

Domain Adaptation and Zero-Shot Learning

Lluís Castrejón

CSC2523 Tutorial

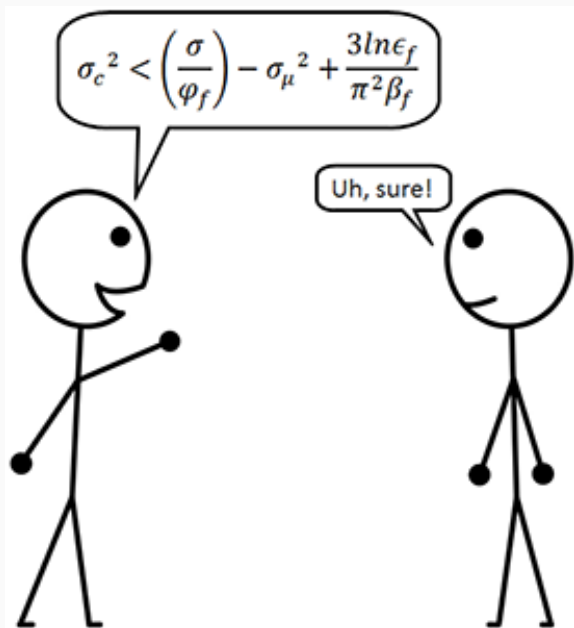
What is this?

Game: Caption the following images using one short sentence.

What is this?



What is this?



What is this?



What is this?

Let's now let a CNN play this game

What is this?

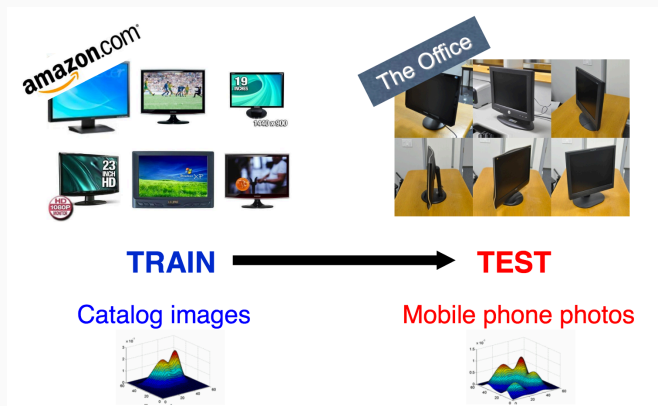
Let's now let a CNN play this game

Current computer vision models are affected by domain changes

Domain Adaptation

Use the same model with different data distributions in training
and test

$$P(X) \neq P(X'); P(Y|X) \approx P(Y'|X')$$



Credit: Kristen Grauman

Domain adaptation

Domain shift:

- ▶ Dataset shift in machine learning [Quionero-Candela 2009]
- ▶ Adapting visual category models to new domains [Saenko 2010]

Dataset bias:

- ▶ Unbiased look at dataset bias [Torralba 2011]
- ▶ Undoing the damage of dataset bias [Khosla 2012]

One-Shot Learning

Learn a classifier using only one (or fewer than normal) examples.



Credit: *Russ Salakhutdinov*

- ▶ A Bayesian approach to unsupervised one-shot learning of object categories [Fei-Fei 2003]
- ▶ Object classification from a single example utilizing class relevance pseudo-metrics [Fink 2004]

One-Shot Learning

Training

- ▶ Many labeled images for seen categories

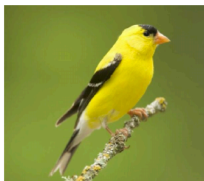
Test

- ▶ One (or a few) training images for new categories
- ▶ Infer new classifiers
- ▶ Test on a testing set (often combining images from seen and unseen categories)

Zero-shot learning

American Goldfinch

Intuitive!



Attribute	Has?
Beak longer than head	✗
Solid yellow belly	✓
Black and white wings	✓
⋮	⋮

Credit: *Stanislaw Antol*

- ▶ Zero-shot learning with semantic output codes [Palatucci 2009]
- ▶ Learning to detect unseen object classes by between-class attribute transfer [Lampert 2009]

Zero-Shot Learning

Training

- ▶ Images for seen classes
- ▶ Additional knowledge for seen classes (attributes, descriptions, ...)
- ▶ Train mapping knowledge to classes

Test

- ▶ Additional knowledge for unseen classes
- ▶ Infer new classifiers
- ▶ Test on a testing set (often combining images from seen and unseen categories)

Word of caution



All these terms are related one to another and many tasks involve a combination of them, often leading to the terms being mixed up in the literature.

