DOC2PPT: Automatic Slide Deck Generation from Documents

AAAI’22

Tsu-Jui Fu  William Wang  Daniel McDuff  Yale Song
• Generate a **slide** from an academic **paper**
DOC2PPT

- Multi-modal summarizer
- Text Summarization + Figure Retrieval + Multi-Page

Text Summarization

Figure Retrieval

Retrieval Models
- Two major types of dialogue model:
  - In the retrieval model, the three modalities are fed into a combiner module.
  - Reset IS2: reset denserets.
  - Dialogue decoder: dialogue decoder the encoding from the image
  - Style encoder to obtain its representation is.

Multi-Page
Dataset Building

- Crawl **paper-slide pairs** from AI conferences
  - Computer Vision (CVPR, ECCV, ...)
  - Natural Language Processing (ACL, NAACL, ...)
  - Machine Learning (ICLR, ICML, ...)

- **5,873** in total
  - 4,686 / 592 / 595 (train / val / test)

- To prepare the data for training, needs some **preprocessing** in advance
Dataset Building

- Extract **text content** from a slide
  - Azure CV to do **Optical Character Recognition (OCR)**

- Learning Over-Parametrized –Neural Networks on Structured Data
- Yingyu Liang@UWLMadison
- Joint work with Yuanzhi Li@Princeton -Y Stanford

- Our Work
  - Is there a simple theoretical explanation?
  - Our work: Yes for two-layer NN on clustered data!
  - Poster: Tue Poster Session A #143
Dataset Building

- **Match sentences** from slide to paper
- **Extractive-based summarization**

```
sentence 1
sentence 2
sentence 3
sentence 4
sentence 5
sentence 6
sentence 7
sentence 8
```
Dataset Building

- **Match sentences** from slide to paper
- **Extractive-based summarization**

---

**Paper**

- “The Pima Indians Diabetes data set contains information about 768 diabetes patients, recording features like glucose, blood pressure, age and skin thickness.”
- “Finally, can the idea of proportionality as a group fairness concept be adapted for supervised learning tasks like classification and regression?”
- “Can fairness as proportionality be adapted for supervised”

**Slide**

- “This data set contains 768 diabetes patients, recording features like glucose, blood”

---

**Extractive-Summarization**

Sentence 1
Sentence 2
Sentence 3
Sentence 4
Sentence 5
Sentence 6
Sentence 7
Sentence 8
Dataset Building

- **Match figures** from slide to paper
- **CNN feature** to do similarity matching
Dataset Building

- **Match figures** from slide to paper

- **Not** always perfect (currently 50.5% F1)
  - Leave as future work for better label to learn from

<table>
<thead>
<tr>
<th>Method</th>
<th>NYT Precision</th>
<th>NYT Recall</th>
<th>NYT F1</th>
<th>WebNLG Precision</th>
<th>WebNLG Recall</th>
<th>WebNLG F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>NovelTagging</td>
<td>62.4%</td>
<td>31.7%</td>
<td>42.0%</td>
<td>52.5%</td>
<td>19.3%</td>
<td>28.3%</td>
</tr>
<tr>
<td>OneDecoder</td>
<td>59.4%</td>
<td>53.1%</td>
<td>56.0%</td>
<td>32.2%</td>
<td>28.9%</td>
<td>30.5%</td>
</tr>
<tr>
<td>MultiDecoder</td>
<td>61.0%</td>
<td>56.6%</td>
<td>58.7%</td>
<td>37.7%</td>
<td>36.4%</td>
<td>37.1%</td>
</tr>
<tr>
<td>GraphRel1p</td>
<td>62.9%</td>
<td>57.3%</td>
<td>60.0%</td>
<td>42.3%</td>
<td>39.2%</td>
<td>40.7%</td>
</tr>
<tr>
<td>GraphRel2p</td>
<td><strong>63.9%</strong></td>
<td><strong>60.0%</strong></td>
<td><strong>61.9%</strong></td>
<td><strong>44.7%</strong></td>
<td><strong>41.1%</strong></td>
<td><strong>42.9%</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>NovelTag</td>
<td>62.4%</td>
<td>31.7%</td>
<td>42.0%</td>
<td>28.3%</td>
</tr>
<tr>
<td>CopyRE</td>
<td>61.0%</td>
<td>56.6%</td>
<td>58.7%</td>
<td>-</td>
</tr>
<tr>
<td>GraphRel1p</td>
<td>62.9%</td>
<td>57.3%</td>
<td>60.0%</td>
<td>88.8%</td>
</tr>
<tr>
<td>GraphRel2p</td>
<td>63.9%</td>
<td>60.0%</td>
<td>61.9%</td>
<td>89.2%</td>
</tr>
</tbody>
</table>
Dataset Building

