Medical Image Report Generation and Beyond

Zhiting Hu, Pengtao Xie, Xiaodan Liang, and Eric Xing

February 25, 2019
A Tsunami of Healthcare Data

Volume

- 153 exabytes (one exabyte = one billion gigabytes) were produced in 2013
- An estimated 2,314 exabytes will be produced in 2020

Complexity

- Notes
- Image
- Lab values
- Vital signs
- Test
- Genomics
- Billing
- Literature
- Social media

153 exabytes (one exabyte = one billion gigabytes) were produced in 2013.

An estimated 2,314 exabytes will be produced in 2020.
Machine Learning for Healthcare

Clinical Data

Note

Image

Lab values

Vital signs

Test

Genomics

Billing

Literature

Social

Healthcare Insurance

PubMed

PatientsLikeMe
Machine Learning for Healthcare

Clinical Data

Machine Learning

Components

Patterns
Machine Learning for Healthcare

Clinical Data

- Note
- Image
- Lab values
- Vital signs
- Test
- Genomics
- Billing
- Literature
- Social

Machine Learning

Components

Patterns

Actionable Insights

Extracted information

Recommended diagnosis and treatment

Suggested ICD codes

- Detected lung nodule
- Predicted mortality rate
- Detected arrhythmia
Medical Imaging and Report

- Medical imaging is widely used in clinical practice
Medical Imaging and Report

- Medical imaging is widely used in clinical practice
- Specialized physicians read medical images and write text reports
Medical Imaging and Report

- Medical imaging is widely used in clinical practice
- Specialized physicians read medical images and write text reports

Knowledge demanding:
- 1) normal anatomy of, e.g., thorax, basic physiology of chest diseases;
- 2) analyzing radio graph;
- 3) evaluating evolution;
- 4) correlation with other diagnostic results; …
Medical Imaging and Report

- Medical imaging is widely used in clinical practice
- Specialized physicians read medical images and write text reports

Knowledge demanding:
- 1) normal anatomy of, e.g., thorax, basic physiology of chest diseases;
- 2) analyzing radio graph; 3) evaluating evolution; 4) correlation with other diagnostic results; …

Time consuming:
- 5-10 mins per image
- 100s of images per day
General Image-to-Text Problems at a Glance

- Traditionally:
  - Labeling (classification to known labels)
  - Tagging (ROI, bounding-boxes, etc)
  - Simple description (one sentence ...)


General Image-to-Text Problems at a Glance

Traditionally:
- Labeling (classification to known labels)
- Tagging (ROI, bounding-boxes, etc)
- Simple description (one sentence …)

In need:
- Full textual summary
- Report
  - different image genre
  - a full image collection, not just a single image
  - videos
Outline

- Medical image report generation
  - Co-attention, hierarchical generation, multi-task
  - Further improvement: retrieval+generation, structured knowledge

- Paragraph description of natural images

- Text generation under control
  - Various text properties, granularities, amount of supervision

- All in one toolkit: Texar
Outline

- Medical image report generation
  - Co-attention, hierarchical generation, multi-task
  - Further improvement: retrieval+generation, structured knowledge

- Paragraph description of natural images

- Text generation under control
  - Various text properties, granularities, amount of supervision

- All in one toolkit: Texar
Chest X-ray Report

Findings:
There are no focal areas of consolidation.
No suspicious pulmonary opacities.
Heart size within normal limits.
No pleural effusions.
There is no evidence of pneumothorax.
Degenerative changes of the thoracic spine.

Impression:
No acute cardiopulmonary abnormality.
Chest X-ray Report

- A paragraph consists of
  - **Findings:** radiology observations regarding the body area examined
  - **Impression:** most prominent observation or conclusion

**Findings:**
There are no focal areas of consolidation.  
No suspicious pulmonary opacities.  
Heart size within normal limits.  
No pleural effusions.  
There is no evidence of pneumothorax.  
Degenerative changes of the thoracic spine.

**Impression:**
No acute cardiopulmonary abnormality.
Solution Overview

Key challenges

Findings:
There are no focal areas of consolidation.
No suspicious pulmonary opacities.
Heart size within normal limits.
No pleural effusions.
There is no evidence of pneumothorax.
Degenerative changes of the thoracic spine.

Impression:
No acute cardiopulmonary abnormality.

[Jing et al., 2018]
Solution Overview

Key challenges

- Identify abnormal regions (lesions)

---

Findings:
There are no focal areas of consolidation.
No suspicious pulmonary opacities.
Heart size within normal limits.
No pleural effusions.
There is no evidence of pneumothorax.
Degenerative changes of the thoracic spine.

Impression:
No acute cardiopulmonary abnormality.

[Jing et al., 2018]
Solution Overview

Key challenges

- Identify abnormal regions (lesions)
- Reports are usually lengthy paragraphs

Findings:
There are no focal areas of consolidation.
No suspicious pulmonary opacities.
Heart size within normal limits.
No pleural effusions.
There is no evidence of pneumothorax.
Degenerative changes of the thoracic spine.

Impression:
No acute cardiopulmonary abnormality.

[Jing et al., 2018]
Solution Overview

Key challenges

- Identify abnormal regions (lesions)
- Reports are usually lengthy paragraphs
- Each sentence has visual evidence

[Jing et al., 2018]
Solution Overview

Key challenges
- Identify abnormal regions (lesions)
- Reports are usually lengthy paragraphs
- Each sentence has visual evidence

Findings:
There are no focal areas of consolidation.
No suspicious pulmonary opacities.
Heart size within normal limits.
No pleural effusions.
There is no evidence of pneumothorax.
Degenerative changes of the thoracic spine.

Impression:
No acute cardiopulmonary abnormality.

[Jing et al., 2018]
Solution Overview

Key challenges
- Identify abnormal regions (lesions)
- Reports are usually lengthy paragraphs
- Each sentence has visual evidence

Solutions
<--- Multi-task tag classification to inform lesions

Findings:
- There are no focal areas of consolidation.
- No suspicious pulmonary opacities.
- Heart size within normal limits.
- No pleural effusions.
- There is no evidence of pneumothorax.
- Degenerative changes of the thoracic spine.

Impression:
- No acute cardiopulmonary abnormality.

MTI Tags: degenerative change

[Jing et al., 2018]
Solution Overview

Key challenges

- Identify abnormal regions (lesions)
- Reports are usually lengthy paragraphs
- Each sentence has visual evidence

Solutions

--- Multi-task tag classification to inform lesions

--- Hierarchical text generation: sentence/word

Findings:
There are no focal areas of consolidation.
No suspicious pulmonary opacities.
Heart size within normal limits.
No pleural effusions.
There is no evidence of pneumothorax.
Degenerative changes of the thoracic spine.

Impression:
No acute cardiopulmonary abnormality.

