Unsupervised Domain Adaptation via Regularized Conditional Alignment

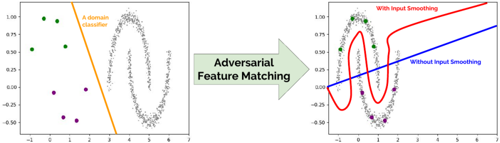
Cicek et al., ICCV 2019

Outline

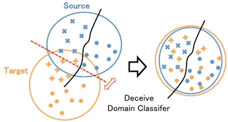
- Overview of domain adversarial adaptation method (DA) and one of its issue
  - Domain adversarial adaptation tries to align/map source and target examples into a common representation so that a class predictor can classify source examples can also perform on target examples
  - However, it adapts *only* the domains *not* the classes
    e.g., a natural image of a cat can be mapped to a synthesis image of a dog
- This paper proposes a new method: joint domain and class adversarial adaptation (JDCA)
  - Instead of imposing a binary domain adversarial loss, it imposes a K-way binary adversarial loss (2K classification, the first K are the known source classes, and the second K are the unknown target classes)
  - The encoder will try to fool the predictor by extracting invariant features from a synthesis image (source) and from a natural image (target) of examples of a specific class

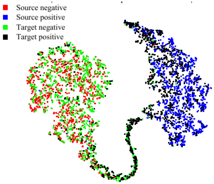
Issues of Class Misalignment

- Simple scenario, when feature distributions are aligned (i.e. a domain classifier cannot distinguish which domain the extracted features belong to), when target examples are mapped into source examples with correct classes, the class predictor performs well in both domains



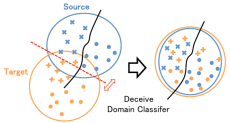
- Realistic scenario, when target examples are mapped into source examples, some target examples from one class are mapped into source examples from a different class

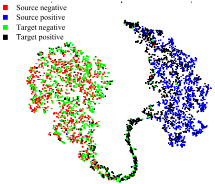




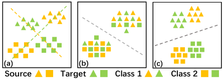
Issues of Class Misalignment: Ideal Solution

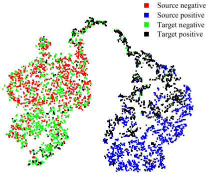
- Issue: when target examples are mapped into source examples, some target examples from one class are mapped into source examples from a different class





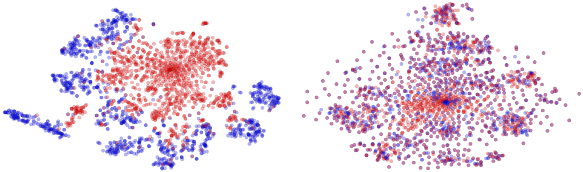
- Ideally, we want a model that can simultaneously align examples from two domain and align examples' classes



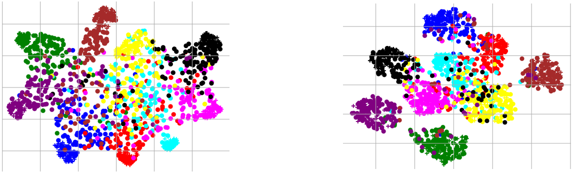


Intuitive Visualization between DA and JDCA

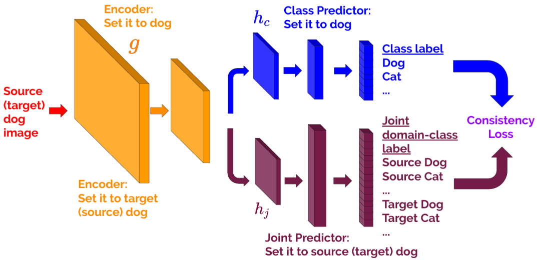
- DA



- JDCA



Joint Domain and Class Adaptation Model



- Three components
  - Feature encoder g
  - Class predictor h_c is trained only source examples, but is used for inference for both source and target examples
  - Joint predictor h_j is trained to jointly align domains and classes of examples
- Note: the model is trained in adversarially using GAN style, not gradient reversal layer, so the model does not have the domain predictor

Joint Domain and Class Adaptation Model

- Class predictor h_c is trained only source examples, but is used for inference for both source and target examples

$$L_{sc}(\theta_g, \theta_{h_c}) = \mathrm{CE}(h_c(g(x^s)), y^s)$$

- The encoder g is trained to jointly align domains and classes of examples via adversarial mechanism with the help of the joint predictor
  - The joint predictor will try to distinguish true source and target labels

$$L_{jsc}(\theta_{h_j}) = \mathrm{CE}(h_j(g(x^s)),\ [y^s, \mathbf{0}])$$
$$L_{jtc}(\theta_g, \theta_{h_j}) = \mathrm{CE}(h_j(g(x^t)),\ [\mathbf{0}, \hat{y}^t]),$$
$$\text{where } \hat{y}^t = \arg\max h_c(g(x^t))$$

  - The encoder will try to fool the joint predictor by generating adversarial source and target labels

$$L_{jsa}(\theta_g) = \mathrm{CE}(h_j(g(x^s)),\ [\mathbf{0}, y^s])$$
$$L_{jta}(\theta_g) = \mathrm{CE}(h_j(g(x^t)),\ [\hat{y}^t, \mathbf{0}]),$$
$$\text{where } \hat{y}^t = \arg\max h_c(g(x^t))$$

Post Adaptation: Semi-supervised Learning

- Once source and target examples are aligned, we train the model solely on target examples (to focus on target representations) using semi-supervised learning model
  - Model can automatically predict pseudo-labels and train on its own
- This paper uses a entropy minimization and virtual adversarial training models for semi-supervised learning

Thank you !

Image Classification Benchmark

Source dataset Target dataset	MNIST SVHN	SVHN MNIST	CIFAR STL	STL CIFAR	SYN-DIGITS SVHN	MNIST MNIST-M
[10] DANN*	60.6	68.3	78.1	62.7	90.1	94.6
[11] DRCN	40.05	82.0	66.37	58.86	NR	NR
[38] kNN-Ad	40.3	78.8	NR	NR	NR	86.7
[36] ATT	52.8	86.2	NR	NR	92.9	94.2
[9] Π-model**	33.87	93.33	77.53	71.65	96.01	NR
[40] VADA	47.5	97.9	80.0	73.5	94.8	97.7
[40] DIRT-T	54.5	99.4	NR	75.3	96.1	98.9
[40] VADA + IN	73.3	94.5	78.3	71.4	94.9	95.7
[40] DIRT-T +IN	76.5	99.4	NR	73.3	96.2	98.7
[18] Co-DA	81.7	99.0	81.4	76.4	96.4	99.0
[18] Co-DA + DIRT-T	88.0	**99.4**	NR	77.6	**96.4**	99.1
Ours	**89.19**	99.33	**81.65**	**77.76**	96.22	**99.47**
Source-only (baseline)	44.21	70.58	79.41	65.44	85.83	70.28
Target-only	94.82	99.28	77.02	92.04	96.56	99.87