Paper #1: Domain Adaptation - Tzeng et al.

Simultaneous Deep Transfer Across Domains and Tasks

Paper #1: Domain Adaptation - Tzeng et al.

Simultaneous Deep Transfer Across Domains and Tasks

Goal: Adapt classifiers to work across domains.



digital SLR camera



low-cost camera, flash



amazon.com



consumer images

Credit: *Kate Saenko*

Paper #1: Domain Adaptation - Tzeng et al.

Wait! Doesn't fine-tuning take care of that?

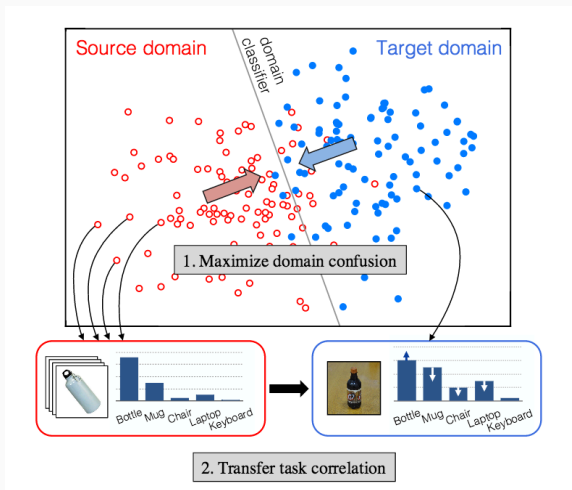
Paper #1: Domain Adaptation - Tzeng et al.

Wait! Doesn't fine-tuning take care of that?

Yes, but with two caveats:

- ▶ A considerable amount of LABELED data is still required
- ▶ Alignment across domains is lost

Paper #1: Domain Adaptation - Tzeng et al.



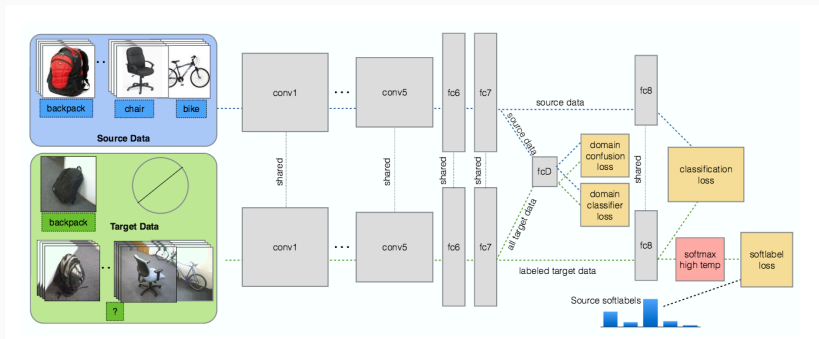
Credit: Tzeng et al.

Paper #1: Domain Adaptation - Tzeng et al.

Assumptions:

- ▶ We have a (small) amount of labeled data for (a subset of) the categories
- ▶ Source and target label spaces are the same

Paper #1: Domain Adaptation - Tzeng et al.



Credit: Tzeng et al.

Paper #1: Domain Adaptation - Tzeng et al.

Domain confusion

Classify a mug



Domain Confusion

Goal: Learn a domain-invariant representation

- ▶ Add a fully-connected layer f_{cD} and train a classifier to discriminate domains: Domain Classifier loss \mathcal{L}_D

$$\mathcal{L}_D(x_S, x_T, \theta_{\text{repr}}; \theta_D) = - \sum_d \mathbb{1}[y_D = d] \log q_d \quad (3)$$

Domain Confusion

Goal: Learn a domain-invariant representation

- ▶ Add another loss that quantifies the domain invariance of the representation: Confusion loss \mathcal{L}_{conf}

$$\mathcal{L}_{conf}(x_S, x_T, \theta_D; \theta_{repr}) = - \sum_d \frac{1}{D} \log q_d. \quad (4)$$

Paper #1: Domain Adaptation - Tzeng et al.

