Deep Learning for Computer Vision

Fall 2020

http://vllab.ee.ntu.edu.tw/dlcv.html (Public website)
https://cool.ntu.edu.tw/courses/3368 (NTU COOL; for grade, etc.)

Yu-Chiang Frank Wang 王鈺強, Professor
Dept. Electrical Engineering, National Taiwan University

2020/12/15
<table>
<thead>
<tr>
<th>Week</th>
<th>Date</th>
<th>Topic</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9/15</td>
<td>Course Logistics</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>9/22</td>
<td>Machine Learning 101</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>9/29</td>
<td>Intro to Neural Networks; Convolutional Neural Network (I)</td>
<td>HW #1 out</td>
</tr>
<tr>
<td>4</td>
<td>10/6</td>
<td>Convolutional Neural Network (II): Visualization &amp; Extensions of CNN</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>10/13</td>
<td>Tutorials on Python, Github, etc. (by TAs)</td>
<td>HW #1 due</td>
</tr>
<tr>
<td>6</td>
<td>10/20</td>
<td>Visualization of CNN (II); Object Detection &amp; Segmentation</td>
<td>HW #2 out</td>
</tr>
<tr>
<td>7</td>
<td>10/27</td>
<td>Image Segmentation; Generative Models</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>11/3</td>
<td>Generative Adversarial Network (GAN)</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>11/10</td>
<td>Transfer Learning for Visual Classification &amp; Synthesis; Representation Disentanglement</td>
<td>HW #2 due; HW #3 out</td>
</tr>
<tr>
<td>10</td>
<td>11/17</td>
<td>Guest Lectures (Dr. Trista Chen &amp; David Chou)</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>11/24</td>
<td>Representation Disentanglement; Recurrent Neural Networks &amp; Transformer (I)</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>12/1</td>
<td>Recurrent Neural Networks &amp; Transformer (II)</td>
<td>HW #3 due</td>
</tr>
<tr>
<td>13</td>
<td>12/8</td>
<td>Meta-Learning; Few-Shot and Zero-Shot Classification (I)</td>
<td>HW #4 out</td>
</tr>
<tr>
<td>14</td>
<td>12/15</td>
<td>Meta-Learning; Few-Shot and Zero-Shot Classification (II) From Domain Adaptation to Domain Generalization</td>
<td>Team-up for Final Projects (2~4 ppl per group, preferably 3+)</td>
</tr>
<tr>
<td>15</td>
<td>12/22</td>
<td>Beyond 2D vision (3D and Depth); Inpainting &amp; outpainting</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>12/29</td>
<td>Guest Lectures (Prof. Yen-Yu Lin &amp; Wei-Chen Chiu)</td>
<td>HW #4 due</td>
</tr>
<tr>
<td>17</td>
<td>1/5</td>
<td>Guest Lectures (NTU alumni at Google, Waymo, etc.)</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>1/18-22</td>
<td>Presentation for Final Projects</td>
<td>TBD</td>
</tr>
</tbody>
</table>
What to Cover Today...

• Meta-Learning for Few-Shot Learning
  • Few-Shot Classification
    • Metric Learning vs. Data Hallucination
    • Few-Shot Image Segmentation

• Zero-Shot Learning

• Domain Generalization
Meta Learning = Learning to Learn

• Let’s consider the following “2-way 1-shot” learning scheme:

**Meta-Training**

- **Task i**
  - **Train**
    - Support set: Cat (+) Dog (-)
  - **Test**
    - Query set: Cat (+) Dog (-)
    - Predict: + or -

- **Task i+1**
  - **Train**
    - Support set: Apple (+) Orange (-)
  - **Test**
    - Query set: Apple (+) Orange (-)
    - Predict: + or -

**Meta-Testing**

- **Train**
  - Support set: Bike (+) Car (-)
- **Test**
  - Query set: Bike Car
  - Predict: Bike as + or -?
Non-Parametric Meta-Learning Approach

• OK, we know how to learn to generalize. Can we just learn to compare?
• Parametric Meta-Learners → Non-Parametric Learners
• Siamese Network
  • Learn a network to determine whether a pair of images are of the same category.

