Bilim Akademisi - Bilkent Üniversitesi Yapay Öğrenme Yaz Okulu 2020

Gökberk Cinbiş

2. Kısım
Eksik Gözetimli Öğrenme
Part 1: Gradient Matching Networks

IEEE / CVF Conf. on Computer Vision and Pattern Recognition (CVPR), June 2019
Part 2: Zero-shot Detection and Captioning of Unseen Objects

♦: A couple of **elephants** standing next to each other.
★: A couple of **zebra** standing next to each other.

♦: A piece of **cake** on a white plate.
★: A piece of **pizza** on a white plate.
Part 3: Do we really need ZSL in practice?

British Machine Vision Conference (BMVC), September 2019
Part 4: Partially-supervised domain transfer for face recognition in the wildest
Part I

Gradient Matching Networks
Zero-shot object recognition

**Seen Classes**
- cow
- bird

**Unseen Classes**
- bat
- monkey

Training samples

**i** - Learn a classification model on **seen** classes

**ii** - Use the model for both sets
Semantic Class Embedding Space

(semantic feature - 1)

A

2

cow
dog

(semantic feature - 2)

A

1

bird

zebra

horse

monkey
Mainstream approach

Image Embedding

Class Embedding

$$f(x, \alpha ; \theta)$$
A weakness in purely discriminative approaches

Image Embedding

Class Embedding

\[ f(x, a; \theta) \]

Generative-model-based approaches

Seen Classes
- cow
- bird

Unseen Classes
- bat
  - has-wing
  - has-teeth
- monkey
  - has-arm
  - has-tail

Class embeddings

$\phi$  \hspace{1cm}  $G$

$\mathbf{f}(\mathbf{x}, \mathbf{a}; \theta)$

Examples:
First attempt: conditional GAN

A naive idea: *just train a conditional GAN model* (or another implicit generative model), which takes `concat(noise, class-embedding)` as the input.
First attempt: conditional GAN

A naive idea: *just train a conditional GAN model* (or another implicit generative model), which takes concat(noise, class-embedding) as the input.

.. but there are three important inter-connected challenges:

- **Semantics**: How do we enforce producing samples that truly belong to the target class?
- **Variance**: How do we enforce producing a variety of samples for a given embedding?
- **Data quality**: How do we make sure that the resulting training examples is actually useful? (ie. will the classifier trained over them be accurate?)
A second attempt

Train a conditional GAN using **GAN loss + loss function of a classifier** over the training classes.

At test time: simply synthesize training examples by feeding class-embeddings of test (unseen) classes to the GAN model.

**Good**: can leverage unsupervised data through the GAN loss.
**Good**: can enforce generating examples that are classified to the right class.
A second attempt

Train a conditional GAN using **GAN loss + loss function of a pre-trained classifier** over the training classes.

At test time: simply synthesize training examples by feeding class-embeddings of test (unseen) classes to the GAN model.

**Good**: can leverage unsupervised data through the GAN loss.

**Good**: can enforce generating examples that are classified to the right class.

However,

- The generated samples are not necessarily informative (**like support vectors**) ones (Likely, the generative model will learn to synthesize the "easy" samples.)
- The generated samples may contain **artifacts** detrimental for training purposes.
3rd attempt: a meta-learning approach

Assume that the synthetic (+real) examples will be used to train a classifier using a first-order gradient optimization technique.
3rd attempt: a meta-learning approach

Assume that the synthetic (+real) examples will be used to train a classifier using a first-order gradient optimization technique.

**Seen Classes**
- cow
- bird

**Unseen Classes**
- bat
- has-wing
- has-teeth
- monkey
- has-arm
- has-tail

Class embeddings

$$\theta^{(0)} \xrightarrow{-\nabla_\theta \mathcal{L}} \theta^{(1)} \xrightarrow{-\nabla_\theta \mathcal{L}} \ldots$$

$$f(x, a; \theta^{(N)})$$
Can we optimize $G$ such that we minimize the loss of final $f$?

Assume that the synthetic (+real) examples will be used to train a classifier using a first-order gradient optimization technique.
Can we optimize $G$ such that we minimize the loss of final $f$?

Assume that the synthetic (+real) examples will be used to train a classifier using a first-order gradient optimization technique.

Not naively. To measure the impact of $G$ on the final loss, we need make many model updates.

That is, just for a single $G$ update, we need to make many $f$ updates. **Very inefficient!**
Our idea

Assume that the synthetic (+real) examples will be used to train a classifier using a first-order gradient optimization technique.

Focus on learning to generate examples that maximizes the correctness of individual model updates.