- **Match figures** from slide to paper

- **Not** always perfect (currently 50.5% F1)

- Apply **human labeling** for testing set
  - **Golden testing set** for fair evaluation
Dataset Building

- Remove the **progressive** page
- **OCR cover rate** > 80% (Acc ~90%)
- Keep the last one
Dataset Building

• Generate pages **for each section** and combine them all
• BERT to match **text** (page) with **paragraph** (section)
• Consider **continuity**
Dataset Building

• Generate pages **for each section** and combine them all
• BERT to match **text** (page) with **paragraph** (section)
Dataset Building
Dataset Building

Section Matching

Paper

Figure

Text

OCR

Slide

Text

Figure

Sentence Matching

Figure Matching

Progressive Removing
## Dataset Building

<table>
<thead>
<tr>
<th></th>
<th>num</th>
<th>#section</th>
<th>#sentence (per section)</th>
<th>#figure</th>
<th>#page</th>
<th>#sentence (per section)</th>
<th>#figure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>4,686</td>
<td>6.9</td>
<td>42.9</td>
<td>8.3</td>
<td>16.9</td>
<td>8.1</td>
<td>2.4</td>
</tr>
<tr>
<td>Val</td>
<td>592</td>
<td>6.9</td>
<td>42.6</td>
<td>8.3</td>
<td>16.8</td>
<td>8.1</td>
<td>2.5</td>
</tr>
<tr>
<td>Test</td>
<td>595</td>
<td>6.9</td>
<td>42.4</td>
<td>8.4</td>
<td>16.5</td>
<td>8.1</td>
<td>2.6</td>
</tr>
<tr>
<td>Test (Human)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.3</td>
</tr>
</tbody>
</table>
Dataset Building

- Distribution of **sentence** and **figure** in slide
- Similar between train, val, and test
Model (Baseline)

- Recurrent extractor to build the slide step-by-step
  - \textbf{[OBJ]}, \textbf{[PAGE]}, \textbf{[SECTION]} token
  - \textbf{Section-based} generation and \textbf{classification} for extraction
Model (Baseline)

- Recurrent extractor to build the slide step-by-step
  - \texttt{[OBJ]}, \texttt{[PAGE]}, \texttt{[SECTION]} token
  - \textbf{Section-based} generation and \textbf{classification} for extraction
Model (Baseline)

- Recurrent extractor to build the slide step-by-step
- \([\text{OBJ}], [\text{PAGE}], [\text{SECTION}]\) token
- **Section-based** generation and **classification** for extraction

![Diagram of Model (Baseline)]
Model (Baseline)

- Recurrent extractor to build the slide step-by-step
  - \([\text{OBJ}], [\text{PAGE}], [\text{SECTION}]\) token
  - **Section-based** generation and **classification** for extraction
Model (Baseline)

- Recurrent extractor to build the slide step-by-step
  - [OBJ], [PAGE], [SECTION] token
  - Section-based generation and classification for extraction
Model (Baseline)

- Recurrent extractor to build the slide step-by-step
- **[OBJ]**, **[PAGE]**, **[SECTION]** token
- **Section-based** generation and **classification** for extraction
Model (Baseline)

• Recurrent extractor to build the slide step-by-step
  • [OBJ], [PAGE], [SECTION] token
  • Section-based generation and classification for extraction
Model (Baseline)

- Recurrent extractor to build the slide step-by-step
- \([OBJ], [PAGE], [SECTION]\) token
- **Section-based** generation and **classification** for extraction
Model (HSE)

- Hierarchical Slide Extractor (HSE)
- Different RNNs for section-, page-, and object-level
TextFigure Module

- Constrain the **coherence** between figure-text
- Co-train with HSE
- Related figure-text should be **close on embedding space**

“The **learning framework** of our adversarial path sampler (APS), where Speaker denotes the back-translated speaker model.”