Domain Confusion

Goal: Learn a domain-invariant representation

- ▶ Optimize them alternatively in iterative updates (this is hard because these objectives are in contradiction, similar to adversarial networks!)

$$\min_{\theta_D} \mathcal{L}_D(x_S, x_T, \theta_{\text{repr}}; \theta_D) \quad (5)$$

$$\min_{\theta_{\text{repr}}} \mathcal{L}_{\text{conf}}(x_S, x_T, \theta_D; \theta_{\text{repr}}). \quad (6)$$

Attention: This does not ensure that features represent the same *concepts* across domains.

Paper #1: Domain Adaptation - Tzeng et al.

Alignment of source and target classes

Goal: Force the representation to be aligned between source and target domains

Simple implementation: Use the same category classifier for both domains and use subset of labels available for target domain

High-level idea: My features need to tell me that this represents a mug regardless of the domain in order to obtain good classification accuracy.

Paper #1: Domain Adaptation - Tzeng et al.

Alignment of source and target classes

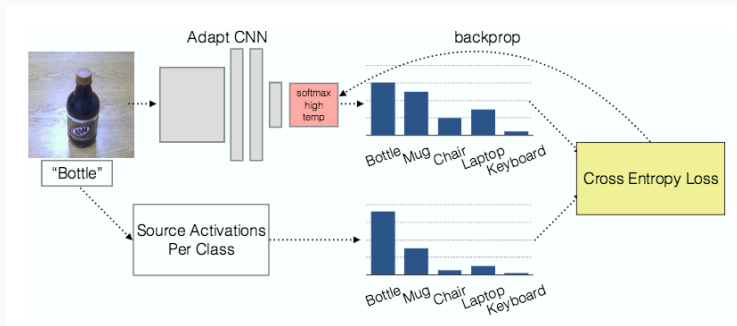
Goal: Force the representation to be aligned between source and target domains

Paper alternative: Use a soft-label loss \mathcal{L}_{soft} in which the probabilities for each label are tried to be replicated. Soft-labels are computed as average of predictions in the source CNN

Paper #1: Domain Adaptation - Tzeng et al.

Alignment of source and target classes

Goal: Force the representation to be aligned between source and target domains



Credit: Tzeng et al.

This can be seen as having a prior on the class labels. But it might not be right for some domains!

Paper #1: Domain Adaptation - Tzeng et al.

	$A \rightarrow W$	$A \rightarrow D$	$W \rightarrow A$	$W \rightarrow D$	$D \rightarrow A$	$D \rightarrow W$	Average
DLID [7]	51.9	-	-	89.9	-	78.2	-
DeCAF ₆ S+T [9]	80.7 ± 2.3	-	-	-	-	94.8 ± 1.2	-
DaNN [13]	53.6 ± 0.2	-	-	83.5 ± 0.0	-	71.2 ± 0.0	-
Source CNN	56.5 ± 0.3	64.6 ± 0.4	42.7 ± 0.1	93.6 ± 0.2	47.6 ± 0.1	92.4 ± 0.3	66.22
Target CNN	80.5 ± 0.5	81.8 ± 1.0	59.9 ± 0.3	81.8 ± 1.0	59.9 ± 0.3	80.5 ± 0.5	74.05
Source+Target CNN	<u>82.5 ± 0.9</u>	<u>85.2 ± 1.1</u>	65.2 ± 0.7	96.3 ± 0.5	<u>65.8 ± 0.5</u>	93.9 ± 0.5	81.50
Ours: dom confusion only	82.8 ± 0.9	85.9 ± 1.1	64.9 ± 0.5	97.5 ± 0.2	66.2 ± 0.4	95.6 ± 0.4	82.13
Ours: soft labels only	<u>82.7 ± 0.7</u>	<u>84.9 ± 1.2</u>	65.2 ± 0.6	98.3 ± 0.3	<u>66.0 ± 0.5</u>	95.9 ± 0.6	82.17
Ours: dom confusion+soft labels	<u>82.7 ± 0.8</u>	86.1 ± 1.2	<u>65.0 ± 0.5</u>	97.6 ± 0.2	66.2 ± 0.3	<u>95.7 ± 0.5</u>	82.22

Table 1. Multi-class accuracy evaluation on the standard supervised adaptation setting with the *Office* dataset. We evaluate on all 31 categories using the standard experimental protocol from [28]. Here, we compare against three state-of-the-art domain adaptation methods as well as a CNN trained using only source data, only target data, or both source and target data together.