The core idea: for classes with a good training set, the model updates over real versus synthetic examples shall be similar.
Gradient matching loss

\[ \mathcal{L}_{GM} = \mathbb{E}_\theta \left[ 1 - \frac{g_r(\theta)^T g_f(\theta)}{||g_r(\theta)||_2 \ ||g_f(\theta)||_2} \right] \]

Gradient by real

\[
g_r(\theta) = \mathbb{E}_{(x,a) \sim p_{\text{data}}} \left[ \nabla_\theta \mathcal{L}(x, a, f_\theta) \right]
\]

Gradient by generated

\[
g_f(\theta) = \mathbb{E}_{\tilde{x} \sim \mathcal{G}(z,a), a \sim p_{\text{data}}} \left[ \nabla_\theta \mathcal{L}(\tilde{x}, a, f_\theta) \right]
\]
To approximate the expectation over $\theta$

$$
\mathcal{L}_{GM} = \mathbb{E}_{\theta} \left[ 1 - \frac{g_r(\theta)^T g_f(\theta)}{\|g_r(\theta)\|_2 \|g_f(\theta)\|_2} \right]
$$

Repeatedly:

- train the classification model $N$ epochs,
- re-initialize all parameters and reset the optimizer state.
Gradient matching network (GMN)

Gradient matching loss
+ adversarial loss
(can be used for unsupervised learning)

\[ \mathcal{L}_{\text{GMN}} \]

\[ \mathcal{D}(x, a) \]

\[ \nabla \theta \]

\[ \mathcal{L}_{\text{GM}} \]

\[ f(x, a; \theta) \]

\[ \phi \]

\[ \mathcal{G} \]

Image Embedding

\( \text{has-wing} \)

\( \text{has-beak} \)

(proposed)
Experiments - Datasets

- Caltech-UCSD Birds-200-2011 (CUB) - 200 bird species - 12k

- SUN Attribute (SUN) - 717 scene categories - 14k

- Animals with Attributes (AWA) - 50 animal categories - 30k

Evaluation Metrics

Normalized score (NS) : average of the top-1 per-class scores

- **T-1** : NS of unseen classes in ZSL setting
- **u** : NS of unseen classes in GZSL setting
- **s** : NS of seen classes in GZSL setting
- **h** : harmonic mean of u and s \( \frac{2 \times u \times s}{u + s} \)
# Zero-shot prediction (unseen classes)

<table>
<thead>
<tr>
<th></th>
<th>CUB</th>
<th>SUN</th>
<th>AWA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T-1</td>
<td>T-1</td>
<td>T-1</td>
</tr>
<tr>
<td>1</td>
<td>Zhang et al. ’18</td>
<td>52.6</td>
<td>61.7</td>
</tr>
<tr>
<td>2</td>
<td>Bucher et al. ’17</td>
<td>57.8</td>
<td>60.4</td>
</tr>
<tr>
<td>3</td>
<td>Xian et al. - DEVISE ’18</td>
<td>60.3</td>
<td>60.9</td>
</tr>
<tr>
<td>4</td>
<td>Xian et al. - ALE ’18</td>
<td>61.5</td>
<td>62.1</td>
</tr>
<tr>
<td>5</td>
<td>Xian et al. - Softmax ’18</td>
<td>57.3</td>
<td>60.8</td>
</tr>
<tr>
<td>6</td>
<td>Verma et al. ’18</td>
<td>59.6</td>
<td>63.4</td>
</tr>
<tr>
<td>7</td>
<td>Felix et al. - cycle-WGAN ’18</td>
<td>57.8</td>
<td>59.7</td>
</tr>
<tr>
<td>8</td>
<td>Felix et al. - cycle-CLSWGAN ’18</td>
<td>58.4</td>
<td>60.0</td>
</tr>
<tr>
<td>9</td>
<td>Bilinear</td>
<td>LN</td>
<td>$\mathcal{L}_{\text{S\text{-WGAN}}}$</td>
</tr>
<tr>
<td>10</td>
<td>Bilinear</td>
<td>LN</td>
<td>$\mathcal{L}<em>{\text{S\text{-WGAN}} + \mathcal{L}</em>{\text{CLS}}}$</td>
</tr>
<tr>
<td>11</td>
<td>Bilinear</td>
<td>LN</td>
<td>$\mathcal{L}<em>{\text{S\text{-WGAN}} + \mathcal{L}</em>{\text{CYCLE}}}$</td>
</tr>
<tr>
<td>12</td>
<td>Bilinear</td>
<td>LN</td>
<td>$\mathcal{L}<em>{\text{S\text{-WGAN}} + \mathcal{L}</em>{\text{GM}}}$ (Ours)</td>
</tr>
<tr>
<td>13</td>
<td>Linear</td>
<td>LN</td>
<td>$\mathcal{L}<em>{\text{S\text{-WGAN}} + \mathcal{L}</em>{\text{GM}}}$ (Ours)</td>
</tr>
<tr>
<td>14</td>
<td>Bilinear</td>
<td>AC</td>
<td>$\mathcal{L}<em>{\text{S\text{-WGAN}} + \mathcal{L}</em>{\text{GM}}}$ (Ours)</td>
</tr>
<tr>
<td>15</td>
<td>Linear</td>
<td>AC</td>
<td>$\mathcal{L}<em>{\text{S\text{-WGAN}} + \mathcal{L}</em>{\text{GM}}}$ (Ours)</td>
</tr>
</tbody>
</table>
## Generalized zero-shot prediction (seen + unseen classes)