“**R2R results** for Seq2Seq, Speaker-Follower, and RCM under testing set.”
TextFigure Module

- Right figures put with right texts
  - Filter out unrelated and add unused related figures

Result

- Randomly sampled stop improving when using more than 60%
- APS sampled helps both seen and unseen
- Pre-Exploration further helps unseen environments
Paraphrasing Module

- Rewrite extracted sentences as **slide-style**
- Seq2seq model (w/ copy attention)

"to understand the spread of individual judgements on a sentence, we compute the standard deviation of **ratings for each sentence** and then **take the mean** over all sentences."

"we perform **empirical evaluation** and analysis of a variety of **classification methods** for the above task."

"we collect multiple ratings for a sentence and take the mean."

"empirical evaluation of classification methods"
3. Hierarchical Vector Quantization

Several optimization techniques have been employed to speed up deformable contour models. A popular approach is the algorithm proposed by Borgefors and Forss 2]. This is several orders of magnitude faster than the original implementation of [2]. The key to their success is vector quantization techniques. The computational demand is much lower. They vector quantize HOG features and compute angular scores by looking up vector tables and adding a weighted score.

We use vector quantization for the same purpose but with a slightly different approach. The main computational bottleneck is [2] by vector quantization. They found that vectorizing the vectors (HOG features of an image) the HOG color descriptors for all regions of the contour (Figure 3, a). We use a hierarchical clustering technique to speed up this process. We first create each cell into 16 clusters (Figure 2, a). Then, according to the nearest cluster is the first step we compare against its other clusters to find the nearest cluster (Figure 3, b). We pre-compute clusters using a nearest algorithm.

Our experience shows that the proposed hierarchical clustering technique works best on a cell size of 0.011 in size. In contrast, the opening can be significant.
3. Hierarchical Vector Quantization

Several optimization techniques have been employed to speed up Delaunay triangulation. The fastest one proposed by Delaunay and Fortun [2] is to match two centers of nontangential facets of the Delaunay triangulation. This is much faster than the original implementation of Delaunay triangulation. The fast is based on a technique that reduces the computation done by a large factor. They introduce Delaunay triangulation and compute triangle areas by using euclidean distance and angle between vectors.

We use vector quantization for the same purpose but with a slightly different approach. The node computation fast is to vector quantization. They based their technique on previous work by Fowlkes [1]. They take advantage of the fact that the Delaunay triangulation is the dual of the Voronoi diagram, which is easier to compute. The Delaunay triangulation can be computed in linear time using a divide and conquer approach. We use a hierarchical clustering technique to speed up this process. We first cluster each cell into its nearest neighbors, and then we cluster the resulting clusters together. This process is repeated until the desired level of clustering is reached. We use a hierarchical clustering technique, which is a divide and conquer approach.
Hierarchical Vector Quantization

- We use a hierarchical clustering technique to speed up the process.
- We use vector quantization for each node purpose
  but with a slightly different approach.

- Then according to the nearest cluster to the first step we compare against
  64 other clusters to find the nearest cluster (Figure 3, 3).

- We pre-compute clusters using k-means algorithm.

- Our experiments show that the proposed hierarchical clustering technique
  leads to a negligible loss of 0.001 in mAP. In contrast, the opening gap is
  quite large.
HSE w/ TextFigure & Paraphrasing

3. Hierarchical Vector Quantization

Several optimization techniques have been employed to speed up Dithermark, including several heuristic techniques. The fastest implementation is achieved by using the following approach:

We use vector quantization for the same purpose but with a slightly different approach. We use hierarchical clustering to speed up the process. We first cluster each cell into 4 clusters (Figure 3, a). Then, according to the nearest cluster in the first step, we assign clusters to other clusters to yield the nearest clusters (Figure 3, b). We pre-compute clusters using k-means algorithm.

Our experiments show that the proposed hierarchical clustering technique leads to a negligible loss of 0.05% in mAP. In contrast, the standard approach leads to a negligible loss of 0.05% in mAP.
Experiments

• Evaluation metrics

Text

- the cat is sleeping on bed
- the brown cat is sitting on bed

Rouge-L: 83.3 / 71.4 / 76.9
Experiments

• Evaluation metrics

Text

the cat is sleeping on bed

the brown cat is sitting on bed

Rouge-L: 83.3 / 71.4 / 76.9

\[
\text{Rouge} \times e^{-\frac{|P-Q|}{Q}}
\]