Koch et al., Siamese Neural Networks for One-Shot Image Recognition, ICML WS 2015
Learn to Compare...with the Representative Ones!

- **Prototypical Networks**
  - Learn a model which properly describes data in terms of intra/inter-class info.
  - It learns a prototype for each class, with data similarity/separation guarantees.

Snell et al., Prototypical Networks for Few-Shot Learning, NIPS 2017
• **Prototypical Networks** (cont’d)
  
  • Learn a model which properly describes data in terms of intra/inter-class info.
  • It learns a prototype for each class, with data similarity/separation guarantees.
  • For DL version, the above embedding space is derived by a non-linear mapping $f_\phi$ and the representatives (or anchors) of each class is the **mean feature vector** $c_k$.

\[
c_k = \frac{1}{|S_k|} \sum_{(x_i, y_i) \in S_k} f_\phi(x_i), \text{ where } S_k \subset S \text{ is the subset of support set } S \text{ with class } k
\]

Snell et al., Prototypical Networks for Few-Shot Learning, NIPS 2017
What to Cover Today...

• Meta-Learning for Few-Shot Learning
  • Few-Shot Classification
    • Metric Learning vs. Data Hallucination
    • Few-Shot Image Segmentation

• Zero-Shot Learning

• Domain Generalization

Selected slide credits: C. Finn, S. Levine, H.-Y. Lee, & W.-L. H. Chao
Learn to Augment…Data Hallucination for FSL

• Data Hallucination
  • Many modes of intra-class variation (e.g., camera pose, translation, lighting changes, and even articulation) are shared across categories.
  • As humans, our knowledge of such intra-class variations allow us to visualize what a novel object might look like in other poses or surroundings.
  • We can thus hallucinate additional examples for novel classes by transferring variation modes from the base classes.
  • Typical data augmentation techniques only use a limited amount of a priori known invariances (e.g., translations, rotations, flips, addition of Gaussian noise, etc.).

Wang et al., "Low-Shot Learning from Imaginary Data," CVPR, 2018
Learn to Augment...Data Hallucination for FSL

- Jointly Trained Hallucinator
  - The hallucinated examples should be **useful** for classification tasks, rather than just being **diverse** or **realistic** (that may fail to improve FSL performances).
  - The authors proposed to train a **conditional-GAN-based** data hallucinator ($G(x, z)$) **jointly** with the meta-learning module ($h$) in an **end-to-end** manner.

**Diagram**:

- **G**: Conditional GAN
- **Sample**: Randomly sampled images
- **h**: Meta-learning module
- **S_{train}**: Training set
- **S_{aug}**: Augmented training set
- **S_{test}**: Test set

Forward pass and backward propagation as shown in the diagram.

Wang et al., Low-shot learning from imaginary data, CVPR 2018
A Closer Look at FSL (1/3)

- **Idea**
  - **Deeper backbones** significantly reduce the gap across existing FSL methods. (with decreased domain shifts between base and novel classes)
  - A slightly modified baseline method (**baseline++**) surprisingly achieves competitive performance.
  - Simple baselines (**baseline** and **baseline++**: trained on base and fine-tuned on novel) outperform representative FSL methods when the domain shift grows larger.

Chen et al., A Closer Look at Few-shot Classification, *ICLR*, 2019

Use **cosine distances** between the input feature and the weight vector for each class to reduce intra-class variations.
A Closer Look at FSL (2/3)

- Performance with deeper backbones
  - For CUB, gaps among different methods diminish as the backbone gets deeper.
  - For mini-ImageNet, some meta-learning methods are even beaten by baselines with a deeper backbone.

Chen et al., A Closer Look at Few-shot Classification, *ICLR*, 2019
A Closer Look at FSL (3/3)

- Performance with domain shifts (using ResNet-18)
  - Existing FSL methods fail to address large domain shifts (e.g., mini-ImageNet → CUB) and are inferior to the baseline methods.
  - This highlights the importance of learning to adapt to domain differences in FSL.