<table>
<thead>
<tr>
<th></th>
<th>CUB</th>
<th></th>
<th></th>
<th></th>
<th>SUN</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>AWA</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>u</td>
<td>s</td>
<td>h</td>
<td>u</td>
<td>s</td>
<td>h</td>
<td>u</td>
<td>s</td>
<td>h</td>
<td>u</td>
<td>s</td>
<td>h</td>
<td>u</td>
</tr>
<tr>
<td>1</td>
<td>31.5</td>
<td>40.2</td>
<td>35.3</td>
<td>41.2</td>
<td>26.7</td>
<td>32.4</td>
<td>38.7</td>
<td>74.6</td>
<td>51.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>28.8</td>
<td>55.7</td>
<td>38.0</td>
<td>40.5</td>
<td>37.2</td>
<td>38.8</td>
<td>2.3</td>
<td>90.2</td>
<td>4.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>52.2</td>
<td>42.4</td>
<td>46.7</td>
<td>38.4</td>
<td>25.4</td>
<td>30.6</td>
<td>35.0</td>
<td>62.8</td>
<td>45.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>40.2</td>
<td>59.3</td>
<td>47.9</td>
<td>41.3</td>
<td>31.1</td>
<td>35.5</td>
<td>47.6</td>
<td>57.2</td>
<td>52.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>43.7</td>
<td>57.7</td>
<td>49.7</td>
<td>42.6</td>
<td>36.6</td>
<td>39.4</td>
<td>57.9</td>
<td>61.4</td>
<td>59.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>41.5</td>
<td>53.3</td>
<td>46.7</td>
<td>40.9</td>
<td>30.5</td>
<td>34.9</td>
<td>56.3</td>
<td>67.8</td>
<td>61.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>46.0</td>
<td>60.3</td>
<td>52.2</td>
<td>48.3</td>
<td>33.1</td>
<td>39.2</td>
<td>56.4</td>
<td>63.5</td>
<td>59.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>45.7</td>
<td>61.0</td>
<td>52.3</td>
<td>49.4</td>
<td>33.6</td>
<td>40.0</td>
<td>56.9</td>
<td>64.0</td>
<td>60.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bilinear</td>
<td>LN</td>
<td>$\mathcal{L}_{cWGAN}^S$</td>
<td>45.6</td>
<td>59.2</td>
<td>51.5</td>
<td>50.6</td>
<td>30.3</td>
<td>37.3</td>
<td>53.5</td>
<td>72.0</td>
<td>61.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bilinear</td>
<td>LN</td>
<td>$\mathcal{L}<em>{cWGAN}^S + \mathcal{L}</em>{CLS}$</td>
<td>45.5</td>
<td>58.9</td>
<td>51.4</td>
<td>50.6</td>
<td>30.3</td>
<td>37.3</td>
<td>52.7</td>
<td>71.0</td>
<td>60.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bilinear</td>
<td>LN</td>
<td>$\mathcal{L}<em>{cWGAN}^S + \mathcal{L}</em>{CYCLE}$</td>
<td>51.1</td>
<td>54.9</td>
<td>52.9</td>
<td>50.6</td>
<td>30.3</td>
<td>37.3</td>
<td>55.4</td>
<td>70.1</td>
<td>61.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bilinear</td>
<td>LN</td>
<td>$\mathcal{L}<em>{cWGAN}^S + \mathcal{L}</em>{GM}$ (Ours)</td>
<td><strong>54.7</strong></td>
<td><strong>58.4</strong></td>
<td><strong>56.5</strong></td>
<td>42.5</td>
<td>35.5</td>
<td>38.7</td>
<td><strong>61.1</strong></td>
<td>71.3</td>
<td>65.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear</td>
<td>LN</td>
<td>$\mathcal{L}<em>{cWGAN}^S + \mathcal{L}</em>{GM}$ (Ours)</td>
<td><strong>48.5</strong></td>
<td><strong>62.8</strong></td>
<td>54.7</td>
<td>42.0</td>
<td><strong>39.3</strong></td>
<td>40.7</td>
<td>57.1</td>
<td>81.3</td>
<td><strong>67.1</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bilinear</td>
<td>AC</td>
<td>$\mathcal{L}<em>{cWGAN}^S + \mathcal{L}</em>{GM}$ (Ours)</td>
<td>53.8</td>
<td>58.2</td>
<td>55.9</td>
<td>43.2</td>
<td>36.2</td>
<td>39.4</td>
<td>54.8</td>
<td>74.1</td>
<td>63.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear</td>
<td>AC</td>
<td>$\mathcal{L}<em>{cWGAN}^S + \mathcal{L}</em>{GM}$ (Ours)</td>
<td>45.8</td>
<td>65.5</td>
<td>53.9</td>
<td><strong>53.2</strong></td>
<td>33.0</td>
<td><strong>42.8</strong></td>
<td>46.8</td>
<td>84.8</td>
<td>60.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Gokberk Cinbis - 2020
In summary