**MTI Tags:** degenerative change

[Jing et al., 2018]
Solution Overview

Key challenges

- Identify abnormal regions (lesions)
- Reports are usually lengthy paragraphs
- Each sentence has visual evidence

Solutions

- Multi-task tag classification to inform lesions
- Hierarchical text generation: sentence/word
- Visual-semantic co-attention

Findings:
There are no focal areas of consolidation.
No suspicious pulmonary opacities.
Heart size within normal limits.
No pleural effusions.
There is no evidence of pneumothorax.
Degenerative changes of the thoracic spine.

Impression:
No acute cardiopulmonary abnormality.

MTI Tags: degenerative change

[Jing et al., 2018]
Model Architecture
(I) Lesion tag classification
Model Architecture

(I) Lesion tag classification

(II) Hierarchical text generation
Model Architecture

(I) Lesion tag classification

(II) Hierarchical text generation

(III) Visual-semantic co-attention
Lesion Tag Classification

Large amount of (image, tags) data available for training.
(II) Hierarchical Text Generation

**Predicted Tags**
- Hyperdistention
- Emphysema
- Cicatrix
(II) Hierarchical Text Generation

Sentence-level LSTM generates sentence-level hidden vectors

Predicted Tags
- Hyperdistention
- Emphysema
- Cicatrix
(II) Hierarchical Text Generation

- **Predicted Tags**
  - Hyperdistention
  - Emphysema
  - Cicatrix

- **Sentence-level LSTM**
  generates sentence-level hidden vectors

- **Visual-semantic co-attention module**
  calculates contextual vectors

![Diagram](image-url)
(II) Hierarchical Text Generation

Predicted Tags
- Hyperdistention
- Emphysema
- Cicatrix

Sentence-level LSTM generates sentence-level hidden vectors

Visual-semantic co-attention module calculates contextual vectors

Topic layer calculates sentence topic vectors
(II) Hierarchical Text Generation

Predicted Tags
- Hyperdistention
- Emphysema
- Cicatrix

Sentence-level LSTM generates sentence-level hidden vectors

Visual-semantic co-attention module calculates contextual vectors

Topic layer calculates sentence topic vectors

Word-level LSTM generates word-level hidden vectors
(II) Hierarchical Text Generation

Predicted Tags
• Hyperdistention
• Emphysema
• Cicatrix

Sentence-level LSTM generates sentence-level hidden vectors

Visual-semantic co-attention module calculates contextual vectors

Topic layer calculates sentence topic vectors

Word-level LSTM generates word-level hidden vectors

Generated words
There is chronic pleural-parenchymal scarring within the lung bases. No lobar consolidation is seen. ...
(III) Visual-Semantic Co-Attention

… There is chronic pleural-parenchymal scarring within the lung bases. No lobar consolidation is seen. …

Predicted Tags
• Hyperdistention
• Emphysema
• Cicatrix
Experiments

- **Image-report data**
  - Indiana University Chest X-ray Collection (IU X-Ray)
  - 7,470 image-report pairs
  - 5.7 sentences/image
Experiments

- **Image-report data**
  - Indiana University Chest X-ray Collection (IU X-Ray)
  - 7,470 image-report pairs
  - 5.7 sentences/image

- **Image-tag data**
  - NIH Chest X-ray images
  - 108,948 image-tags pairs
  - 14 lesion tags
## Evaluation of Report Quality

<table>
<thead>
<tr>
<th>Baselines</th>
<th>Methods</th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
<th>METEOR</th>
<th>ROUGE</th>
<th>CIDER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CNN-RNN [1]</td>
<td>0.316</td>
<td>0.211</td>
<td>0.140</td>
<td>0.095</td>
<td>0.159</td>
<td>0.267</td>
<td>0.111</td>
</tr>
<tr>
<td></td>
<td>LRCN [2]</td>
<td>0.369</td>
<td>0.229</td>
<td>0.149</td>
<td>0.099</td>
<td>0.155</td>
<td>0.278</td>
<td>0.190</td>
</tr>
<tr>
<td></td>
<td>Soft ATT [3]</td>
<td>0.399</td>
<td>0.251</td>
<td>0.168</td>
<td>0.118</td>
<td>0.167</td>
<td>0.323</td>
<td>0.302</td>
</tr>
<tr>
<td></td>
<td>ATT-RK [4]</td>
<td>0.369</td>
<td>0.226</td>
<td>0.151</td>
<td>0.108</td>
<td>0.171</td>
<td>0.323</td>
<td>0.155</td>
</tr>
<tr>
<td>Our methods</td>
<td>No-Attention</td>
<td>0.505</td>
<td>0.383</td>
<td>0.290</td>
<td>0.224</td>
<td>0.200</td>
<td>0.420</td>
<td>0.259</td>
</tr>
<tr>
<td></td>
<td>Semantic-Only</td>
<td>0.504</td>
<td>0.371</td>
<td>0.291</td>
<td>0.230</td>
<td>0.207</td>
<td>0.418</td>
<td>0.286</td>
</tr>
<tr>
<td></td>
<td>Visual-Only</td>
<td>0.507</td>
<td>0.373</td>
<td>0.297</td>
<td>0.238</td>
<td>0.211</td>
<td>0.426</td>
<td>0.300</td>
</tr>
<tr>
<td></td>
<td>Co-Attention</td>
<td>0.517</td>
<td>0.386</td>
<td>0.306</td>
<td>0.247</td>
<td>0.217</td>
<td>0.447</td>
<td>0.327</td>
</tr>
</tbody>
</table>

Comparison with the state-of-the-art image captioning methods
Evaluation of Clinical Correctness

- Presence/absence of lesions in the report
Evaluation of Clinical Correctness

- Presence/absence of lesions in the report

- Human evaluators
  - manually read the report
  - check whether a lesion exists
Evaluation of Clinical Correctness

- Presence/absence of lesions in the report
- Human evaluators
  - manually read the report
  - check whether a lesion exists

Generated report:
Normal cardiomediastinal silhouette. Interval improvement in lung volumes bilaterally. Improved aeration of the right and left lung bases. Bilateral small pleural effusions and left base atelectatic change, with interval improvement. Visualized XXXX of the chest XXXX are within normal limits.
Evaluation of Clinical Correctness

- Presence/absence of lesions in the report
- Human evaluators
  - manually read the report
  - check whether a lesion exists

Generated report:
Normal cardiomiastinal silhouette. Interval improvement in lung volumes bilaterally. Improved aeration of the right and left lung bases. Bilateral small pleural effusions and left base atelectatic change, with interval improvement. Visualized XXXX of the chest XXXX are within normal limits.

