Compositional Zero-Shot Learning via Fine-Grained Dense Feature Composition

Dat Huynh and Ehsan Elhamifar

Khoury College of Computer Sciences
Northeastern University
Motivation

• **Fine-Grained Recognition**: recognize visually similar classes
  • Classes differ in a few attributes
  • **Costly**: require expert annotator
  • **Cannot handle** unseen classes
Motivation

- **Fine-Grained Recognition**: recognize visually similar classes
  - Classes *differ in a few attributes*
  - **Costly**: require expert annotator
  - **Cannot handle** unseen classes

- **Zero-Shot Learning**: recognize unseen classes without training samples
  - **Reduce** annotation cost

Using attribute descriptions
State of the Art

• **Generative Methods:** train a classifier on unseen class features synthesized by a generator

• **Challenge:**
  - Generate holistic features **lacking attribute details**

[M. Bucher et. al. '17, G. Arora et. al. '18, Y. Xian et. al. '19, E. Schönfeld et. al. '19, H. Huang et. al. '19, H. Yu et. al. '19, J. Ni et. al. '19]
Contributions

• Address fine-grained ZSL
• Propose a *compositional feature learning* framework:
  • Generate *dense features* preserving fine-grained details
  • Directly train a classifier *without learning separate generative models*
  • Outperform SOTA on DeepFashion, AWA2 and CUB

Extract one feature per attribute
*[D. Huynh and E. Elhamifar, CVPR20]*
Challenges

- Generate **dense feature** $H$: collection of **attribute features**

  ![Diagram showing dense feature and attribute features]

- **Challenge**: Difficult to learn generative models
  - **High dimension**
  - **No sample** for unseen classes
Feature Composition

• **Transform** a discriminative model \( p(y|H, z) \) into a generative model

\[
\arg\max_H p(H|y, z) = \arg\max_H p(y|H, z)p(H|z)
\]

• **Challenge:**
  • Search for most probable feature in high dimension space of \( H \) is intractable
Feature Composition

- Limit our search in combinations of **attribute features across samples**
  - Simulate **unseen attribute combinations**
  - **Tractable** due to finite number of combinations

\[
\mathcal{U} \triangleq \begin{cases} 
\vdots & \vdots & \vdots & \vdots & \cdots \\
\vdots & \vdots & \vdots & \vdots & \cdots \\
\vdots & \vdots & \vdots & \vdots & \cdots \\
\vdots & \vdots & \vdots & \vdots & \cdots \\
\end{cases}
\]

Search in all combinations is **expensive**

How to construct **feature prior**?

\[
\arg\max_H p(H | y, z) \approx \arg\max_{H \in \mathcal{U}} p(y | H, z)p(H | z)
\]
Feature Prior

- **Restrict the search** to combinations from related sample sets $Q_u$
  - **Set of samples best reconstruct** the unseen class attributes

$$Q_u \triangleq \arg\min_S \left( \min_\gamma \| z - \sum_{i \in S} z^i \gamma_i \|_2^2 \right)$$

- **Construct feature prior**
  - **Independence** among attribute features
  - The more related a sample, the more probable its feature will be used

$$p(H | z^u) \triangleq \prod_{a=1}^A p(h_{i_a}^a | z^u), \quad p(h_{i_a}^a | z^u) \triangleq \begin{cases} \exp \left( \langle z_{i_a}^a, z^u \rangle \right), & \text{if } i_a \in Q_u, \\ 0, & \text{otherwise,} \end{cases}$$
**Proposed Method**

1. **Generate candidate combinations** from feature prior

2. **Pick the most probable combination** among candidates

\[
M_u = \{H | H \sim p(H | z^u)\}
\]

\[
H_u \triangleq \arg\max_{H \in M_u} p(u | H, z^u)p(H | z^u)
\]
Proposed Method

1. Generate candidate combinations from feature prior
2. Pick the most probable combination among candidates
3. Train the discriminative model on real and composed features

$$\min \mathbb{E}_S \left[ -\frac{1}{|S|} \sum_{i \in S} y_i \log p(y_i | H_i z^{y_i}) - \frac{1}{|C_u|} \sum_{u \in C_u} u \log p(u | H_u, z^u) \right]$$

Cross-entropy with seen class features
Cross-entropy with composed features
## Experiments

- **Outperform SOTA** on DeepFashion, AWA2, CUB

<table>
<thead>
<tr>
<th>Method</th>
<th><strong>DFashion</strong> (5691 images/class)</th>
<th><strong>AWA2</strong> (588 images/class)</th>
<th><strong>CUB</strong> (47 images/class)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$u \rightarrow u$</td>
<td>$a \rightarrow s$</td>
<td>$a \rightarrow u$</td>
</tr>
<tr>
<td>CVC</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TripletLoss</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>f-VAEGAN-d2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CADA-VAE</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>f-Translator</td>
<td>40.7</td>
<td>30.5</td>
<td>23.9</td>
</tr>
<tr>
<td>DAZLE</td>
<td>38.4</td>
<td><strong>38.1</strong></td>
<td>21.5</td>
</tr>
<tr>
<td>Composer (Ours)</td>
<td><strong>43.0</strong></td>
<td>32.9</td>
<td><strong>31.2</strong></td>
</tr>
</tbody>
</table>

**Note:** Figures indicate percentage improvements over baseline performances.
Qualitative Results

• Generated features are **interpretable**

Attributes from seen images used for composition

- Composed Features
  - head color blue
  - wing pattern striped
  - under tail color gray

Unseen Class

- Composed Features
  - forehead color black
  - belly color yellow
  - throat color yellow