- **a novel** proxy loss for **zero-shot learning**
  - better estimation of class distributions
- **state of the art** on CUB, AWA and SUN

Source code: [https://mbsariyildiz.github.io/](https://mbsariyildiz.github.io/)
Part II

Zero-shot Detection and Captioning of Unseen Objects
Why study zero-shot detection?

Detection in the Wild using text-based queries

Robotic
Our approach

➔ Our method consists of two components:
  ◆ (i) utilize a convex combination of class embeddings,
  ◆ (ii) directly learn to map regions to the space of class embeddings.
➔ Zero-shot object detection within the YOLO detection framework.
Convex Combination of Class Embeddings

- Represent a given image region (i.e. a bounding box) as the convex combination of training class embeddings.

\[
f_{CC}(x, b, y) = \frac{\phi_{CC}(x, b)^T \eta(y)}{\|\phi_{CC}(x, b)\| \|\eta(y)\|}
\]

\[
\phi_{CC}(x, b) = \frac{1}{\sum_{y \in \mathcal{Y}_s} p(y|x, b)} \sum_{y \in \mathcal{Y}_s} p(y|x, b) \eta(y)
\]
Convex Combination of Class Embeddings

- Represent a given image region (i.e. a bounding box) as the convex combination of training class embeddings.

\[
f_{CC}(x, b, y) = \frac{\phi_{CC}(x, b)^T \eta(y)}{\|\phi_{CC}(x, b)\| \|\eta(y)\|}
\]

\[
\phi_{CC}(x, b) = \frac{1}{\sum_{y \in \mathcal{Y}_s} p(y|x, b)} \sum_{y \in \mathcal{Y}_s} p(y|x, b) \eta(y)
\]

Sum of class embeddings, weighted by posterior probability
Region Scoring by Label Embedding

- The goal is to directly model the compatibility between the visual features of image regions and class embeddings.
- The equation can be interpreted as a dot product between L2-normalized image region descriptors and class embeddings.
Hybrid region embedding

- The two scores are accumulated within the loss function:

\[ L_{LE}(x, b, y) = \frac{1}{|\mathcal{Y}_s| - 1} \sum_{y' \in \mathcal{Y}_s \setminus \{y\}} \max (0, 1 - f_{LE}(x, b, y) + f_{LE}(x, b, y')) \]
Experimental Results on PASCAL VOC

- Select 16 of the 20 classes as the training set.
- Remaining 4 classes as the test set. These test classes are car, dog, sofa and train respectively.
- Class-attribute relations of aPaY dataset are used for semantic descriptions.
- 65.6% mAP on seen classes, 54.6% mAP on unseen ones.