Lesions:
- Has effusion and atelectasis
- No other lesions
## Evaluation of Clinical Correctness (cont’d)

<table>
<thead>
<tr>
<th></th>
<th>No-Attention</th>
<th>Visual-Only</th>
<th>Semantic-Only</th>
<th>Co-Attention</th>
</tr>
</thead>
<tbody>
<tr>
<td>Macro-F1</td>
<td>0.49</td>
<td>0.51</td>
<td>0.75</td>
<td>0.79</td>
</tr>
</tbody>
</table>
Visualization of Co-attention

Figure 4: Visualization of co-attention for three examples. Each example is comprised of four things: (1) image and visual attentions; (2) ground truth tags and semantic attention on predicted tags; (3) generated descriptions; (4) ground truth descriptions. For the semantic attention, three tags with highest attention scores are highlighted. The underlined tags are those appearing in the ground truth.

The image at the bottom is a failure case of Ours-CoAttention. However, even though the model makes the wrong judgment about the major abnormalities in the image, it does find some unusual regions: “lateral lucency” and “left lower lobe”.

To further understand models’ ability of detecting abnormalities, we present the portion of sentences which describe the normalities and abnormalities in Table 2. We consider sentences which contain “no”, “normal”, “clear”, “stable” as sentences describing normalities. It is clear that Ours-CoAttention best approximates the ground truth distribution over normality and abnormality.

<table>
<thead>
<tr>
<th>Method</th>
<th>Normality</th>
<th>Abnormality</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soft Attention</td>
<td>0.510</td>
<td>0.490</td>
<td>1.0</td>
</tr>
<tr>
<td>Ours-no-Attention</td>
<td>0.753</td>
<td>0.247</td>
<td>1.0</td>
</tr>
<tr>
<td>Ours-CoAttention</td>
<td>0.471</td>
<td>0.529</td>
<td>1.0</td>
</tr>
<tr>
<td>Ground Truth</td>
<td>0.385</td>
<td>0.615</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table 2: Portion of sentences which describe the normalities and abnormalities in the image.

4.5.2 Co-Attention Learning

Figure 4 presents visualizations of co-attention. The first property shown by Figure 4 is that the sentence LSTM can generate different topics at different time steps since the model focuses on different image regions and tags for different sentences. The next finding is that visual attention can guide our model to concentrate on relevant regions.
Generated Examples

Ground Truth
No active disease. The heart and lungs have in the interval. Both lungs are clear and expanded. Heart and mediastinum normal.

Ours-CoAttention
No active disease. The heart and lungs have in the interval. Lungs are clear and expanded. Cardi mediastinal silhouette is within normal limits. No pleural effusion or pneumothorax is seen. No pleural effusion. No cavitary or pneumothorax.

Figure 3: Illustration of paragraph generated by Ours-CoAttention, Ours-no-Attention, and Soft Attention models. The underlined sentences are the descriptions of detected abnormalities. The second image is a lateral x-ray image. Top two images are positive results, the third one is a partial failure case and the bottom one is failure case. These images are from test dataset.

Semantic attention is inadequate of localizing small abnormal image-regions. Finally, our full model (Ours-CoAttention) achieves the best results on all of the evaluation metrics, which demonstrates the effectiveness of the proposed co-attention mechanism.

4.5 Qualitative Results
4.5.1 Paragraph Generation
An illustration of paragraph generation by three models (Ours-CoAttention, Ours-no-Attention and Soft Attention models) is shown in Figure 3.

We can find that different sentences have different topics. The first sentence is usually a high level description of the image, while each of the following sentences is associated with one area of the image (e.g. "lung", "heart"). Soft Attention and Ours-no-Attention models detect only a few abnormalities of the images and the detected abnormalities are incorrect. In contrast, Ours-CoAttention model is able to correctly describe many true abnormalities (as shown in top three images). This comparison demonstrates that co-attention is better at capturing abnormalities.

For the third image, Ours-CoAttention model successfully detects the area ("right lower lobe") which is abnormal ("eventration"), however, it fails to precisely describe this abnormality. In addition, the model also finds abnormalities about "interstitial opacities" and "atheroscalerotic calcification", which are not considered as true abnormality by human experts. The potential reason for this mis-description might be that this x-ray image is darker (compared with the above images), and our model might be very sensitive to this change.

For the single-sentence generation results (shown in the lower part of Table 1), the ablated versions of our model (Ours-Semantic-only and Ours-Visual-only) achieve competitive scores compared with the state-of-the-art methods. Our full model (Ours-CoAttention) outperforms all of the baseline, which indicates the effectiveness of the proposed co-attention mechanism.

(47)
Failure Cases

Ground Truth
No acute cardiopulmonary abnormality. Normal heart size mediastinal contours. Eventration of the right hemidiaphragm. No focal airspace consolidation. No pleural effusion or pneumothorax.

Generated
No acute cardiopulmonary abnormality. Stable appearance of the thoracic aorta. The right lateral lower lobe is noted in the right lower right midlung. No large pleural effusion or focal airspace disease. Mild interstitial opacities. Atherosclerotic calcifications bony structures bilaterally. There is no pleural effusion or pneumothorax developed in the right lower lobe.

No acute cardiopulmonary abnormality. Heart size appears within normal limits. Pulmonary vasculature appears within normal limits. Overlying the middle cardiac silhouette representing a hiatal hernia. No focal consolidation pleural effusion or pneumothorax. No acute bony abnormality.

No active disease. The heart and lungs have in the interval. Nipple and lateral lucency in the lungs suggestive of focal airspace disease. The lungs are hyperexpanded consistent with emphysema in the left lower lobe. This is most at the upper lobes. This may indicate hypoventilated irregularities or effusions. The lungs are otherwise grossly clear. Resolution of by normal pleural effusion.
Recap: Model Architecture

(I) Lesion tag classification

(III) Visual-semantic co-attention

(II) Hierarchical text generation
Outline

- Medical image report generation
  - Co-attention, hierarchical generation, multi-task
  - Further improvement: retrieval+generation, structured knowledge

- Paragraph description of natural images

- Text generation under control
  - Various text properties, granularities, amount of supervision

- All in one toolkit: Texar
Improving Abnormality Description

- Normal v.s. abnormal findings
  - **Normal findings**: dominate the reports; general, templated descriptions
  - **Abnormal findings**: relatively rare, but critical; more specifically stated
Improving Abnormality Description

- Normal v.s. abnormal findings
  - Normal findings: dominate the reports; general, templated descriptions
  - Abnormal findings: relatively rare, but critical; more specifically stated

Ground Truth: The heart size and mediastinal contours appear within normal limits. There is blunting of the right lateral costophrenic sulcus which could be secondary to a small effusion versus scarring. No focal airspace consolidation or pneumothorax. No acute bony abnormalities.
Improving Abnormality Description

- Normal v.s. abnormal findings
  - **Normal findings**: dominate the reports; general, templated descriptions
  - **Abnormal findings**: relatively rare, but critical; more specifically stated