	$A \rightarrow W$	$A \rightarrow D$	$W \rightarrow A$	$W \rightarrow D$	$D \rightarrow A$	$D \rightarrow W$	Average
MMDT [18]	-	44.6 ± 0.3	-	58.3 ± 0.5	-	-	-
Source CNN	54.2 ± 0.6	63.2 ± 0.4	34.7 ± 0.1	94.5 ± 0.2	36.4 ± 0.1	89.3 ± 0.5	62.0
Ours: dom confusion only	55.2 ± 0.6	63.7 ± 0.9	41.1 ± 0.0	96.5 ± 0.1	41.2 ± 0.1	91.3 ± 0.4	64.8
Ours: soft labels only	56.8 ± 0.4	65.2 ± 0.9	38.8 ± 0.4	96.5 ± 0.2	41.7 ± 0.3	89.6 ± 0.1	64.8
Ours: dom confusion+soft labels	59.3 ± 0.6	68.0 ± 0.5	40.5 ± 0.2	97.5 ± 0.1	43.1 ± 0.2	90.0 ± 0.2	66.4

Table 2. Multi-class accuracy evaluation on the standard semi-supervised adaptation setting with the *Office* dataset. We evaluate on 16 held-out categories for which we have no access to target labeled data. We show results on these unsupervised categories for the source only model, our model trained using only soft labels for the 15 auxiliary categories, and finally using domain confusion together with soft labels on the 15 auxiliary categories.

Paper #2: Zero-shot Learning - Ba et al.

Predicting Deep Zero-Shot Convolutional Neural Networks using Textual Descriptions

Paper #2: Zero-shot Learning - Ba et al.

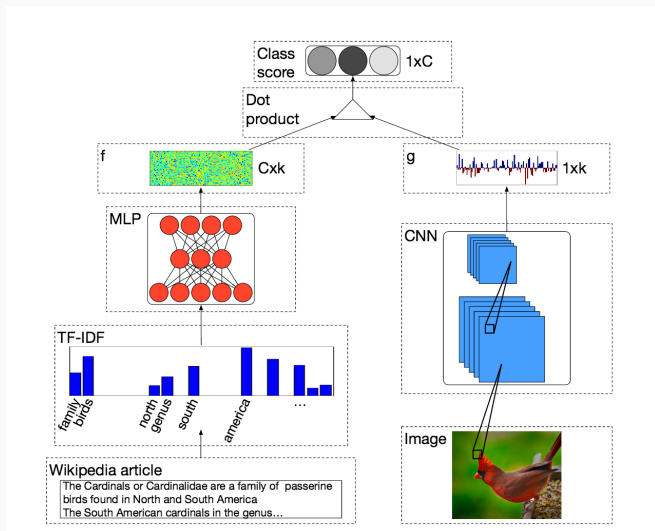
Predicting Deep Zero-Shot Convolutional Neural Networks using Textual Descriptions

Goal: Learn visual classifiers using only textual descriptions.

Globe thistle is one of the most elegantly colored plants around. It has fantastical large blue balls of steel-blue flowers



Paper #2: Zero-shot Learning - Ba et al.



Credit: *Ba et al.*

Paper #2: Zero-shot Learning - Ba et al.

Training:

- ▶ Images and descriptions for seen classes
- ▶ Learn classifiers for classes
- ▶ Learn a mapping from text to classifier weights

Test:

- ▶ Only descriptions for unseen classes
- ▶ Infer classifier weights (fully connected, convolutional or both)
- ▶ Evaluate on unseen images

Paper #2: Zero-shot Learning - Ba et al.

The devil is in the details!



Credit: *Mark Anderson*

Paper #2: Zero-shot Learning - Ba et al.

Implementation details:

- ▶ Dimensionality reduction

We need to reduce the dimensionality of the features since we only have < 200 descriptions and a classifier on fc7 features would have 4096 dimensions!

Fortunately, projections of CNN features still are very informative and we can learn them end-to-end using a MLP.

Paper #2: Zero-shot Learning - Ba et al.

Implementation details:

- ▶ Adam optimizer

In many experiments it has been shown that for architectures that would require different learning rates, Adam learns better and faster!

Adam [13] is used to optimize our proposed models with minibatches of 200 images. We found that SGD does not work well for our proposed models. This is potentially due to the difference in magnitude between the sparse gradient of the text features and the dense gradients in the convolutional layers. This problem is avoided by using adaptive step sizes.