<table>
<thead>
<tr>
<th>Method</th>
<th>Test split</th>
<th>aeroplane</th>
<th>bicycle</th>
<th>bird</th>
<th>boat</th>
<th>bottle</th>
<th>bus</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
<th>dining table</th>
<th>horse</th>
<th>motorbike</th>
<th>person</th>
<th>potted plant</th>
<th>sheep</th>
<th>tvmonitor</th>
<th>mAP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LE</td>
<td>v</td>
<td>.46</td>
<td>.50</td>
<td>.44</td>
<td>.28</td>
<td>.12</td>
<td>.59</td>
<td>.44</td>
<td>.20</td>
<td>.11</td>
<td>.38</td>
<td>.35</td>
<td>.47</td>
<td>.65</td>
<td>.16</td>
<td>.18</td>
<td>.53</td>
<td>36.8</td>
</tr>
<tr>
<td></td>
<td>t</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>47.9</td>
</tr>
<tr>
<td></td>
<td>v+t</td>
<td>.34</td>
<td>.48</td>
<td>.40</td>
<td>.23</td>
<td>.12</td>
<td>.34</td>
<td>.28</td>
<td>.12</td>
<td>.09</td>
<td>.32</td>
<td>.28</td>
<td>.36</td>
<td>.60</td>
<td>.15</td>
<td>.13</td>
<td>.50</td>
<td>27.4</td>
</tr>
<tr>
<td>CC</td>
<td>v</td>
<td>.69</td>
<td>.74</td>
<td>.72</td>
<td>.63</td>
<td>.43</td>
<td>.83</td>
<td>.73</td>
<td>.43</td>
<td>.43</td>
<td>.66</td>
<td>.78</td>
<td>.80</td>
<td>.75</td>
<td>.41</td>
<td>.62</td>
<td>.75</td>
<td>65.0</td>
</tr>
<tr>
<td></td>
<td>t</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>v+t</td>
<td>.67</td>
<td>.73</td>
<td>.70</td>
<td>.59</td>
<td>.41</td>
<td>.61</td>
<td>.58</td>
<td>.32</td>
<td>.32</td>
<td>.65</td>
<td>.74</td>
<td>.68</td>
<td>.72</td>
<td>.39</td>
<td>.57</td>
<td>.72</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>v</td>
<td>.70</td>
<td>.73</td>
<td>.76</td>
<td>.54</td>
<td>.42</td>
<td>.86</td>
<td>.64</td>
<td>.40</td>
<td>.54</td>
<td>.75</td>
<td>.80</td>
<td>.80</td>
<td>.75</td>
<td>.34</td>
<td>.69</td>
<td>.79</td>
<td>53.8</td>
</tr>
<tr>
<td></td>
<td>t</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>52.0</td>
</tr>
<tr>
<td></td>
<td>v+t</td>
<td>.68</td>
<td>.72</td>
<td>.74</td>
<td>.48</td>
<td>.41</td>
<td>.61</td>
<td>.48</td>
<td>.25</td>
<td>.48</td>
<td>.73</td>
<td>.75</td>
<td>.71</td>
<td>.73</td>
<td>.33</td>
<td>.59</td>
<td>.57</td>
<td><strong>65.6</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>54.2</strong></td>
</tr>
</tbody>
</table>

Gokberk Cinbis - 2020
Example detections
Captioning with Unseen Objects

- **Motivation:** Overcome the data collection bottleneck in image captioning.
- **Task:** Define a new paradigm for generating captions of unseen classes.
- **Key Idea:** Use zero-shot object detector with template based sentence generator.
Zero-shot Image Captioning

Visual Input

Textual Input

“a person riding a horse”
**Zero-shot Image Captioning**

<table>
<thead>
<tr>
<th>Visual Input</th>
<th>Image Captioning</th>
<th>(Partial) Zero-Shot Image Captioning</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Visual Input Image" /></td>
<td><img src="caption1" alt="Caption" /></td>
<td><img src="caption2" alt="Caption" /></td>
</tr>
<tr>
<td><img src="image2" alt="Visual Input Image" /></td>
<td><img src="caption3" alt="Caption" /></td>
<td><img src="caption4" alt="Caption" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Textual Input</th>
<th>Image Captioning</th>
<th>(Partial) Zero-Shot Image Captioning</th>
</tr>
</thead>
<tbody>
<tr>
<td>“a person riding a horse”</td>
<td><img src="caption1" alt="Caption" /></td>
<td><img src="caption2" alt="Caption" /></td>
</tr>
<tr>
<td>“a person riding a horse”</td>
<td><img src="caption3" alt="Caption" /></td>
<td><img src="caption4" alt="Caption" /></td>
</tr>
</tbody>
</table>
## Zero-shot Image Captioning

<table>
<thead>
<tr>
<th>Visual Input</th>
<th>Image Captioning</th>
<th>(Partial) Zero-Shot Image Captioning</th>
<th>True Zero-Shot Image Captioning</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Caption" /></td>
<td><img src="image3.png" alt="Caption" /></td>
<td><img src="image4.png" alt="Caption" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Textual Input</th>
<th>“a person riding a horse”</th>
<th>“a person riding a horse”</th>
<th>“a person riding a horse”</th>
</tr>
</thead>
</table>

{person, horse} ∈ unseen classes
Framework - Fully Zero-shot Image Captioning

Generalized Zero-Shot Object Detector

Improving ZSD: Generalized Zero-shot Detection

- Unlike the prior work on ZSD, test captioning images contain a mixture of seen and unseen classes.

- Typically there is a significant bias towards the seen classes.