- A pure **generation-based** model tends to **overfit** to normal findings
  - Make it easier to generate fluent, natural-looking sentences

---

**Ground Truth**: The heart size and mediastinal contours **appear within normal limits**. There is **blunting** of the right lateral costophrenic **sulcus** which could be secondary to a small **effusion** versus scarring. No **focal** airspace consolidation or pneumothorax. No **acute** bony abnormalities.
Improving Abnormality Description

- Normal v.s. abnormal findings
  - **Normal findings**: dominate the reports; general, templated descriptions
  - **Abnormal findings**: relatively rare, but critical; more specifically stated

- A pure generation-based model tends to overfit to normal findings
  - Make it easier to generate fluent, natural-looking sentences

**Ground Truth**: The heart size and mediastinal contours appear within normal limits. There is blunting of the right lateral costophrenic sulcus which could be secondary to a small effusion versus scarring. No focal airspace consolidation or pneumothorax. No acute bony abnormalities.

**G**: The heart size is normal. No pleural effusion or pneumothorax. No acute bony abnormalities.
Improving Abnormality Description

- Normal v.s. abnormal findings
  - Normal findings: dominate the reports; general, templated descriptions
  - Abnormal findings: relatively rare, but critical; more specifically stated

- A pure generation-based model tends to overfit to normal findings
  - Make it easier to generate fluent, natural-looking sentences

- Solution: alleviate the burden of generating natural sentences

Ground Truth: The heart size and mediastinal contours appear within normal limits. There is blunting of the right lateral costophrenic sulcus which could be secondary to a small effusion versus scarring. No focal airspace consolidation or pneumothorax. No acute bony abnormalities.

G: The heart size is normal. No pleural effusion or pneumothorax. No acute bony abnormalities.
Improving Abnormality Description

- Normal v.s. abnormal findings
  - **Normal findings**: dominate the reports; general, templated descriptions
  - **Abnormal findings**: relatively rare, but critical; more specifically stated

- A pure generation-based model tends to overfit to normal findings
  - Make it easier to generate fluent, natural-looking sentences

- **Solution**: alleviate the burden of generating natural sentences
  - Method: retrieval + generation

**Ground Truth**: The heart size and mediastinal contours appear within normal limits. There is blunting of the right lateral costophrenic sulcus which could be secondary to a small effusion versus scarring. No focal airspace consolidation or pneumothorax. No acute bony abnormalities.

**G**: The heart size is normal. No pleural effusion or pneumothorax. No acute bony abnormalities.
Improving Abnormality Description - I: Retrieval + Generation

- **Retrieve** template sentences from a database according to input features
- **Rewrite** the templates for more accurate description

![Diagram of the KERP model](image)

[Li et al., 2018; 2019]
Improving Abnormality Description - I: Retrieval + Generation

- Retrieve template sentences from a database according to input features
- Rewrite the templates for more accurate description

Ground Truth: The heart size and mediastinal contours appear within normal limits. There is blunting of the right lateral costophrenic sulcus which could be secondary to a small effusion versus scarring. No focal airspace consolidation or pneumothorax. No acute bony abnormalities.
Improving Abnormality Description - I: Retrieval + Generation

- Retrieve template sentences from a database according to input features
- Rewrite the templates for more accurate description

Ground Truth: The heart size and mediastinal contours appear within normal limits. There is blunting of the right lateral costophrenic sulcus which could be secondary to a small effusion versus scarring. No focal airspace consolidation or pneumothorax. No acute bony abnormalities.

G: The heart size is normal. No pleural effusion or pneumothorax. No acute bony abnormalities.
Improving Abnormality Description - I: Retrieval + Generation

- Retrieve template sentences from a database according to input features
- Rewrite the templates for more accurate description

**Ground Truth:** The heart size and mediastinal contours appear within normal limits. There is **blunting** of the right lateral costophrenic **sulcus** which could be secondary to a small **effusion** versus scarring. No focal airspace consolidation or pneumothorax. No acute bony abnormalities.

**G:** The heart size is normal. **No pleural effusion** or pneumothorax. No acute bony abnormalities.

**R:** The heart size is normal. **There is mild effusion.** No acute bony abnormalities.
Improving Abnormality Description - I: Retrieval + Generation

- Retrieve template sentences from a database according to input features
- Rewrite the templates for more accurate description

**Ground Truth:** The heart size and mediastinal contours appear within normal limits. There is **blunting** of the right lateral costophrenic sulcus which could be secondary to a small **effusion** versus scarring. No focal airspace consolidation or pneumothorax. No acute bony abnormalities.

**G:** The heart size is normal. No pleural effusion or pneumothorax. No acute bony abnormalities.

**R:** The heart size is normal. There is mild effusion. No acute bony abnormalities.

**R+G:** The heart size is normal. There is blunting of costophrenic sulcus suggesting a small effusion. No acute bony abnormalities.
Improving Abnormality Description - II

- Structured medical knowledge
Improving Abnormality Description - II

- **Structured medical knowledge**
  - Common types of abnormality:
    - presence of abnormal attributes
    - absence of typical attributes
    - abnormal change of object shape or location
Improving Abnormality Description - II

- **Structured medical knowledge**
  - Common types of abnormality:
    - presence of abnormal attributes
    - absence of typical attributes
    - abnormal change of object shape or location
  - Different abnormality exhibits certain correlation patterns
Improving Abnormality Description - II

- Structured medical knowledge
  - Common types of abnormality:
    - presence of abnormal attributes
    - absence of typical attributes
    - abnormal change of object shape or location
  - Different abnormality exhibits certain correlation patterns

Construct a prior abnormality graph from data that captures abnormality co-occurrence patterns
Improving Abnormality Description - II: Incorporating structured knowledge

- Graph Transformer (GTR):
  - A universal transformation model between modalities
Improving Abnormality Description - II: Incorporating structured knowledge

- **Graph Transformer (GTR):**
  - A universal transformation model between modalities

**image-to-graph GTR**

![Diagram of GTR](image)

- **Encode:** visual feature to knowledge graph
- **GTR\textsubscript{i2g}**
- **Retrieve:** abnormality graph to a disease graph in order to predict complications
- **Paraphrase:** the abnormality graph as a template sequence, the words of which are then retrieved and paraphrased by a template module

**Templates**

- Hyperexpansion of lungs (0.78)
- Tortuous aorta (0.12)
- Focal airspace consolidation (0.01)
- Low lung volumes (0.00)
- Enlarged heart size (0.04)
- Degenerative change of spine (0.66)

**Abnormality graph**

**Retrieval-Rewriting Text generation**

- **Learned weights + \lambda \cdot prior weights**
Improving Abnormality Description - II: Incorporating structured knowledge

- Graph Transformer (GTR):
  - A universal transformation model between modalities

*image-to-graph GTR*  *graph-to-graph GTR*

Diagram showing the architecture of KERP (Knowledge-driven Encode, Retrieve, Paraphrase) for medical image report generation.