Chen et al., A Closer Look at Few-shot Classification, ICLR, 2019
What to Cover Today...

• Meta-Learning for Few-Shot Learning
  • Few-Shot Classification
    • Metric Learning vs. Data Hallucination
    • Few-Shot Image Segmentation

• Zero-Shot Learning

• Domain Generalization
Semantic Segmentation

- Goal
  - Assign a class label to each pixel in the input image
  - Don’t differentiate instances, only care about pixels
Few-Shot Segmentation

- A large number of image categories are with pixel-wise ground truth labels, while a small number of them are with limited.
- A **shared backbone** produces feature maps for both **support** and **query** images.
- **Prototypes** for each class is obtained by **masked pooling** from support feature maps.
- Query feature maps are then compared with the pooled prototypes **pixel-by-pixel**.
- Typically, **cosine similarity** is adopted for pixel-wise feature comparison.
• $S$ is an annotated image from a new semantic class
• Input $S$ to a function $g$ that outputs a set of parameters $\theta$
• $\theta$ is used to parameterize part of the segmentation model which produces a segmentation mask given $I_q$

OSLSM [BMVC 2017]

Prototype Learning [BMVC 2018]

- A prototype is learned for each foreground class and the background class.
- Prototypes are used to predict rough segmentation maps for each class.
- The final prediction is optimized using probability fusion.

AMP [ICCV 2019]

- Adaptive masked proxies (i.e., prototypes’) are extracted for each semantic class.
- Proxies update themselves in a continuous stream of data (e.g., video).
- Proxies from different resolution levels are used in multi-resolution imprinting.
AMP [ICCV 2019]

CANet [CVPR 2019]

- Dense comparison module (DCM) concatenates prototypes to each spatial location in query feature map
- Rough segmented maps are produced after comparing with mask-pooled feature prototypes
- The final result is optimized in an iterative manner
**CANet** [CVPR 2019]

- Standard FSL methods (e.g., shared backbone, masked pooling...) are used during training.
- A ‘relevance’ factor is added and taken into account during cosine similarity computation.
During inference, ensemble is utilized to select the best set of parameters.

Prototypes are used to predict the support masks reversely, which can be compared to the ground truth.
PANet [ICCV 2019]

- Extracted prototypes are first used to predict query masks, as standard FSL methods do.
- Predicted query masks are used to generate new prototypes and reversely predict support masks
- Similar concept to that of the ‘cycle consistency’ (support→query; query→support)

PANet [ICCV 2019]

Dataset & Evaluation Metric

- **Datasets**
  - **PASCAL VOC 2012** (main)
    - 20 classes
    - Split: (15 *base* + 5 *novel*)
  - **coco** (secondary)

- **Evaluation Metrics**
  - **Binary-mIoU** (difficult)
  - **FB-mIoU** (easy)
    - Foreground/Background IoU
## Performance Comparisons

