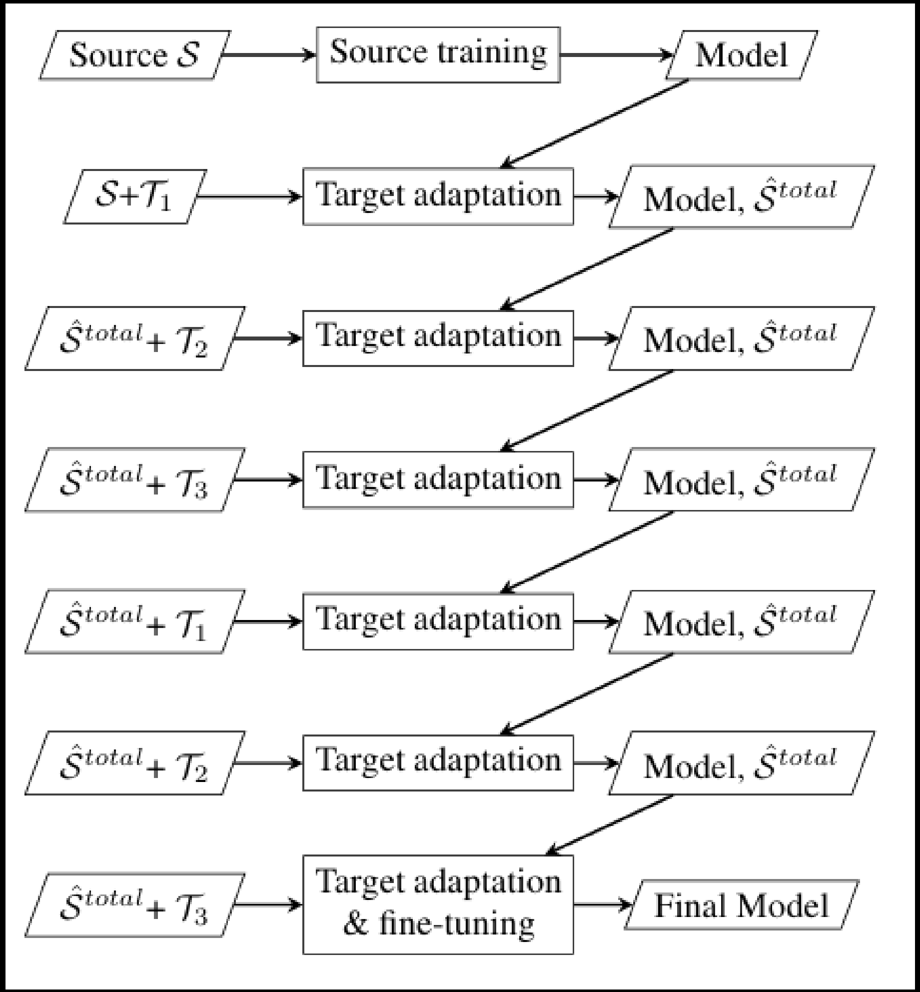
- Train on the source dataset
- Choose a target (from the collection of targets) to adapt on

❑ Feedback from MLP to GNN (during training)

- Train on the source dataset
- Choose a target (from the collection of targets) to adapt on
- [ ] Feedback from MLP to GNN (during training)
- [ ] Feedback from GNN to MLP (after processing each target)

- Train on the source dataset
- Choose a target domain (from the collection of targets) to adapt on
- ❑ Feedback from MLP to GNN (during training)
- ❑ Feedback from GNN to MLP (after processing each target)
- Choose another target domain from the remaining domains and repeat

Reiterative adaptation



What are roles that ViT can play in DA

At bare minimum, it can provide a more effective feature extractor backbone in methods that use domain adversarial training

However, more effectively, we can use the generalizability acquired by foundations models

Foundation Models and DA

- We will see the case of CLIP

Foundation Models and DA

- We will see the case of CLIP

- Let's start from the ideas that we used for CNN-based domain adaptation

❑ Batch normalization

Foundation Models and DA

- We will see the case of CLIP

- Let's start from the ideas that we used for CNN-based domain adaptation

❑ Batch normalization – can we replace by feature stylization

CLIP – Feature stylization

- Maybe even target domain images are not needed?