- **Encode GTR**
- **Retrieval-Rewriting**
- **Text generation**

Graph Transformer (GTR) is used to encode visual features as an abnormality graph and retrieve and paraphrase the abnormality graph as a template sequence. The words of the template sequence are then used to decode the abnormality graph as a template sequence, which is then retrieved and paraphrased by GTR to generate the final report.

Abnormality graph with nodes representing abnormal findings such as hyperexpansion of lungs, tortuous aorta, focal airspace consolidation, low lung volumes, enlarged heart size, degenerative change of spine, and values indicating the probabilities of these findings.

Diagram includes a CNN for extracting image features and a GTR for encoding these features as an abnormality graph.

Learned weights + λ · prior weights are used in the graph transformation process.
Improving Abnormality Description

**image-to-graph GTR  graph-to-graph GTR**

**Abnormality graph**

- Hyperexpansion of lungs (0.78)
- Tortuous aorta (0.12)
- Focal airspace consolidation (0.01)
- Enlarged heart size (0.04)
- Low lung volumes (0.00)
- Degenerative change of spine (0.66)

**CNN**

**Encode**

**GTR _{i2g}**

**Retrieval-Rewriting**

**Text generation**
Improving Abnormality Description

image-to-graph GTR  graph-to-graph GTR  graph-to-seq GTR  seq-to-seq GTR

Abnormality graph

Visu.feature

Visual feature

Encode

GTR_{12g}

Retrieve

GTR_{g2s}

Paraphrase

GTR_{gs2s}

Report

Dege.nerative changes in the spine.
No pleural effusion.
There is hyperexpansion of the lungs suggesting underlying emphysema.
No focal airspace consolidation.
Heart size is normal.

Hyperexpansion of lungs (0.78)
Focal airspace consolidation (0.01)
Enlarged heart size (0.04)
Degenerative change of spine (0.66)
Low lung volumes (0.00)
Tortuous aorta (0.12)

0.03
0.12
0.00
0.19
0.04

Templates
## Empirical Results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>CIDEr</th>
<th>ROUGE-L</th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
<th>Hit (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IU X-Ray</td>
<td>CNN-RNN</td>
<td>0.294</td>
<td>0.307</td>
<td>0.216</td>
<td>0.124</td>
<td>0.087</td>
<td>0.066</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>LRCN</td>
<td>0.285</td>
<td>0.307</td>
<td>0.223</td>
<td>0.128</td>
<td>0.089</td>
<td>0.068</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>AdaAtt</td>
<td>0.296</td>
<td>0.308</td>
<td>0.220</td>
<td>0.127</td>
<td>0.089</td>
<td>0.069</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Att2in</td>
<td>0.297</td>
<td>0.307</td>
<td>0.224</td>
<td>0.129</td>
<td>0.089</td>
<td>0.068</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>CoAtt*</td>
<td>0.277</td>
<td>0.307</td>
<td>0.224</td>
<td>0.129</td>
<td>0.089</td>
<td>0.068</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>HRGR-Agent</td>
<td>0.343</td>
<td>0.322</td>
<td>0.455</td>
<td>0.288</td>
<td>0.205</td>
<td>0.154</td>
<td>24.100</td>
</tr>
<tr>
<td></td>
<td>KER</td>
<td>0.318</td>
<td>0.335</td>
<td>0.455</td>
<td>0.304</td>
<td>0.210</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td></td>
<td>KERP</td>
<td>0.280</td>
<td>0.339</td>
<td>0.482</td>
<td>0.325</td>
<td>0.226</td>
<td>0.162</td>
<td>57.425</td>
</tr>
<tr>
<td>CX-CHR</td>
<td>CNN-RNN</td>
<td>1.580</td>
<td>0.578</td>
<td>0.592</td>
<td>0.506</td>
<td>0.450</td>
<td>0.411</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>LRCN</td>
<td>1.589</td>
<td>0.577</td>
<td>0.593</td>
<td>0.508</td>
<td>0.459</td>
<td>0.413</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>AdaAtt</td>
<td>1.568</td>
<td>0.576</td>
<td>0.588</td>
<td>0.505</td>
<td>0.446</td>
<td>0.409</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Att2in</td>
<td>1.564</td>
<td>0.576</td>
<td>0.587</td>
<td>0.503</td>
<td>0.447</td>
<td>0.403</td>
<td>25.937</td>
</tr>
<tr>
<td></td>
<td>HRG</td>
<td>2.800</td>
<td>0.588</td>
<td>0.629</td>
<td>0.547</td>
<td>0.497</td>
<td>0.463</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>HRGR-Agent</td>
<td>2.895</td>
<td>0.612</td>
<td>0.673</td>
<td>0.587</td>
<td>0.530</td>
<td>0.486</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>KER</td>
<td>0.817</td>
<td>0.552</td>
<td>0.609</td>
<td>0.489</td>
<td>0.400</td>
<td>0.335</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>KERP</td>
<td>2.850</td>
<td>0.618</td>
<td>0.673</td>
<td>0.588</td>
<td>0.532</td>
<td>0.473</td>
<td>67.820</td>
</tr>
</tbody>
</table>

Table 1: Automatic and human evaluation on IU X-Ray (upper part) and CX-CHR dataset (lower part) compared with CNN-RNN (Vinyals et al. 2015), LRCN (Donahue et al. 2015), AdaAtt (Lu et al. 2017), Att2in (Rennie et al. 2017), CoAtt (Jing, Xie, and Xing 2018), and HRGR-Agent (Li et al. 2018). * indicates re-training and evaluation on our data split.
Empirical Examples

The cardiac silhouette is mildly enlarged. Mediastinal contours are within normal limits. The pulmonary vascularity is increased. There is large right - sided pleural effusion and probable underlying associated compressive atelectasis. Mild perihilar xxxx opacities, xxxx edema. No pneumothorax is seen.

There is a small left pleural effusion. No pneumothorax. Heart size normal the lungs are clear.