<table>
<thead>
<tr>
<th>Method</th>
<th>Split-0</th>
<th>Split-1</th>
<th>Split-2</th>
<th>Split-3</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reduced-DFCN8s</td>
<td>39.2</td>
<td>48.0</td>
<td>39.3</td>
<td>34.2</td>
<td>40.2</td>
</tr>
<tr>
<td><strong>OSLSM</strong></td>
<td>33.6</td>
<td>55.3</td>
<td>40.9</td>
<td>33.5</td>
<td>40.8</td>
</tr>
<tr>
<td>co-FCN</td>
<td>ICLRW 2018</td>
<td>36.7</td>
<td>50.6</td>
<td>44.9</td>
<td>32.4</td>
</tr>
<tr>
<td>AMP</td>
<td>ICCV 2019</td>
<td>41.9</td>
<td>50.2</td>
<td>46.7</td>
<td>34.7</td>
</tr>
<tr>
<td><strong>SG-One</strong></td>
<td>40.2</td>
<td>58.4</td>
<td>48.4</td>
<td>38.4</td>
<td>46.4</td>
</tr>
<tr>
<td>PANet</td>
<td>ICCV 2019</td>
<td>42.3</td>
<td>58.0</td>
<td>51.1</td>
<td>41.2</td>
</tr>
<tr>
<td><strong>PRNet</strong></td>
<td>51.6</td>
<td>61.3</td>
<td>53.1</td>
<td>47.6</td>
<td>53.4</td>
</tr>
<tr>
<td>Co-att</td>
<td></td>
<td>49.5</td>
<td>65.5</td>
<td>50.0</td>
<td>49.2</td>
</tr>
<tr>
<td><strong>CANet</strong></td>
<td>52.5</td>
<td>65.9</td>
<td>51.3</td>
<td>51.9</td>
<td>55.4</td>
</tr>
<tr>
<td>PGNet</td>
<td>ICCV 2019</td>
<td>56.0</td>
<td>66.9</td>
<td>50.6</td>
<td>50.4</td>
</tr>
<tr>
<td>FWB</td>
<td>ICCV 2019</td>
<td>51.3</td>
<td>64.5</td>
<td>56.7</td>
<td>52.2</td>
</tr>
</tbody>
</table>
What to Cover Today...

• Meta-Learning for Few-Shot Learning
  • Few-Shot Classification
    • Metric Learning vs. Data Hallucination
    • Few-Shot Image Segmentation

• Zero-Shot Learning

• Domain Generalization

Selected slide credits: C. Finn, S. Levine, H.-Y. Lee, & W.-L. H. Chao
Recent Success

- ImageNet error rate goes down with strong supervision
Supervised Learning

Training data: pairs of inputs: images, outputs: labels

Training environment: seen objects

Supervised learning

Classifier

Slides credit: Wei-Lun (Harry) Chao
Real-World Scenarios...

Environment in the wild: Unseen objects

Slides credit: Wei-Lun (Harry) Chao
Challenges of Generalization

• Large number of classes

• Long-tail Distribution

Objects in SUN dataset

[Zhu et al. CVPR 2014]

Slides credit: Wei-Lun (Harry) Chao
Zero-Shot Learning

Training

In the wild

Zero-Shot learning

Classifier

horse, dog, cat 

(seen)

cat, horse, dog

Slides credit: Wei-Lun (Harry) Chao
Zero-Shot Learning (ZSL)

- **Two** type of classes:
  - **Seen**: with training examples
  - **Unseen**: with no training examples

- **Goal**: *Transfer* classifiers from **seen** classes to **unseen** ones

Slides credit: Wei-Lun (Harry) Chao
How human performs recognition?

**Chimpanzee**
- black
- mid sized
- wild

**Cat**
- stripes
- small sized
- domestic

Slides credit: Wei-Lun (Harry) Chao
How human performs zero-shot learning?

- **Semantic representation**

  - stripes, black body, white head, similar to zebra

- **Okapi**

- **Araripe Manakin**

  - white body, red head, tree, bird

Slides credit: Wei-Lun (Harry) Chao
Semantic Representation

- **Attributes** [Farhadi et al. 09, Lampert et al. 09, Parikh & Grauman 11]

  ![](image1.png)

- **Word vectors of class names** [Socher et al. 13, Frome et al. 13]

  ![](image2.png)

Semantic representation of class $c$: $m$-dimensional vector $\mathbf{a}_c$

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>has beak</td>
<td>0.9</td>
</tr>
<tr>
<td>has snout</td>
<td>0.1</td>
</tr>
<tr>
<td>furry</td>
<td>0.1</td>
</tr>
<tr>
<td>feather</td>
<td>0.9</td>
</tr>
<tr>
<td>has leg</td>
<td>0.9</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Slides credit: Wei-Lun (Harry) Chao
Evaluation

- **Zero-shot learning**: Recognition among unseen classes
- **Generalized ZSL**: Recognition among seen + unseen classes
Evaluation – Conventional Zero-Shot Recognition