- Aim to overcome this problem by learning a scaling coefficient:

\[
    f(x, c, i) = \begin{cases} 
    \alpha f(x, c, i), & \text{if } c \in \hat{Y}_s \\
    f(x, c, i), & \text{otherwise}
    \end{cases}
\]
Improving (G)ZSD - Better embeddings

- Reminder: detection scoring function \( f(x, c, i) \) is defined as follows:
  \[
  f(x, c, i) = \frac{\Omega(x, i)^T \psi(c)}{||\Omega(x, i)|| ||\psi(c)||}
  \]

- Here, \( \psi(c) \) represents the class embedding for class \( c \), which is now obtained in terms of target-class to training-class similarities in the word embedding space:
  \[
  \psi(c) = [\varphi(c)^T \varphi(\bar{c}) + 1]_{\bar{c}}
  \]

- We also drop the convex combination approach to be able to deal with GZSD better.
Experimental Setup

- **Dataset**: MS-COCO splits for evaluating zero-shot image captioning.
- **Evaluation**: F1 score, METEOR, SPICE, ROUGE-L, BLEU metrics.
- **Class embeddings**: Use 300-dim word2vec of class embeddings.
- **Evaluation - GZSD**: Use COCO val5k split, which contains both seen and unseen class instances.
## Generalized-ZSD results

<table>
<thead>
<tr>
<th>Classes</th>
<th>GZSD w/o $\alpha$</th>
<th>GZSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bottle</td>
<td>0</td>
<td>0.8</td>
</tr>
<tr>
<td>Bus</td>
<td>0</td>
<td>21.4</td>
</tr>
<tr>
<td>Couch</td>
<td>2.7</td>
<td>4.9</td>
</tr>
<tr>
<td>Microwave</td>
<td>0</td>
<td>1.2</td>
</tr>
<tr>
<td>Pizza</td>
<td>0</td>
<td>4.8</td>
</tr>
<tr>
<td>Racket</td>
<td>0</td>
<td>0.7</td>
</tr>
<tr>
<td>Suitcase</td>
<td>0</td>
<td>9.1</td>
</tr>
<tr>
<td>Zebra</td>
<td>0</td>
<td>15.8</td>
</tr>
<tr>
<td>U-mAP(%)</td>
<td>0.3</td>
<td>7.3</td>
</tr>
<tr>
<td>S-mAP(%)</td>
<td>27.4</td>
<td>19.2</td>
</tr>
<tr>
<td>Harmonic Mean</td>
<td>0.7</td>
<td>10.6</td>
</tr>
</tbody>
</table>

Typically, an unseen class instance is detected as the instance of some seen class.
Image Captioning Results

Comparison Results

<table>
<thead>
<tr>
<th>Metric</th>
<th>NBT-Baseline</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>METEOR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPICE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROUGE-L</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLEU1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A piece of **cake** on a white plate.

A yellow and black **train** traveling down the road.

A piece of **pizza** on a white plate.

A yellow and black **bus** driving down a road.
Image Captioning Results

Comparison Results

<table>
<thead>
<tr>
<th>Metric</th>
<th>NBT-Baseline</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>METEOR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPICE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROUGE-L</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLEU1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **NBT-Baseline**: A piece of cake on a white plate. A yellow and black train traveling down the road.
- **Proposed**: A piece of pizza on a white plate. A yellow and black bus driving down a road.
Qualitative Results

Image captioning results of images which consist of seen and unseen classes:

A small white **dog** is sitting on a **couch**.

A red **bus** is driving down the street.

A couple of **zebra** standing in a field.

A **tennis player** is about to hit a **racket**.

A white plate topped with a piece of **pizza**.

A kitchen with a **microwave** and a counter.
In summary,

- a **new** problem: generating captions of images with **unseen classes**.
- a **novel** approach for generalized zero-shot object detection problem.
Part III

Do we really need ZSL in practice?
Do we really need ZSL in practice?

- ZSL sometimes sounds like a practically irrelevant problem given that there are several large-scale datasets, such as Google Open Images, ImageNet, etc.

However:

1) These datasets are arguably still very far from capturing **richness** of human vision

2) Large-scale data collection can be **inherently difficult** due to physical constraints, lack of annotation experts, etc. in certain problems.
Do we really need ZSL in practice?

- I will talk about two examples:
  - Fine-grained recognition in remote sensing
  - Sign Language Recognition
Traditional Object Recognition in Remote Sensing

- The mainstream object recognition task in remote sensing:
  - Benchmark datasets: UC Merced, AID, etc.
  - Assign each pixel/patch to one of few categories
  - *eg.* Agricultural vs Beach vs Forest vs Freeway vs Harbor
  - Typically there are a large number of examples per class

- Distinct classes

- Largely an over-simplified categorization of earth surface
Fine-grained Recognition in Remote Sensing

- Under-studied problem: fine-grained, semantically rich recognition
- Our focus: 40 different tree species and their satellite views
- We manually cleaned 48k GPS-tagged samples belonging to 40 top categories
Formulation

- Ovate Leaf
- No Thorn on Trunk Bark
- Regular Crown
- Word2Vec
- Taxonomy Hierarchy

\[ P(y = \text{Narrowleaf Ash}|x) \]
Multisource Region Attention Network
The quantitative annotation advantage of ZSL