There are bilateral pleural effusions with bibasilar airspace disease, right greater than left. No pneumothorax. Cardiac silhouette is at the upper limits of normal. Clear lungs.
Summary: Medical Image Report Generation

- **Challenges:** (1) abnormality detection and description; (2) Long paragraph generation; (3) Accurate, visually-grounded description

- **A set of techniques for solution**
  - **Cross modalities:** images, text, graphs
    - *Graph Transformer*
  - **Long text generation in a more structured way**
    - *Hierarchical generation*
    - *Combination of retrieval + generation*
  - **Integrating structured knowledge, visual grounding**
    - *Structured medical knowledge*
    - *Visual-semantic co-attention*
Outline

- Medical image report generation
  - Co-attention, hierarchical generation, multi-task
  - Further improvement: retrieval+generation, structured knowledge

- Paragraph description of natural images

- Text generation under control
  - Various text properties, granularities, amount of supervision

- All in one toolkit: Texar
Paragraph description of natural images

[Liang et al., 2017]
Paragraph description of natural images

Semantic region detection & captioning

Local Phrases

- people playing baseball
- a man wearing white shirt and pants
- man holding a baseball bat
- person wearing a helmet in the field
- a man bending over
Paragraph description of natural images

Figure 2. Our RTT-GAN alternatively optimizes a structured paragraph generator and two discriminators following an adversarial training implementation. The optimization of the generator and discriminators is given by the following joint objective function:

\[
    \text{max} \sum_{t,i} \log D_{t,i} \left( \hat{s}_{t,i} \right) - \sum_{t,i} \log (1 - D_{t,i}) \left( s_{t,i} \right) + \nu_{t,i} L_{\text{recon}}(s_{t,i})
\]

where

- \( s_{t,i} \) denotes the \( i \)-th sentence of the paragraph at time \( t \),
- \( \hat{s}_{t,i} \) denotes the \( i \)-th sentence generated by the generator,
- \( D_{t,i} \) denotes the discriminator that detects whether \( s_{t,i} \) is real or generated,
- \( L_{\text{recon}}(s_{t,i}) \) denotes the reconstruction loss for generator optimization,
- \( \nu_{t,i} \) is the balancing parameter.

The discrete nature of text samples hinders gradient back-propagation from the discriminators to the generator. To leverage existing image-paragraph pair dataset in the supervised setting, we also incorporate the traditional word reconstruction loss for generator optimization, which is defined as:

\[
    L_{\text{recon}}(s_{t,i}) = \sum_{w_{t,i}} \left( \hat{p}(w_{t,i} | s_{t,i-1}) - p(w_{t,i} | s_{t,i-1}) \right)^2
\]

where

- \( w_{t,i} \) denotes the \( i \)-th word of the paragraph at time \( t \),
- \( \hat{p}(w_{t,i} | s_{t,i-1}) \) denotes the predicted probability distribution of the word given the previous sentence,
- \( p(w_{t,i} | s_{t,i-1}) \) denotes the true probability distribution of the word given the previous sentence.

The following paragraphs describe the details of our RTT-GAN framework.

3.1. Adversarial Objective

The objective of the adversarial framework is written as:

\[
    \text{max} \sum_{t,i} \log D_{t,i} \left( \hat{s}_{t,i} \right) - \sum_{t,i} \log (1 - D_{t,i}) \left( s_{t,i} \right) + \nu_{t,i} L_{\text{recon}}(s_{t,i})
\]

where

- \( s_{t,i} \) denotes the \( i \)-th sentence of the paragraph at time \( t \),
- \( \hat{s}_{t,i} \) denotes the \( i \)-th sentence generated by the generator,
- \( D_{t,i} \) denotes the discriminator that detects whether \( s_{t,i} \) is real or generated,
- \( L_{\text{recon}}(s_{t,i}) \) denotes the reconstruction loss for generator optimization,
- \( \nu_{t,i} \) is the balancing parameter fixed to 0.001.

In practice, we represent a paragraph as a sequence of sentences \( s = \{ s_1, s_2, \ldots, s_T \} \), where each sentence is a sequence of words \( s_t = \{ w_1, w_2, \ldots, w_{N_t} \} \). Each synthesized sentence is then fed into a sentence discriminator and a recurrent topic-transition discriminator for assessing sentence plausibility and topic coherence, respectively. A paragraph description corpus is adopted to provide linguistic knowledge about paragraph generation, which depicts the true data distribution of the discriminators.
Paragraph description of natural images

We construct an adversarial game between the generator and the multi-level discriminator of our RTT-GAN, then describe detailed model design and training. The objective of the adversarial framework is written as:

\[ \mathbb{E}_{P \sim \mathcal{D}} \left[ D \left( P \right) \right] - \mathbb{E}_{G \sim \mathcal{G}} \left[ D \left( \hat{P} \right) \right] = \lambda \mathbb{E}_{P \sim \mathcal{D}} \left[ \sum_{t=1}^{T} \log \left( 1 - D \left( P_{1:t} \right) \right) \right] + \mathbb{E}_{\hat{P} \sim \mathcal{G}} \left[ \sum_{t=1}^{T} \log \left( D \left( \hat{P}_{1:t} \right) \right) \right] \]

where \( D \) is the discriminator, \( G \) is the generator, and \( \lambda \) is the balancing parameter fixed to \( 0.001 \).

The generator recurrently produces each sentence by reasoning about local semantic regions and preceding paragraph state. Each synthesized sentence is then fed into a sentence discriminator and a recurrent topic-transition discriminator for assessing sentence plausibility and topic coherence, respectively. A paragraph description corpus is adopted to provide linguistic knowledge about paragraph plausibility and topic-transition coherence. The objective of the sentence discriminator is written as:

\[ \mathbb{E}_{\hat{S} \sim \mathcal{G}} \left[ \log \left( D \left( \hat{S} \right) \right) \right] \]

The discriminators learn to differentiate real sentences from the synthesized ones. Reasoning sentence generation consists of a sequence of sentences, conditioned on preceding sentences.

Visual-text co-attention

Hierarchical text generation

Semantic region detection & captioning

Semantic Regions

Attentive Reasoning

Generator

Sentence

Sentence

Sentence

Sentence

Sentence

Sentence

Sentence

Paragraph description Corpus

Sentence Discriminator

Topic-Transition Discriminator
Paragraph description of natural images

Semantic Regions

Attentive Reasoning

Generator

Sentence

Sentence

Sentence

Sentence

Sentence Discriminator

Topic-Transition Discriminator

Paragraph description Corpus

Visual-text co-attention

Hierarchical text generation

Multi-level adversarial learning

Semantic region detection & captioning

Paragraph description of natural images
Paragraph description of natural images

Paragraph: A group of people are riding bikes. There are two people riding bikes parked on the sidewalk. He is wearing a black shirt and jeans. A woman is wearing a short sleeve yellow shirt and shorts. There are many other people on the red and black bikes. A woman wearing a shirt is riding a bicycle.
Paragraph description of natural images