Recognition among **unseen** classes

\[
f_{ZSL} : X \rightarrow Y^U
\]

\[
\arg \max_{c \in U} h(a_c)^T x
\]

\[
acc_U = \frac{1}{\|U\|} \sum_{c=1}^{\|U\|} \frac{\# \text{ correct predictions in } c}{\# \text{ samples in } c}
\]

A: zebra, leopard, wolf

Slides credit: Wei-Lun (Harry) Chao
Evaluation – Generalized Zero-Shot Recognition

Generalized: Recognition among seen + unseen classes

\[ \arg \max_{c \in S + U} h^T(a_c) x \]

\[ f_{GZSL} : X \rightarrow Y^U \cup Y^S \]

\[ H = \frac{2 \times acc_S \times acc_U}{acc_S + acc_U} \]

A: horse, dog, cat, zebra, leopard, wolf

Slides credit: Wei-Lun (Harry) Chao
What to Cover Today...

- Meta-Learning for Few-Shot Learning
  - Few-Shot Classification
    - Metric Learning vs. Data Hallucination
    - Few-Shot Image Segmentation
- Zero-Shot Learning
  - Embedding-based Approach
- Domain Generalization

Selected slide credits: C. Finn, S. Levine, H.-Y. Lee, & W.-L. H. Chao
Embedding-based Method

- Goal: align **visual** and **semantic** representations to a shared latent space for effectively performing nearest neighbor comparisons.
Cross and Distribution Aligned VAE (CADA-VAE)

- Intergrade VAE, distribution alignment, and cross-alignment loss
  - VAE: embedding both modality data into shared latent space
  - Distribution alignment: aligning global distribution of different modalities
  - Cross-alignment loss: ensuring the latent features to be modality-invariant

\[
\mathcal{L}_{DA} \min \{ ||\mu_1 - \mu_2||_2^2 + ||\Sigma_1^{1/2} - \Sigma_2^{1/2}||_{Frob}^2 \}
\]

\[
\mathcal{L}_{CA} \left|\frac{x - x'(z_1)}{x - x'(z_2)}\right|
\]

\[
\mathcal{L}_{CA} \left|\frac{c - c'(z_1)}{c - c'(z_2)}\right|
\]
What to Cover Today...

• Meta-Learning for Few-Shot Learning
  • Few-Shot Classification
    • Metric Learning vs. Data Hallucination
    • Few-Shot Image Segmentation
  • Zero-Shot Learning
    • Embedding-based Approach
    • Hallucination-based Approach
• Domain Generalization
Feature Generating Networks for ZSL

• Apply GAN that synthesizes visual features for seen and unseen classes from their semantic representations

• WGAN objective + classification loss

\[
\min_G \max_D \mathcal{L}_{GAN} = E[\log D(x, c(y))] + E[\log (1 - D(\tilde{x}, c(y)))],
\]

\[
\mathcal{L}_{WGAN} = -E_{\tilde{x} \sim p_\tilde{x}} [\log P(y|\tilde{x}; \theta)]
\]

\[
\mathcal{L}_{CLS} = -E_{\tilde{x} \sim p_\tilde{x}} [\log P(y|\tilde{x}; \theta)]
\]
What to Cover Today...

• Meta-Learning for Few-Shot Learning
  • Few-Shot Classification
    • Metric Learning vs. Data Hallucination
    • Few-Shot Image Segmentation

• Zero-Shot Learning
  • Embedding-based Approach
  • Hallucination-based Approach
  • Meta-Learning based Approach

• Domain Generalization
Meta-Learning for Zero-Shot Learning

- Create zero-shot episode to simulate the fake ZSL tasks during training
- The label space of *meta-training set* and *meta-testing set* are disjoint
- The label space of support set and refining set are disjoint (different from FSL)

**Episode-based Prototype Generating Network (E-PGN)**

- Meta-learning for zero-shot learning
- The training stage learns a base model to align the semantic information across different modalities.
- The refining stage updates the model parameters by minimizing the loss between the predicted results and the ground-truth labels.