<table>
<thead>
<tr>
<th>Supervised Classification Results (in %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random guess</td>
</tr>
<tr>
<td>---------------</td>
</tr>
<tr>
<td>Normalized accuracy</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Zero-Shot Learning Results (in %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>---------------</td>
</tr>
<tr>
<td>Normalized accuracy</td>
</tr>
</tbody>
</table>

- ZSL is advantageous up to 256 supervised samples
- Note that (i) ZSL uses no examples, (ii) most categories are hard to distinguish even by visual inspection
Data collection problem, revisited

- Collection of even 256 samples can be very costly

- Just not possible to annotation by looking into the images. It is necessary to physically visit the instances & GPS-tag them.

- 16-test classes, around Seattle (WA), scattered around 217 km²

- Arguably not feasible for scaling up for monitoring tree species all over the world.
Most of the SLR approaches require a large amount of annotated data to recognize predefined sign classes.

Problem 1: Excessive manual annotation

Problem 2: What if we want to recognize other signs?
Learning Signs with Minimal Supervision

- Languages are constantly growing
  - Thousands of new words are added to OED every year.
- Same is true for Sign Languages

New words list March 2019

- anti-suffragism, n.: “Opposition to the extension of the right to vote in political elections to women; the political movement dedicated to this.”
- Aperol, n.: “A proprietary name for: an orange-coloured Italian aperitif flavoured with gentian, rhubarb, and a variety of herbs and roots. Also: a drink of this.”
- archicembalo, n.: “Any of various types of harpsichord having more than twelve keys to the octave and therefore capable of producing intervals smaller than a semitone…”
- Argonautical, adj.: “Of, relating to, or likened to the Argonauts. Cf. Argonautic adj.”
- Assiniboin, n. and adj.: “A member of a Siouan people of the Great Plains, now living mainly in southern Saskatchewan and northern Montana.”
- Auckland, n.: “A native or inhabitant of city or region of Auckland, New Zealand.”
- baff, n.2: “A slipper; = baffie n. Usually in plural.”
- baffie, n.: “A slipper, esp. one that is old and worn out (cf. bauchie n.1). Usually in plural. Cf. baff n.2.”
- baggataway, n.: “The game of lacrosse, esp. as played by certain indigenous peoples of eastern Canada and the midwestern and northeastern United States, using sticks…”
Zero-shot Sign Language Recognition (ZSSLR)

**BICYCLE:** Move both S hands in alternating forward circles, palms facing down, in front of each side of the body.

**HIGH:** Move the right H hand, palm facing left and fingers pointing forward, from in front of the right side of the chest upward to near the right side of the head.
Our ZSSLR model

Beginning with both 4 hands in front of the chest, fingers pointing in opposite directions and overlapping, both palms facing in, move the hands outward to in front of each shoulder.
Our ZSSLR model

Frame embedding

Body stream

Hand stream

Text

Beginning with both 4 hands in front of the chest, fingers pointing in opposite directions and overlapping, both palms facing in, move the hands outward to in front of each shoulder.

3D-CNN (13D)

Spatio-temporal representation (short-term)

Bidirectional LSTM

1024-d

Concatenate

Temporal modeling (longer-term)

Encoder

Encoder

Encoder

F(v, t; W)

Compatibility function
Our ZSSLR Model

Temporal model

Body stream

3D-CNN (13D)

Hand stream

3D-CNN (13D)

Spatio-temporal representation (short-term)

Bidirectional LSTM

1024-d

Concatenate

Text

Beginning with both 4 hands in front of the chest, fingers pointing in opposite directions and overlapping, both palms facing in, move the hands outward to in front of each shoulder.

Encoder

Encoder

Encoder

Beginning with shoulder

BERT Embedding

Compatibility function

\[ F(v, t; W) \]
Our ZSSLR model

Beginning with both 4 hands in front of the chest, fingers pointing in opposite directions and overlapping, both palms facing in, move the hands outward to in front of each shoulder.
Our ZSSLR model

Beginning with both 4 hands in front of the chest, fingers pointing in opposite directions and overlapping, both palms facing in, move the hands outward to in front of each shoulder.
Text description embeddings

**OBSCURE**

Beginning with the left 5 hand in front of the chest, palm facing in, and the right 5 hand by the right side of the body, palm facing forward, bring the right hand in an arc past the left hand, ending with the wrists crossed.

**FRIEND**

Hook the bent right index finger, palm facing down, over the bent left index finger, palm facing up. Then repeat, reversing the position of the hands.

**HAMBURGER**

Clasp the right curved hand, palm facing down, across the upturned left curved hand. Then flip the hands over and repeat with the left hand on top.