<table>
<thead>
<tr>
<th>Method</th>
<th>METEOR</th>
<th>CIDEr</th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence-Concat</td>
<td>12.05</td>
<td>6.82</td>
<td>31.11</td>
<td>15.10</td>
<td>7.56</td>
<td>3.98</td>
</tr>
<tr>
<td>Template</td>
<td>14.31</td>
<td>12.15</td>
<td>37.47</td>
<td>21.02</td>
<td>12.03</td>
<td>7.38</td>
</tr>
<tr>
<td>Image-Flat [14]</td>
<td>12.82</td>
<td>11.06</td>
<td>34.04</td>
<td>19.95</td>
<td>12.20</td>
<td>7.71</td>
</tr>
<tr>
<td>Regions-Hierarchical [16]</td>
<td>15.95</td>
<td>13.52</td>
<td>41.90</td>
<td>24.11</td>
<td>14.23</td>
<td>8.69</td>
</tr>
<tr>
<td>RTT-GAN (Semi- w/o discriminator)</td>
<td>12.35</td>
<td>8.96</td>
<td>33.82</td>
<td>17.40</td>
<td>9.01</td>
<td>5.88</td>
</tr>
<tr>
<td>RTT-GAN (Semi- w/o sentence D)</td>
<td>11.22</td>
<td>10.04</td>
<td>35.29</td>
<td>19.13</td>
<td>11.55</td>
<td>6.02</td>
</tr>
<tr>
<td>RTT-GAN (Semi- w/o topic-transition D)</td>
<td>12.68</td>
<td>12.77</td>
<td>37.20</td>
<td>20.51</td>
<td>12.08</td>
<td>6.91</td>
</tr>
<tr>
<td>RTT-GAN (Semi-)</td>
<td>14.08</td>
<td>13.07</td>
<td>39.22</td>
<td>22.50</td>
<td>13.34</td>
<td>7.75</td>
</tr>
<tr>
<td>RTT-GAN (Fully- w/o discriminator)</td>
<td>16.57</td>
<td>15.07</td>
<td>41.86</td>
<td>24.33</td>
<td>14.56</td>
<td>8.99</td>
</tr>
<tr>
<td>RTT-GAN (Fully-)</td>
<td>17.12</td>
<td>16.87</td>
<td>41.99</td>
<td>24.86</td>
<td>14.89</td>
<td>9.03</td>
</tr>
<tr>
<td>RTT-GAN (Semi + Fully)</td>
<td><strong>18.39</strong></td>
<td><strong>20.36</strong></td>
<td><strong>42.06</strong></td>
<td><strong>25.35</strong></td>
<td><strong>14.92</strong></td>
<td><strong>9.21</strong></td>
</tr>
<tr>
<td>Human</td>
<td>19.22</td>
<td>28.55</td>
<td>42.88</td>
<td>25.68</td>
<td>15.55</td>
<td>9.66</td>
</tr>
</tbody>
</table>
Outline

- Medical image report generation
  - Co-attention, hierarchical generation, multi-task
  - Further improvement: retrieval+generation, structured knowledge

- Paragraph description of natural images

- Text generation under control
  - Various text properties, granularities, amount of supervision

- All in one toolkit: Texar
Beyond Image-to-Text Generation

- Controlled Generation of Text [Hu et al., 2017]
  - With control over content, attributes, stylistic characteristics, …
  - At sentence / discourse level
  - In supervised / unsupervised settings
Beyond Image-to-Text Generation

- Controlled Generation of Text [Hu et al., 2017]
  - With control over content, attributes, stylistic characteristics, ...
  - At sentence / discourse level
  - In supervised / unsupervised settings

- E.g., Text Style Transfer
  [Hu et al., 2017; Yang et al., 2018]

**Original:** if i could give them a zero star review i would !
**Output:** if i can give them a five star review i would !

**Original:** the food is fresh and the environment is good
**Output:** the food is bland and the environment is bad

**Original:** i thought the movie was too bland and too much
**Output-1:** i guess the movie is too bland and too much
**Output-2:** i guess the film will have been too bland
Beyond Image-to-Text Generation

- Controlled Generation of Text [Hu et al., 2017]
  - With control over content, attributes, stylistic characteristics, ...
  - At sentence / discourse level
  - In supervised / unsupervised settings

- E.g., Text Style Transfer
  [Hu et al., 2017; Yang et al., 2018]

- E.g., Content Manipulation
  [Wang, Hu, et al., 2019]

<table>
<thead>
<tr>
<th>Content Record</th>
<th>PLAYER</th>
<th>PT</th>
<th>RB</th>
<th>AS</th>
<th>PLAYER</th>
<th>PT</th>
</tr>
</thead>
<tbody>
<tr>
<td>LeBron_James</td>
<td></td>
<td>32</td>
<td>4</td>
<td>7</td>
<td>Kyrie_Irving</td>
<td>20</td>
</tr>
</tbody>
</table>

Reference Sentence:
Jrue_Holiday led the way with 26 points and 6 assists, while Goran_Dragic scored 23 points and pulled down 8 rebounds.

Output:
LeBron_James led the way with 32 points, 7 assists and 4 rebounds, while Kyrie_Irving scored 20 points.
Beyond Image-to-Text Generation

- Controlled Generation of Text [Hu et al., 2017]
  - With control over content, attributes, stylistic characteristics, ...
  - At sentence / discourse level
  - In supervised / unsupervised settings

- E.g., Text Style Transfer
  [Hu et al., 2017; Yang et al., 2018]

- E.g., Content Manipulation
  [Wang, Hu, et al., 2019]

- E.g., Conversation Strategy Control
  [Tang et al., 2019]

---

Guide conversation to a target topic: e-books

Agent: hi how are you doing
Human: not so good. i am really tired
Agent: oh i'm sorry to hear
Human: i have a lot of work to do
Agent: what kind was it?
Human: i am writing a chatbot program
Agent: interesting. i read about coding from a book
Human: really
Agent: i prefer e-books over paperback book
Outline

- Medical image report generation
  - Co-attention, hierarchical generation, multi-task
  - Further improvement: retrieval+generation, structured knowledge

- Paragraph description of natural images

- Text generation under control
  - Various text properties, granularities, amount of supervision

- All in one toolkit: Texar
Text Generation Tasks

- Generates *natural language* from input *data or machine representations*
- Spans a broad set of natural language processing (NLP) tasks:

<table>
<thead>
<tr>
<th>Task</th>
<th>Input X</th>
<th>Output Y (Text)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chatbot / Dialog System</td>
<td>Utterance</td>
<td>Response</td>
</tr>
<tr>
<td>Machine Translation</td>
<td>English</td>
<td>Chinese</td>
</tr>
<tr>
<td>Summarization</td>
<td>Document</td>
<td>Short paragraph</td>
</tr>
<tr>
<td>Description Generation</td>
<td>Structured data</td>
<td>Description</td>
</tr>
<tr>
<td>Captioning</td>
<td>Image/video</td>
<td>Description</td>
</tr>
<tr>
<td>Speech Recognition</td>
<td>Speech</td>
<td>Transcript</td>
</tr>
</tbody>
</table>

