Multi-Source Augmentation and Composite Prompts for Visual Recognition with Missing Modality

Zhirui Kuai, Yulu Zhou, Qi Xie, Li Kuang*

• **Motivations:**
  - The prevalence of missing modalities in visual recognition tasks poses a significant challenge to model performance. Existing approaches, which often address only single-modality losses and require substantial retraining, are inadequate for environments with constrained computational resources.

• **key idea:**
  - We are aimed to enhance vision-language models’ robustness against missing modalities through a combination of multi-source data augmentation and composite prompts, enabling improved performance with limited computational resources.

• **Contributions:**
  - We present a dynamic framework that combines diverse data augmentation techniques, enabling models to select optimal augmented data autonomously during training.
  - The MACP method optimizes computational efficiency, suitable for limited-resource environments, by fine-tuning minimal model parameters.
  - Experiments on three datasets (MM-IMDb, Food101, MVSA-Single*) validate MACP’s effectiveness in handling single and multiple missing modalities.

• **Methodology:**
  • **Problem definition:**
    The dataset $\mathcal{D}_\epsilon$ is represented as a collection of multimodal samples, where each sample consists of a pair of modalities and a label.

    The parameter $\eta$ is used to control the proportion of missing modalities in the dataset. It is an adjustable value that helps simulate the scenario of missing data.

• **The Overall Framework of MACP:**

• **Multi-Source Data Augmentation:**
  Integration of multiple data augmentation techniques with a selector for tailored enhancements.

### Table 1: Performance of the ViLT model in the MM-IMDb classification task using augmented data under a condition of 50% modal data missing, with a comparative reference to its training on a complete dataset (0% missing).

<table>
<thead>
<tr>
<th>Missing rate</th>
<th>ViLT</th>
<th>ViLT-T</th>
<th>ViLT-P</th>
</tr>
</thead>
<tbody>
<tr>
<td>0% (Complete)</td>
<td>94.13</td>
<td>94.94</td>
<td>94.95</td>
</tr>
<tr>
<td>30% (Image)</td>
<td>83.11</td>
<td>83.67</td>
<td>83.67</td>
</tr>
<tr>
<td>30% (Text)</td>
<td>48.65</td>
<td>54.07</td>
<td>54.07</td>
</tr>
</tbody>
</table>

**Motivations:**
- The prevalence of missing modalities in visual recognition tasks poses a significant challenge to model performance. Existing approaches, which often address only single-modality losses and require substantial retraining, are inadequate for environments with constrained computational resources.

**key idea:**
- We are aimed to enhance vision-language models’ robustness against missing modalities through a combination of multi-source data augmentation and composite prompts, enabling improved performance with limited computational resources.

**Contributions:**
- We present a dynamic framework that combines diverse data augmentation techniques, enabling models to select optimal augmented data autonomously during training.
- The MACP method optimizes computational efficiency, suitable for limited-resource environments, by fine-tuning minimal model parameters.
- Experiments on three datasets (MM-IMDb, Food101, MVSA-Single*) validate MACP’s effectiveness in handling single and multiple missing modalities.

**Methodology:**
- **Problem definition:**
  - The dataset $\mathcal{D}_\epsilon$ is represented as a collection of multimodal samples, where each sample consists of a pair of modalities and a label.

  - The parameter $\eta$ is used to control the proportion of missing modalities in the dataset. It is an adjustable value that helps simulate the scenario of missing data.

**The Overall Framework of MACP:**

**Multi-Source Data Augmentation:**
Integration of multiple data augmentation techniques with a selector for tailored enhancements.

### Table 1: Performance of the ViLT model in the MM-IMDb classification task using augmented data under a condition of 50% modal data missing, with a comparative reference to its training on a complete dataset (0% missing).

<table>
<thead>
<tr>
<th>Missing rate</th>
<th>ViLT</th>
<th>ViLT-T</th>
<th>ViLT-P</th>
</tr>
</thead>
<tbody>
<tr>
<td>0% (Complete)</td>
<td>94.13</td>
<td>94.94</td>
<td>94.95</td>
</tr>
<tr>
<td>30% (Image)</td>
<td>83.11</td>
<td>83.67</td>
<td>83.67</td>
</tr>
<tr>
<td>30% (Text)</td>
<td>48.65</td>
<td>54.07</td>
<td>54.07</td>
</tr>
</tbody>
</table>

**Motivations:**
- The prevalence of missing modalities in visual recognition tasks poses a significant challenge to model performance. Existing approaches, which often address only single-modality losses and require substantial retraining, are inadequate for environments with constrained computational resources.

**key idea:**
- We are aimed to enhance vision-language models’ robustness against missing modalities through a combination of multi-source data augmentation and composite prompts, enabling improved performance with limited computational resources.

**Contributions:**
- We present a dynamic framework that combines diverse data augmentation techniques, enabling models to select optimal augmented data autonomously during training.
- The MACP method optimizes computational efficiency, suitable for limited-resource environments, by fine-tuning minimal model parameters.
- Experiments on three datasets (MM-IMDb, Food101, MVSA-Single*) validate MACP’s effectiveness in handling single and multiple missing modalities.

**Methodology:**
- **Problem definition:**
  - The dataset $\mathcal{D}_\epsilon$ is represented as a collection of multimodal samples, where each sample consists of a pair of modalities and a label.

  - The parameter $\eta$ is used to control the proportion of missing modalities in the dataset. It is an adjustable value that helps simulate the scenario of missing data.

**The Overall Framework of MACP:**

**Multi-Source Data Augmentation:**
Integration of multiple data augmentation techniques with a selector for tailored enhancements.

### Table 1: Performance of the ViLT model in the MM-IMDb classification task using augmented data under a condition of 50% modal data missing, with a comparative reference to its training on a complete dataset (0% missing).

<table>
<thead>
<tr>
<th>Missing rate</th>
<th>ViLT</th>
<th>ViLT-T</th>
<th>ViLT-P</th>
</tr>
</thead>
<tbody>
<tr>
<td>0% (Complete)</td>
<td>94.13</td>
<td>94.94</td>
<td>94.95</td>
</tr>
<tr>
<td>30% (Image)</td>
<td>83.11</td>
<td>83.67</td>
<td>83.67</td>
</tr>
<tr>
<td>30% (Text)</td>
<td>48.65</td>
<td>54.07</td>
<td>54.07</td>
</tr>
</tbody>
</table>

**Motivations:**
- The prevalence of missing modalities in visual recognition tasks poses a significant challenge to model performance. Existing approaches, which often address only single-modality losses and require substantial retraining, are inadequate for environments with constrained computational resources.

**key idea:**
- We are aimed to enhance vision-language models’ robustness against missing modalities through a combination of multi-source data augmentation and composite prompts, enabling improved performance with limited computational resources.

**Contributions:**
- We present a dynamic framework that combines diverse data augmentation techniques, enabling models to select optimal augmented data autonomously during training.
- The MACP method optimizes computational efficiency, suitable for limited-resource environments, by fine-tuning minimal model parameters.
- Experiments on three datasets (MM-IMDb, Food101, MVSA-Single*) validate MACP’s effectiveness in handling single and multiple missing modalities.

**Methodology:**
- **Problem definition:**
  - The dataset $\mathcal{D}_\epsilon$ is represented as a collection of multimodal samples, where each sample consists of a pair of modalities and a label.

  - The parameter $\eta$ is used to control the proportion of missing modalities in the dataset. It is an adjustable value that helps simulate the scenario of missing data.

**The Overall Framework of MACP:**

**Multi-Source Data Augmentation:**
Integration of multiple data augmentation techniques with a selector for tailored enhancements.