CLIP – Feature stylization

- Target domain textual descriptions – passed through text encoder – provides us target domain features – feature statistics

CLIP – Feature statistics alignment

- Target domain textual descriptions – passed through text encoder – provides us target domain features – feature statistics

- Similarly, source domain feature statistics from source domain images and the image encoder

CLIP – Feature statistics alignment

- Target domain textual descriptions – passed through text encoder – provides us target domain features – feature statistics
- Similarly, source domain feature statistics from source domain images and the image encoder
- Feature alignment

Foundation Models and DA

- We will see the case of CLIP

- Let’s start from the ideas that we used for CNN-based domain adaptation

❑ Batch normalization – we have just seen

❑ Adversarial training – we already saw with multi-target adaptation case

Foundation Models and DA

- We will see the case of CLIP

- Let’s start from the ideas that we used for CNN-based domain adaptation

❑ Batch normalization – we have just seen

❑ Adversarial training – we already saw with multi-target adaptation case

❑ Pseudo-label some data from target domain and then finetune using those pseudo-labeled data

Pseudo-label and fine-tune

- Risk of catastrophic forgetting – how to deal with it?

Pseudo-label and fine-tune

- Risk of catastrophic forgetting – how to deal with it?

❑ If you have access to source domain data during target adaptation – simply keep checking performance on the source domain data

Pseudo-label and fine-tune

- Risk of catastrophic forgetting – how to deal with it?

❑ If you have access to source domain data during target adaptation – simply keep checking performance on the source domain data

❑ Only using target domain - Decrease the learning rate according to the difference between the original CLIP and fine-tuned CLIP representations. Large differences indicate that CLIP forgets the pre-trained knowledge (resulting in a new representation).

PADCLIP: Pseudo-labeling with Adaptive Debiasing in CLIP for Unsupervised Domain Adaptation

A step back – how to pseudo label

- In addition to what we already know, another idea – (strong) augment and (weak) augment same image and if the prediction matches, then added to pseudo label.

Don't confuse this with the idea of strong/weak augment in PADCLIP: Pseudo-labeling with Adaptive Debiasing in CLIP for Unsupervised Domain Adaptation

CLIP for multi-target adaptation

- With the idea of pseudo label

CLIP for multi-target adaptation

- With the idea of pseudo label

- Text input in format “a [DOMAIN] photo of a [CLASS]”

PADCLIP: Pseudo-labeling with Adaptive Debiasing in CLIP for Unsupervised Domain Adaptation

Weights - vectors

We can edit models via task arithmetic

Editing models with task arithmetic, 2023

Obtaining task vector

Forgetting via Subtraction

Method	ViT-B/32	
	Target (↓)	Control (↑)
Pre-trained	48.3	63.4
Fine-tuned	90.2	48.2
Gradient ascent	2.73	0.25
Random vector	45.7	61.5
Negative task vector	24.0	60.9

Learning via Addition

Learning via Addition

- Better multi-task model

Learning via Addition

- Better multi-task model
- Single resulting model can be competitive with using multiple specialized models

Domain Generalization

Objective

Use the information from:

Pretrained ($\theta_{pre, D1}$) and finetuned ($\theta_{ft, D1}$) models on dataset in domain D1 (Amazon/DSLR)

To obtain a model that:

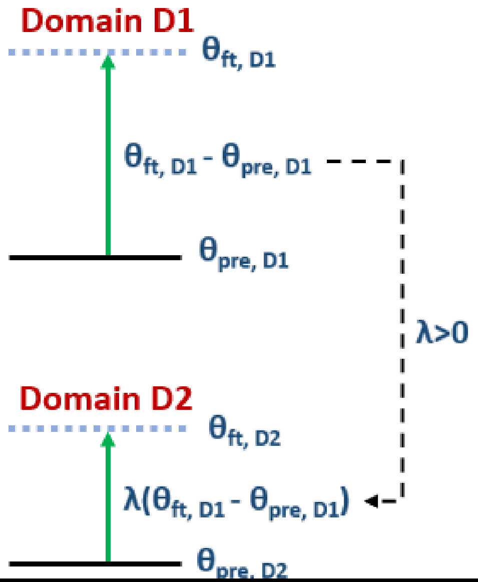
Performs well on dataset in domain D2 (DSLR/Amazon)

For the task of:

Image Classification

Without:

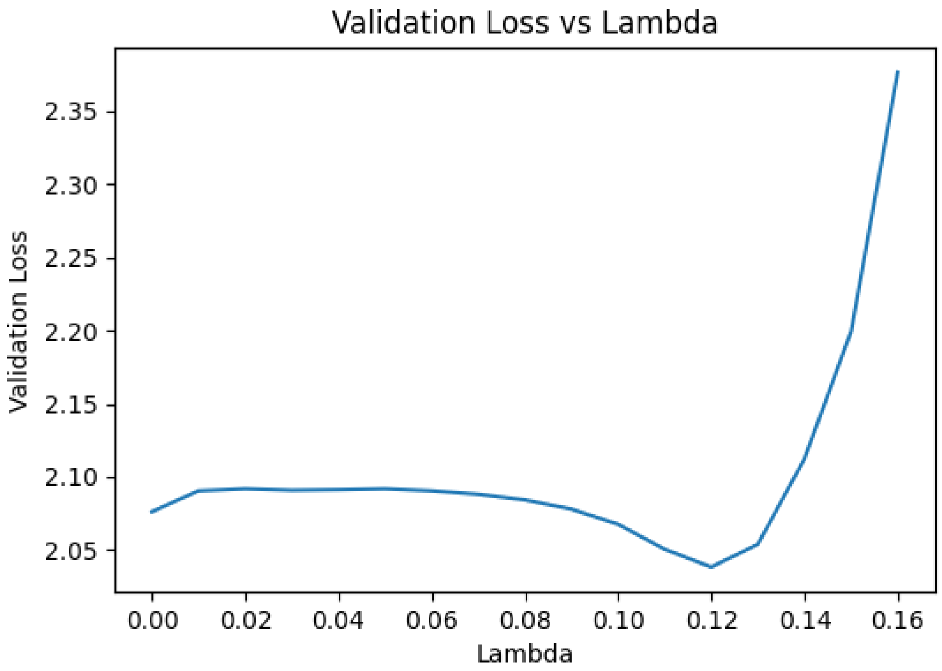
Explicit finetuning on dataset in domain D2 (DSLR/Amazon)

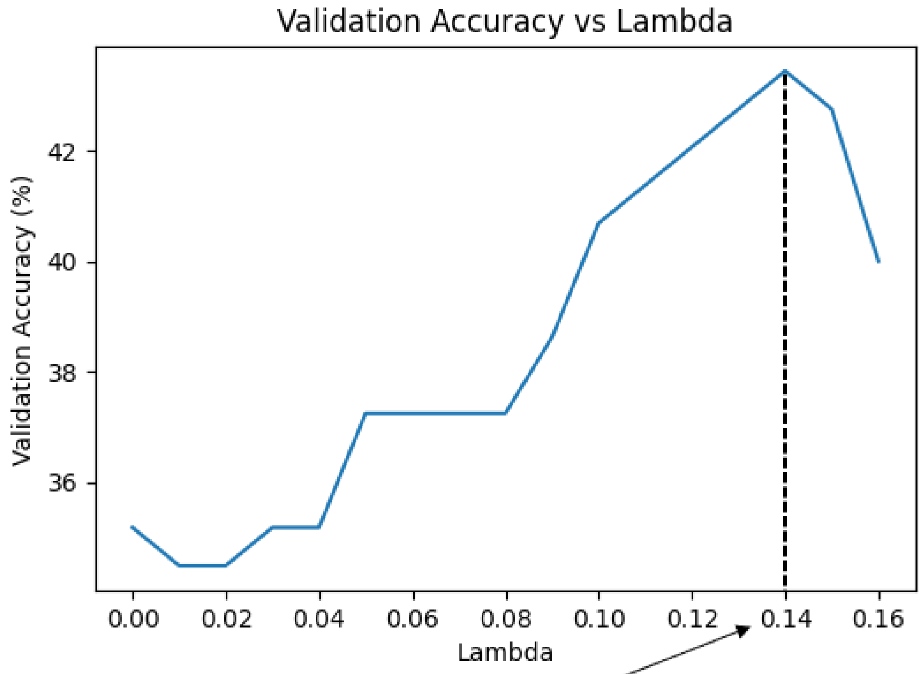


Thought Process

$$\theta_{ft, D2} = \theta_{pre, D2} + \lambda(\theta_{ft, D1} - \theta_{pre, D1}) \quad \text{where } \lambda>0$$

Finding Optimal λ using Validation Dataset





Optimal λ = 0.14 @ max. Validation Accuracy

Federated Learning

Task Arithmetic Through The Lens Of One-Shot Federated Learning, 2024