Courtesy: Neubig, 2017
Various (Deep Learning) Techniques

- Various model architectures

E: encoder, D: decoder, C: Classifier, A: attention, Prior: prior distribution, M: memory
Texar

A General-Purpose Text Generation Toolkit
A General-Purpose Text Generation Toolkit

Hundreds of thousands of lines of code
A General-Purpose Text Generation Toolkit

With the toolkit

Hundreds of thousands of lines of code
Texar Overview

- A **unified platform** aiming to cover many text generation tasks
  - Provide the most **comprehensive** set of well-tailored and ready-to-use modules
  - Enable **reuse** of common components and functionalities
  - **Standardize** design, implementation, and experimentation
  - Encourage **technique sharing** among different tasks

- Based on TensorFlow

- Open-source under Apache License 2.0
Assemble any complex model like playing building blocks
Texar Highlights

Modularized
Assemble any complex model like playing building blocks

Versatile
Supports a large variety of applications/models/algorithms...
...
Texar Highlights

**Modularized**
Assemble any complex model like playing building blocks

**Versatile**
Supports a large variety of applications/models/algorithm...

**Extensible**
Allows to plug in any customized or external modules
# Texar Stack

**Texar stack**

<table>
<thead>
<tr>
<th>Applications</th>
<th>Library APIs</th>
<th>Model templates + Config files</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>Evaluation</td>
<td>Prediction</td>
</tr>
</tbody>
</table>

## Models

<table>
<thead>
<tr>
<th>Architectures</th>
<th>Losses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encoder</td>
<td>(Seq) MaxLikelihood</td>
</tr>
<tr>
<td>Decoder</td>
<td>Adversarial</td>
</tr>
<tr>
<td>Embedder</td>
<td>Dialog</td>
</tr>
<tr>
<td>Classifier</td>
<td>Numerical</td>
</tr>
<tr>
<td>Connector</td>
<td>Multi-field/type Parallel</td>
</tr>
<tr>
<td>Policy</td>
<td>Ir decay / grad clip / ...</td>
</tr>
<tr>
<td>QNet</td>
<td></td>
</tr>
<tr>
<td>Rewards</td>
<td></td>
</tr>
<tr>
<td>RL-related</td>
<td></td>
</tr>
<tr>
<td>Regularize</td>
<td></td>
</tr>
</tbody>
</table>

## Data

<table>
<thead>
<tr>
<th></th>
<th>MonoText</th>
<th>PairedText</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Executor</td>
<td>Optimizer</td>
</tr>
<tr>
<td></td>
<td>Seq/Episodic RL Agent</td>
<td></td>
</tr>
</tbody>
</table>

## Trainer

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimizer</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Memory

- ...
Example: Build a sequence-to-sequence model

```
# Read data
dataset = PairedTextData(data_hparams)
batch = DataIterator(dataset).get_next()

# Encode
embedder = WordEmbedder(dataset.vocab.size, hparams=embedder_hparams)
encoder = TransformerEncoder(hparams=encoder_hparams)
enc_outputs = encoder(embedder(batch['source_text_ids']),
                      batch['source_length'])

# Decode
decoder = AttentionRNNDecoder(memory=enc_outputs,
                               hparams=decoder_hparams)
outputs, length, _ = decoder(inputs=embedder(batch['target_text_ids']),
                             seq_length=batch['target_length']-1)

# Loss
loss = sequence_sparse_softmax_cross_entropy(
       labels=batch['target_text_ids'][1:, :],
       logits=outputs.logits, seq_length=length)
```

(1) Customize model template
via a YAML config file

(2) Program with Texar Python Library APIs
Resources

- Website: https://texar.io
- GitHub: https://github.com/asyml/texar
- Examples: https://github.com/asyml/texar/blob/master/examples
- Documentation: https://texar.readthedocs.io/
- Blog: https://medium.com/@texar
The Petuum Vision

Industry Agnostic

Building AI Like Lego

AI With No Tears

Completed software

10% White-Glove Assembly

90% Completed building blocks

Petuum OS

Robo-radiologist
Insurance Auto-Report
Virtual EA
Smart Expense Reports
Smart Catalog
Robot Store Staff

Raw Data Enrichment
Model/Algorithm
System/Infra
Summary

- Medical image report generation
- Cross modalities: images, text, graphs
- Structured long text generation: retrieval/gen; hierarchical
- Integrating medical knowledge

- Paragraph description of natural images

- Text generation under control
  - Various text properties, granularities, amount of supervision

- All in one toolkit: Texar

**Findings:**

- There are no focal areas of consolidation.
- No suspicious pulmonary opacities.
- Heart size within normal limits.
- No pleural effusions.
- There is no evidence of pneumothorax.
- Degenerative changes of the thoracic spine.

**Table 2. Ablation studies on the effectiveness of key components**

<table>
<thead>
<tr>
<th>Method</th>
<th>METEOR</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>RTT-GAN (Fully-w/o phrase att)</td>
<td>16.08</td>
<td>15.13</td>
</tr>
<tr>
<td>RTT-GAN (Fully-w/o att)</td>
<td>16.92</td>
<td>16.15</td>
</tr>
<tr>
<td>RTT-GAN (Fully-20 regions)</td>
<td>17.01</td>
<td>16.26</td>
</tr>
<tr>
<td>RTT-GAN (Fully-10 regions)</td>
<td>16.91</td>
<td>16.14</td>
</tr>
<tr>
<td>RTT-GAN (Fully)</td>
<td>16.92</td>
<td>16.15</td>
</tr>
</tbody>
</table>

**Paragraph:** A group of people are riding bikes. There are two people riding bikes parked on the sidewalk. He is wearing a black shirt and pants. A woman is wearing a short sleeve yellow shirt and shorts. There are many other people on the red and black bikes. A woman wearing a shirt is riding a bicycle.

**Table 3. Human evaluation results**

<table>
<thead>
<tr>
<th>Content</th>
<th>Reference</th>
<th>Desired</th>
<th>PLAyer</th>
<th>PT</th>
<th>PT</th>
</tr>
</thead>
<tbody>
<tr>
<td>LeBron</td>
<td>Jru Holiday led the way with 26 points and 6 assists , while Goran Dragic scored 23 points and pulled down 8 rebounds .</td>
<td>LeBron James led the way with 32 points , 7 assists and 4 rebounds , while Kyrie Irving scored 20 points .</td>
<td>32</td>
<td>4</td>
<td>7</td>
</tr>
</tbody>
</table>

**Image:**

A group of people are riding bikes. There are two people riding bikes parked on the sidewalk. He is wearing a black shirt and pants. A woman is wearing a short sleeve yellow shirt and shorts. There are many other people on the red and black bikes. A woman wearing a shirt is riding a bicycle.