• consider Page Difference
• P: \#Page_{pd}
• Q: \#Page_{gd}
Experiments

- Evaluation metrics

**Text**

- the cat is sleeping on bed
- the brown cat is sitting on bed

Rouge-L: 83.3 / 71.4 / 76.9

**Figure**

- A / D / C / F / E
- A / F / B / E

LC-P/R/F: 60.0 / 75.0 / 66.7

$$\text{Rouge} \times e^{-\frac{|P-Q|}{Q}}$$

- consider Page difference
- P: \#Page_{pd}
- Q: \#Page_{gd}
Experiments

Evaluation metrics

- **Text**
  - the cat is sleeping on bed
  - the brown cat is sitting on bed

Rouge-L: 83.3 / 71.4 / 76.9

- **Figure**
  - A / D / C / F / E
  - A / F / B / E

LC-P/R/F: 60.0 / 75.0 / 66.7

- **TextFigure**
  - the cat is sleeping on bed
  - the brown cat is sitting on bed

Rouge-L

- consider Page Difference
- P: #Page\(_{pd}\)
- Q: #Page\(_{gd}\)
## Experiments

<table>
<thead>
<tr>
<th>Model</th>
<th>Co-Train</th>
<th>w/ Module</th>
<th>Text</th>
<th>Figure</th>
<th>TextFigure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TextFigure</td>
<td>Paraphrase</td>
<td>TextFigure</td>
<td>Rouge-L</td>
<td>w/ Page</td>
</tr>
<tr>
<td>Baseline</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>27.2</td>
<td>21.8</td>
</tr>
<tr>
<td></td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>27.7</td>
<td>22.9</td>
</tr>
<tr>
<td></td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
<td>32.3</td>
<td>26.7</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>28.7</td>
<td>24.0</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>28.7</td>
<td>24.0</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>33.6</td>
<td>28.2</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>33.6</td>
<td>28.2</td>
</tr>
</tbody>
</table>

1st / 2nd
## Experiments

<table>
<thead>
<tr>
<th>Model</th>
<th>Co-Train</th>
<th>w/ Module</th>
<th>TextFigure</th>
<th>Paraphrase</th>
<th>TextFigure</th>
<th>Rouge-L</th>
<th>w/ Page</th>
<th>LC-P</th>
<th>LC-R</th>
<th>LC-F1</th>
<th>Rouge-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>27.2</td>
<td>21.8</td>
<td>13.2</td>
<td>21.9</td>
<td>16.5</td>
<td>3.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>27.7</td>
<td>22.9</td>
<td>14.6</td>
<td>23.7</td>
<td>18.1</td>
<td>4.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>32.3</td>
<td>26.7</td>
<td>14.6</td>
<td>23.7</td>
<td>18.1</td>
<td>4.7</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>28.7</td>
<td>24.0</td>
<td>14.8</td>
<td>32.4</td>
<td>20.3</td>
<td>7.9</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>28.7</td>
<td>24.0</td>
<td>24.6</td>
<td>40.5</td>
<td>30.6</td>
<td>13.8</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>33.6</td>
<td>28.2</td>
<td>14.8</td>
<td>32.4</td>
<td>20.3</td>
<td>8.2</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>33.6</td>
<td>28.2</td>
<td>24.6</td>
<td>40.5</td>
<td>30.6</td>
<td>15.5</td>
</tr>
</tbody>
</table>

- **Hierarchical architecture** extracts slide
- Helps both **text quality** and **figure retrieval**
<table>
<thead>
<tr>
<th>Model</th>
<th>Co-Train</th>
<th>w/ Module</th>
<th>Text</th>
<th>Figure</th>
<th>TextFigure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Rouge-L w/ Page</td>
<td>LC-P</td>
<td>LC-R</td>
</tr>
<tr>
<td>Baseline</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>27.2</td>
<td>21.8</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>✓</td>
<td>X</td>
<td>27.7</td>
<td>22.9</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>X</td>
<td>32.3</td>
<td>26.7</td>
</tr>
<tr>
<td>HSE</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>28.7</td>
<td>24.0</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>28.7</td>
<td>24.0</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>33.6</td>
<td>28.2</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>33.6</td>
<td>28.2</td>
</tr>
</tbody>
</table>

- **Paraphrasing module** rewrites sentences into slide-style
- Better **text** as a slide
## Experiments