Episode-based Prototype Generating Network (E-PGN)

• Design a cycle-consistency feature generation network to achieve semantic-visual interaction

• WGAN objective & Multi-modal Cross-Entropy loss

\[ L_{WGAN} = \mathbb{E}[D(x, a)] - \mathbb{E}[D(\tilde{x}, \tilde{a})] - \lambda \mathbb{E}[\|\nabla_\tilde{x} D(\tilde{x}, \tilde{a})\|_2 - 1)^2] \]

\[ L_{MCE} = - \sum_x \log p_i^y(x) - \sum_x \log p_i^S(x) \]

\[ p_i^y(x) = \frac{\exp(x^T G(a_j))}{\sum_j \exp(x^T G(a_j))} \quad p_i^S(x) = \frac{\exp(F(x)^T a_j)}{\sum_j \exp(F(x)^T a_j)} \]

What to Cover Today...

• Meta-Learning for Few-Shot Learning
  • Few-Shot Classification
    • Metric Learning vs. Data Hallucination
    • Few-Shot Image Segmentation

• Zero-Shot Learning
  • Embedding-based Approach
  • Hallucination-based Approach
  • Meta-Learning based Approach

• Domain Generalization
  • Domain-Generalized Image Classification
  • Cross-Domain (Domain-Generalized) Few-Shot Classification
Recap: Domain Confusion

• Domain-Adversarial Training of Neural Networks (DANN)
  • Y. Ganin et al., ICML 2015
  • Maximize domain confusion = maximize domain classification loss
  • Minimize source-domain data classification loss
  • The derived feature $f$ can be viewed as a disentangled & domain-invariant feature.

\[
\frac{\partial L_y}{\partial \theta_f} + \lambda \frac{\partial L_d}{\partial \theta_f}
\]

\[
d_i \log \hat{d}_i + (1 - d_i) \log(1 - \hat{d}_i)
\]
Domain Generalization

- Input: Images and labels from multiple source domains
- Output: A well-generalized model for unseen target domains

\[ D_S = \{\text{Photo, Painting, Cartoon}\} \]
\[ D_T = \{\text{Sketch}\} \]
Episodic Training

- Episodic training for domain generalization (ICCV’19)
- Generalize across domains via Meta-Learning
Episodic Training (cont’d)

- Motivation
Episodic Training (cont’d)

- Random sample two domains, e.g., Photo and Cartoon
Episodic Training (cont’d)

- Random sample two domains, e.g., Photo and Cartoon
Episodic Training (cont’d)
Episodic Training (cont’d)

- Experiments and Results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0,1,2,3</td>
<td>4</td>
<td>61.5</td>
<td>75.8</td>
<td>78.0</td>
<td>75.8</td>
<td><strong>79.1</strong></td>
<td>75.0</td>
<td>70.7</td>
<td>71.6</td>
<td>74.2</td>
<td>73.1</td>
<td>76.9</td>
</tr>
<tr>
<td>0,1,2,4</td>
<td>3</td>
<td>72.5</td>
<td>86.9</td>
<td>92.3</td>
<td>92.3</td>
<td>94.5</td>
<td>94.1</td>
<td>93.6</td>
<td>93.8</td>
<td>94.0</td>
<td>94.2</td>
<td><strong>94.8</strong></td>
</tr>
<tr>
<td>0,1,3,4</td>
<td>2</td>
<td>74.7</td>
<td>84.5</td>
<td>91.2</td>
<td>94.5</td>
<td>95.6</td>
<td>97.3</td>
<td>97.5</td>
<td>95.7</td>
<td>96.9</td>
<td>95.7</td>
<td><strong>99.0</strong></td>
</tr>
<tr>
<td>0,2,3,4</td>
<td>1</td>
<td>67.0</td>
<td>83.4</td>
<td>90.1</td>
<td>91.2</td>
<td>93.4</td>
<td>95.4</td>
<td>95.4</td>
<td>94.2</td>
<td>97.0</td>
<td>95.7</td>
<td><strong>98.0</strong></td>
</tr>
<tr>
<td>1,2,3,4</td>
<td>0</td>
<td>71.4</td>
<td>92.3</td>
<td>93.4</td>
<td><strong>96.7</strong></td>
<td><strong>96.7</strong></td>
<td>95.7</td>
<td>93.6</td>
<td>94.0</td>
<td>94.7</td>
<td>94.4</td>
<td>96.3</td>
</tr>
<tr>
<td>Ave.</td>
<td></td>
<td>69.4</td>
<td>84.6</td>
<td>87.0</td>
<td>90.1</td>
<td>91.9</td>
<td>91.5</td>
<td>90.2</td>
<td>89.9</td>
<td>91.4</td>
<td>90.6</td>
<td><strong>93.0</strong></td>
</tr>
</tbody>
</table>