**MOST**

Beginning with the palm sides of both 10 hands together in front of the chest, bring the right hand upward, ending with the right hand in front of the right shoulder, palm facing left.

**COMB**

Drag the fingertips of the right curved 5 hand through the hair on the right side of the head with a short double movement.

**BOSS**

Tap the fingertips of the right curved 5 hand on the right shoulder with a repeated movement.
### ZSSLR Experimental Results

<table>
<thead>
<tr>
<th>Temporal Representation</th>
<th>top-1</th>
<th>top-2</th>
<th>top-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>AvePool</td>
<td>18.0</td>
<td>27.4</td>
<td>43.8</td>
</tr>
<tr>
<td>LSTM</td>
<td>18.2</td>
<td>28</td>
<td>47.2</td>
</tr>
<tr>
<td>GRU</td>
<td>19.7</td>
<td>31.8</td>
<td>50.0</td>
</tr>
<tr>
<td>bi-LSTM</td>
<td><strong>20.9</strong></td>
<td><strong>32.5</strong></td>
<td><strong>51.4</strong></td>
</tr>
</tbody>
</table>

Accuracies are still quite low, large room for improvement

Dataset available for download: [https://ycbilge.github.io/zsslr.html](https://ycbilge.github.io/zsslr.html)
Conclusions

● Presented two problems where the need for zero-shot learning naturally emerges:

1. Fine-grained recognition in remote sensing – towards globally monitoring all tree species

2. Sign Language Recognition – towards recon – towards recognizing all words in all sign languages, with quick adaptation to novel words
Part IV

Partially-supervised domain transfer for face recognition in the wildest
Face Recognition in the “Wildest”

- Face Recognition is largely solved in controlled cases (> 95% accuracy).
- People in criminal activity expose a diverse set of facial expressions
- These people may not necessarily have prior criminal records
  - Only have passport or Facebook type photos
WildestFaces Dataset

Collected video scenes from YouTube
- Car chase, fist fights, gun fights, heated arguments
- 64 actors
- 2186 shots - 64,242 frames

Clean images from IMDB-WIKI & Internet
- 64 actors
- 8069 images

Some may not have prior violent footage → Partially supervised

GZSL-like setting
- **Train**: clean images of 64 classes, videos of 40 seen classes
- **Val**: videos of 40 seen classes, 10 unseen classes
- **Test**: videos of 64 classes (seen + unseen)
Partially Supervised Domain Transfer

\[
\min_W R(W) + \sum_{i=1}^{n_x} \ell(\phi(x_i)^T W, y_i)
\]

Image representation

\[
\min_\tau R(\tau) + \sum_{j=1}^{n_v} \ell(\tau(\Psi(v_j))^T W, y_j)
\]

Classifier transfer + temporal adaptation
Classifier transfer layer

\[
\min_{\tau} R(\tau) + \sum_{j=1}^{n_v} \ell(\tau(\Psi(v_j))^T W, y_j)
\]

**Fully-connected classifier transfer**

\[
\tau_{fc}(\Psi(v)) = Q\Psi(v)
\]

**Affine classifier transfer**

\[
\tau_{affine}(\Psi(v)) = \alpha \odot \Psi(v) + \beta
\]

**Residual stacked affine**

\[
\tau_{rsa}(\Psi(v)) = \alpha_2 \odot \max(\alpha_1 \odot \Psi(v) + \beta_1, 0) + \beta_2 + \Psi(v)
\]
**Temporal adaptation**

\[
\min_{\tau} R(\tau) + \sum_{j=1}^{n_v} \ell(\tau(\Psi(v_j))^\top W, y_j)
\]

\[
\Psi_{\text{AvgPool}}(v) = \frac{1}{|v|} \sum_{t=1}^{|v|} \phi(v[t])
\]

\[
\Psi_{\text{ATP}}(v) = \frac{1}{K} \Phi(v) \Gamma(v) 1_K
\]

\[
[\Gamma(v)]_{t,k} = \frac{\exp [\phi(v[t])^\top a_k]}{\sum_{t'=1}^{|v|} \exp [\phi(v[t'])^\top a_k]}
\]

- Temporal average pooling
- Attentive temporal pooling
- Attention matrix
Classifier Transfer

seen  unseen  harmonic

random  w/o transfer  fc  MMD [30]  MMD-Affine  Ours (Affine)
Comparison to literature

- DAN[37]
- AvgPool
- Self-attention[50]
- ATP(Ours)

seen
unseen
harmonic
Summary

- Towards semantically rich recognition systems, build models that are
  - more flexible
  - more tightly integrated with language
  - requires less supervision
- Presented:
  - Gradient Matching Networks (currently for ZSL)
  - A zero-shot object detection approach, with application to image captioning
  - Two real-world applications of ZSL
  - A partially supervised model domain transfer problem
Thank you!