<table>
<thead>
<tr>
<th>Model</th>
<th>Co-Train</th>
<th>w/ Module</th>
<th>Text</th>
<th>Figure</th>
<th>TextFigure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Rouge-L w/ Page</td>
<td>LC-P</td>
<td>LC-R</td>
</tr>
<tr>
<td>Baseline</td>
<td>X</td>
<td>X</td>
<td>27.2</td>
<td>13.2</td>
<td>21.9</td>
</tr>
<tr>
<td>HSE</td>
<td>✓</td>
<td>✓</td>
<td>27.7</td>
<td>14.6</td>
<td>23.7</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>32.3</td>
<td>14.6</td>
<td>23.7</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>28.7</td>
<td>14.8</td>
<td>32.4</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>28.7</td>
<td>24.6</td>
<td>40.5</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>33.6</td>
<td>14.8</td>
<td>32.4</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>33.6</td>
<td>24.6</td>
<td>40.5</td>
</tr>
</tbody>
</table>

- Co-train with **TextFigure constrain**
- Learns the **correlation** between text and figure
## Experiments

<table>
<thead>
<tr>
<th>Model</th>
<th>Co-Train</th>
<th>w/ Module</th>
<th>Text</th>
<th>Figure</th>
<th>TextFigure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TextFigure</td>
<td>Paraphrase</td>
<td>TextFigure</td>
<td>Rouge-L</td>
<td>w/ Page</td>
</tr>
<tr>
<td>Baseline</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>27.2</td>
<td>21.8</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>27.7</td>
<td>22.9</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>✓</td>
<td>X</td>
<td>32.3</td>
<td>26.7</td>
</tr>
<tr>
<td>HSE</td>
<td>✓</td>
<td>X</td>
<td>X</td>
<td>28.7</td>
<td>24.0</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>X</td>
<td>✓</td>
<td>28.7</td>
<td>24.0</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>33.6</td>
<td>28.2</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>33.6</td>
<td>28.2</td>
</tr>
</tbody>
</table>

- **TextFigure module** removes unrelated or adds related
- Benefits **figure retrieval** a lot
## Experiments

<table>
<thead>
<tr>
<th>Model</th>
<th>Co-Train</th>
<th>w/ Module</th>
<th>Text</th>
<th>Figure</th>
<th>TextFigure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>TextFigure</td>
<td>Paraphrase</td>
<td>TextFigure</td>
</tr>
<tr>
<td>Baseline</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>27.2</td>
<td>21.8</td>
</tr>
<tr>
<td>HSE</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>27.7</td>
<td>22.9</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>✓</td>
<td>X</td>
<td>32.3</td>
<td>26.7</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>X</td>
<td>X</td>
<td>28.7</td>
<td>24.0</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>X</td>
<td>✓</td>
<td>28.7</td>
<td>24.0</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>X</td>
<td>33.6</td>
<td>28.2</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>33.6</td>
<td>28.2</td>
</tr>
</tbody>
</table>

- Combines both **Paraphrasing** and **TextFigure** module
- Fully improves on **all aspects of metrics**
Qualitative Examples

Introduction
- We propose a novel multi-modal conditional alignment methodology to bridge domain divergence while preserving the discriminability of the features.
- Idea: multi-modal conditional distribution alignment and detection/regularization model

Category Prediction based Regularization
- What is the structure of the graph?
- We propose a novel category prediction mechanism for object detection.
- Each proposal will be classified in a regression.
- Region proposal network (RPN) 23-30 August 2019,
- Loss function: l + l_

Adaptation from Clear to Foggy Scenes.
- Cross-domain detection from real to virtual image scenarios.
- Domain adaptation from normal / clear images to foggy image.
- Recall, precision, no semantic.

Ablation Study
- Ablation: adaptive feature visualizations with mutual regularization. - Zhao et al.
- We use the foggy chameleon dataset as the target domain.
- Focus on bilateral data in the target domain.
- Multiple auxiliary loss terms in the proposed learning objective.

Introduction
- What is a good emotion classification task?
- We use the context principle for emotion recognition.
- Context 1: incorporating cues from different modalities.
- Multimodal emotion recognition (Apr 2020)
- Not asking for the meaning of a word in isolation and instead of finding the meaning in isolation.

Network Architecture
- How to train your neural network?
- To make the soft margin loss function.
- We combine the loss function, multiplicative (from eq. 1).
- ∑ i (x_i).