Table 1: Cross-view action recognition results (accuracy, %) on IXMAS dataset. Best result in bold.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>L,C,S</td>
<td>V</td>
<td>63.7</td>
<td>60.6</td>
<td>63.9</td>
<td>67.1</td>
<td><strong>67.7</strong></td>
<td>66.4</td>
<td>67.7</td>
<td>65.5</td>
<td>65.0</td>
<td>65.4</td>
<td>67.1</td>
</tr>
<tr>
<td>V,C,S</td>
<td>L</td>
<td>58.2</td>
<td>59.7</td>
<td>60.1</td>
<td>62.1</td>
<td>62.6</td>
<td>64.0</td>
<td>61.3</td>
<td>60.0</td>
<td>60.2</td>
<td>60.6</td>
<td><strong>64.3</strong></td>
</tr>
<tr>
<td>V,L,S</td>
<td>C</td>
<td>79.7</td>
<td>88.1</td>
<td>89.1</td>
<td>92.3</td>
<td><strong>94.4</strong></td>
<td>92.6</td>
<td><strong>94.4</strong></td>
<td>92.0</td>
<td>92.3</td>
<td>93.1</td>
<td>94.1</td>
</tr>
<tr>
<td>V,L,C</td>
<td>S</td>
<td>61.0</td>
<td>54.9</td>
<td>61.3</td>
<td>59.1</td>
<td>64.4</td>
<td>63.6</td>
<td><strong>65.9</strong></td>
<td>64.7</td>
<td>64.2</td>
<td>65.8</td>
<td><strong>65.9</strong></td>
</tr>
<tr>
<td>Ave.</td>
<td></td>
<td>65.7</td>
<td>65.8</td>
<td>68.6</td>
<td>70.2</td>
<td>72.3</td>
<td>71.7</td>
<td>72.3</td>
<td>70.5</td>
<td>70.4</td>
<td>71.2</td>
<td><strong>72.9</strong></td>
</tr>
</tbody>
</table>

Table 2: Cross-dataset object recognition results (accuracy, %) on VLCS benchmark. Best in bold.
What to Cover Today...

• Meta-Learning for Few-Shot Learning
  • Few-Shot Classification
    • Metric Learning vs. Data Hallucination
    • Few-Shot Image Segmentation

• Zero-Shot Learning
  • Embedding-based Approach
  • Hallucination-based Approach
  • Meta-Learning based Approach

• Domain Generalization
  • Domain-Generalized Image Classification
  • Cross-Domain (Domain-Generalized) Few-Shot Classification

Selected slide credits: C. Finn, S. Levine, H.-Y. Lee, & W.-L. H. Chao
Cross-Domain (or Domain Generalized) FSL

- Tackle N-way K-shot problems on **unseen domains**
- Novel classes (few-shot) and domains (cross-domain)
- Personally, I prefer to call it “domain-generalized” FSL instead of cross-domain FSL.
Recap: AdaIN

• Last week, we talked about style transfer methods like Pix2Pix or CycleGAN...
• **Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization (ICCV’17)**
  • Adaptive instance normalization (AdaIN) for arbitrary style transfer in real-time
AdaIN