Credit: *Ba et al.*

Paper #2: Zero-shot Learning - Ba et al.

Implementation details:

- ▶ Convolutional classifier

We can further reduce the dimensionality of the classifier features by learning a convolutional classifier.

$$\hat{y}'_c = o\left(\sum_{i=1}^{K'} w'_{c,i} \check{*} a'_i\right), \quad (4)$$

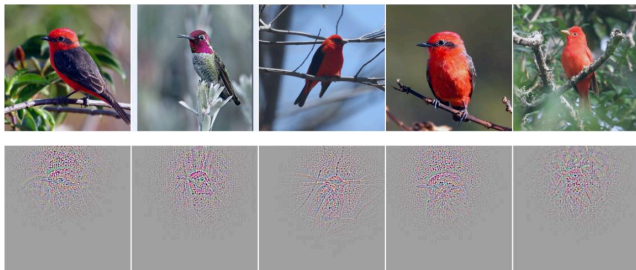
Credit: *Ba et al.*

Paper #2: Zero-shot Learning - Ba et al.

Implementation details:

- ▶ Convolutional classifier

It also allows us to see which part of the image is relevant to classify a species!



Credit: *Ba et al.*

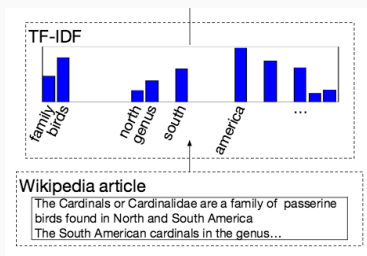
Paper #2: Zero-shot Learning - Ba et al.

Implementation details:

- ▶ TF-IDF and no predefined attributes

Can we improve the model by using distributed language representations?

TF-IDF allows us to easily find the most important features



Credit: *Ba et al.*

Paper #2: Zero-shot Learning - Ba et al.

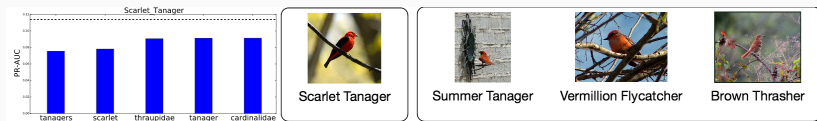
Evaluation of zero-shot learning is not straightforward

Accuracy and AUC measures are predominant measures, but
dataset splits are not standard

Paper #2: Zero-shot Learning - Ba et al.

Evaluation of zero-shot learning is not straightforward









Accuracy and AUC measures are predominant measures, but dataset splits are not standard



Credit: *Ba et al.*









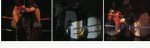
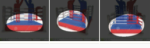


Learning Aligned Cross-Modal Representations from Weakly Aligned Data

Goal: Extend domain adaptation methods to more extreme and abstract domain/modality shifts

	Real	Clip art	Sketches	Spatial text	Descriptions
Bedroom					<p>There is a bed with a striped bedspread. Beside this is a nightstand with a drawer. There is also a tall dresser and a chair with a blue cushion. On the dresser is a jewelry box and a clock.</p> <p>I am inside a room surrounded by my favorite things. This room is filled with pillows and a comfortable bed. There are stuffed animals everywhere. I have posters on the walls. My jewelry box is on the dresser.</p>
	Kindergarten classroom				

Current work

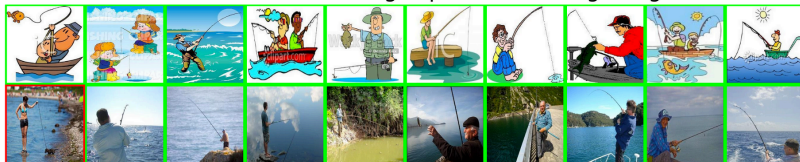
Method: Learn a multi-modal representation in which abstract concepts are aligned.

	Real	Clip art	Sketches	Spatial text	Descriptions
Unit 31 (Fountain)					we, water, fishes, you, drink, formed, greek, would, ball, have
Unit 50 (Arcade)					play, children, there, equipment, are, for, train, hole, games, path
Unit 81 (Ring)					ropes, recess, seats, dug, that, square, down, each, fight, it

Current work

Many applications: Cross-modal retrieval, zero-shot/transfer learning, etc.

AP=0.638332476203 Training clipart: 45 Testing images: 173



Current work

Demo

End

Thank you!

Questions?