Datasets
- We present a comparison with other datasets.
- The apparent emotional states of the people.
- How do we evaluate the abstraction process?
- How do we evaluate the friendliness?

SELF-ADVERSARIAL LEARNING
- For a training set with a real sample, we have
- Self (x_i, y_i) = red
- How to suffer from the reward sparsity?
- Self (x_i) = blue

TRAINING
- The comparative discriminator can offer more informative learning signals from the comparative discriminator.
- How to enhance the generalization ability of the comparative discriminator?
- C (u_i, v_i)

COMPARATIVE DISCRIMINATOR
- The self-improvement mechanism corresponds to the comparative discriminator.
- How to construct the model to supervise the model?
- Guo et al., 2016

RESULTS IN REAL DATA
- Table 3. The results of coco emotion classification.
Qualitative Examples

- **TextFigure Module (w/o vs w/)**
Qualitative Examples

- Paraphrasing Module (w/o vs w/)

Introduction
- Since the phrase representations are produced and attended at each encoder layer, the encoding of each layer is also enhanced with phrase-level attention computation.

Base Architecture
- The feed-forward layers capture the domain-specific and -independent information by using pure output layers for each domain and one shared output layer.
- Word embeddings are derived from a combination of GloVe (Pennington et al., 2014) and fastText (Bojanowski et al., 2017) pre-trained word embeddings, as used in (Ma and Hovy, 2016).
- The global objective function is the combination of the non-loss and domain loss.
- The domain classification objective is to minimize the cross-entropy loss (underlla, 2018) for an input with domain label.
- We propose a new architecture based on the BERT/MAVERIK model utilized in the three previous experimental setups.

Related work
- Action proposals is an essential part of many methods for action detection, realized by a number of recent.
- More related to our work, previous methods [9, 10, 14, 15, 16] expect the temporal order, either by predicting the exact order of consecutive frames [9, 10] or verifying the partial order [14, 15].
- In the video domain, motion has been used as a cue for learning video representations in [5, 6, 7, 8].
- The notice of actions was first introduced in [9] as a confidence measure of intentional bodily movement of biological agents.

Conclusions
- We proposed two parameterized benchmark games in which ETCE exhibits interesting behaviors.
- We also provided an alternative saddle-point formulation of ETCE and demonstrated its merit with a simple subgradient method which surpasses standard LP-based methods.
- We analyzed those behaviors both qualitatively and quantitatively, and isolated two ways through which a mediator is able to control the agents to follow the recommendations.
- We hope that our analysis will bring attention to some of the computational and practical uses of ETCE, and that our benchmark games will be useful for evaluating future algorithms for computing ETCE in large games.

Related work
- Action detection is a key tool for action detection.
- Predicting the exact order of consecutive frames [9, 10] or verifying the partial order [14, 15].
- The notice of actions was first introduced in [9] as a confidence measure of intentional bodily movement of biological agents.
Qualitative Examples

• Applying Design Ideas

Introduction
- Weak supervision in text classification has the burden of human experts
- How to train a deep neural network?
- We have performed experiments on real-world datasets
- Identify words that are discriminative and highly label-inducive

Face photo to drawing generator G
- No need for existing between human domains
- Requires inverse generator to reconstruct a face photo
- A strict loss function for cycle-consistency loss
- P (p. 3)

Multi-task Learning with Self-supervision
- Depending on the type of training samples, the statistical characteristics of the augmented training samples
- Remove unnecessary invariant property of the classifier
- Aggregate the corresponding conditional probabilities to improve the classification accuracy

Language Model Baselines
- Feeding the tokens in the input sequence
- We can obtain the contextualized embeddings

Detection Results on CityPersons
- Malicious mix of PBM
- Breaking the curse of many agents with events

Multi-task Learning with Self-supervision
- Depending on the type of training samples, the statistical characteristics of the augmented training samples
- Remove unnecessary invariant property of the classifier
- Aggregate the corresponding conditional probabilities to improve the classification accuracy

FAD Frequency-Aware Decomposition
- The frequencies of frequency-aware components can be inversely transformed
- Number of filters (N=152)
- The frequency filtering is a special case of the input image
Conclusion

• DOC2PPT serves as a **multi-modal summarizer** to generate slide from academic documents

• We propose **hierarchical architecture, text-figure constrain**, and **paraphrasing module** to improve the quality of slide generation

• DOC2PPT **provides useful outline and flow** to make building a slide more efficiency
thanks!