Adaptive Instance Normalization

\[ \text{AdaIN}(x, y) = \sigma(y) \left( \frac{x - \mu(x)}{\sigma(x)} \right) + \mu(y) \]

- \( x \): content input, \( y \): style input
- No learnable affine parameters
- Perform style transfer in the feature space

Huang et al. ”Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization.” ICCV 2017
AdaIN

- $f$: Encoder, $g$: Decoder

\[
t = \text{AdaIN}(f(c), f(s))
\]

\[
\mathcal{L}_c = \|f(g(t)) - t\|_2
\]

Content loss: content consistency

\[
\mathcal{L}_s = \sum_{i=1}^{L} \|\mu(\phi_i(g(t))) - \mu(\phi_i(s))\|_2 + \sum_{i=1}^{L} \|\sigma(\phi_i(g(t))) - \sigma(\phi_i(s))\|_2
\]

Style loss: Gram matrix loss

($\phi_i$ denotes a layer in VGG – 19 used to compute style loss)

Huang et al. “Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization.” ICCV 2017
Cross-Domain FSL

- Cross-domain Few-shot Classification via Learned Feature-wise Transformation
- Feature-wise transformation layer for feature level augmentation

Cross-Domain FSL

- **Alternately** update the backbone model \((E, M)\) and the transformation layer.

Cross-Domain FSL

- Experiments
  - FT: hand-tuned feature transformer ($\gamma$ and $\beta$)
  - LFT: learned feature transformer

<table>
<thead>
<tr>
<th>5-way 5-Shot</th>
<th>CUB</th>
<th>Cars</th>
<th>Places</th>
<th>Plantae</th>
</tr>
</thead>
<tbody>
<tr>
<td>MatchingNet</td>
<td>-</td>
<td>51.92 ± 0.80%</td>
<td>39.87 ± 0.51%</td>
<td>61.82 ± 0.57%</td>
</tr>
<tr>
<td></td>
<td>FT</td>
<td>56.29 ± 0.80%</td>
<td>39.58 ± 0.54%</td>
<td>62.32 ± 0.58%</td>
</tr>
<tr>
<td></td>
<td>LFT</td>
<td>61.41 ± 0.57%</td>
<td>43.08 ± 0.55%</td>
<td>64.99 ± 0.59%</td>
</tr>
<tr>
<td>RelationNet</td>
<td>-</td>
<td>62.13 ± 0.74%</td>
<td>40.64 ± 0.54%</td>
<td>64.34 ± 0.57%</td>
</tr>
<tr>
<td></td>
<td>FT</td>
<td>63.64 ± 0.77%</td>
<td>42.24 ± 0.57%</td>
<td>65.42 ± 0.58%</td>
</tr>
<tr>
<td></td>
<td>LFT</td>
<td>64.99 ± 0.54%</td>
<td>43.44 ± 0.59%</td>
<td>67.35 ± 0.54%</td>
</tr>
<tr>
<td>GNN</td>
<td>-</td>
<td>69.26 ± 0.68%</td>
<td>48.91 ± 0.67%</td>
<td>72.59 ± 0.67%</td>
</tr>
<tr>
<td></td>
<td>FT</td>
<td>70.37 ± 0.68%</td>
<td>47.68 ± 0.63%</td>
<td>74.48 ± 0.70%</td>
</tr>
<tr>
<td></td>
<td>LFT</td>
<td>73.11 ± 0.68%</td>
<td>49.88 ± 0.67%</td>
<td>77.05 ± 0.65%</td>
</tr>
</tbody>
</table>

Cross-Domain FSL

- Visualization
  - Features derived from different domains are pulled together by LFT

What We’ve Covered Today...

• Meta-Learning for Few-Shot Learning
  • Few-Shot Classification
    • Metric Learning vs. Data Hallucination
    • Few-Shot Image Segmentation

• Zero-Shot Learning

• Domain Generalization

• Remember to...
  • Team up for the final project
  • Choose your preferable topic(s)