Towards Image Caption Generation for Art Historical Data

Gupta et al.

AI methods for digital heritage, KI2020

Prathmesh Madhu, M.Sc.,
PhD Student

Pattern Recognition Lab, FAU
Who am I?

- PhD student at the Chair of Pattern Recognition, FAU, Germany.
- Research Focus: **Scene understanding** in digital humanities, with a major focus on Iconographics and Narratives
- Email: prathmesh.madhu@fau.de
- Lab: https://lme.tf.fau.de/person/madhu/
- Personal: https://prathmeshrmadhu.github.io/
Motivation

The Virgin, kneeling on the left, an open book on a table behind her, bowing her head, one hand at her breast, the other gesturing outwards, while the angel Gabriel steps down over clouds on the right, holding a lily, pointing upwards towards the dove, representing the Holy Spirit, from which a shaft of light floods over the Virgin; after Cosway. 1800

Mezzotint

In his etchings sometimes Bellange deliberately makes the space vague, as in the great Annunciation.
Motivation

Once we have the captions,

a) Objects, Characters in the scene?
b) Positions and Gestures?
c) Type of background/Foreground?
d) Which regions are important/highlighted in the scene?

Captioning very important from
Scene understanding → Scene Describing
Outline

1. Preliminary Contributions
2. Dataset generation
3. Method and Ablation studies
4. Results and Discussion
5. Conclusion
6. Further Research Directions
Preliminary Contributions

- Preliminary caption dataset for ~4000 images with across 9 iconographies in art history from Web gallery of Art
  - Iconographies -
    - Annunciation, Adoration, Baptism, Noli me Tangere, Still-life, nativity, virgin and child, rape, tower of babel

- Ablation Study
  - recent SOTA DL Architectures
  - Impact of embedding sizes
Dataset Generation [2]
## Dataset Generation

**Title:** annunciation

464 pictures found, 20 shown (41-60).

<table>
<thead>
<tr>
<th>Preview</th>
<th>Picture Data</th>
<th>File Info</th>
<th>Comment</th>
</tr>
</thead>
</table>
| ![BARTOLOMEO, Fra](image1.jpg) | BARTOLOMEO, Fra  
Annunciation  
1487  
Panel, 176 x 170 cm  
Cathedral of Santa Maria Assunta, Volterra  
Other works by the artist... | 1000\*1019  
True Color  
171 Kb | ![i](image2.png) |
| ![BARTOLOMEO, Fra](image3.jpg) | BARTOLOMEO, Fra  
The Annunciation (front), Circumcision and Nativity (back)  
c. 1600  
Tempera on wood, 19.5 x 9 cm  
Galleria degli Uffizi, Florence  
Other works by the artist... | 921\*443  
True Color  
70 Kb | ![i](image4.png) |
Dataset Generation

The Annunciation

Oil on copper, 62 x 47 cm
The Hermitage, St. Petersburg

Send this picture as postcard
Friendly format for printing and bookmarking
Order oil painting

Albani painted many versions of this subject. One of the best is the large altarpiece in the church of San Bartolomeo in Bologna, to which this painting is closely related.
Dataset Generation

- Histogram of number of images versus iconography

![Histogram of number of images versus iconography](image)

- Iconography:
  - Virgin and Child: 1341
  - Rape: 1201
  - Adultery: 1132
  - Assumption: 875
  - Adoration: 704
  - Sнятие: 363
  - Baptism: 315
  - Icon of Jesus: 28
  - Tower of Babel: 20

No. of images per iconography: 0 to 1400
In this painting - which may belong to a series of landscapes painted in 1570 for the Cabinet du Roi in the Château de Fontainebleau - the artist applies the formula of a vast landscape animated by small, graceful figures. He takes as his source works by Joachim Patenier.

The picture depicts a still-life with peaches and grapes fallen from an over-turned basket, resting on a partly-draped marble ledge. The painting has been reduced from a vertical to horizontal format, resulting in the loss of the signature.

This is one of the four predella scenes belonging to the altarpiece with the central panel depicting the Virgin Enthroned between Sts John the Baptist and Anthony of Padua.

Albani painted many versions of this subject. One of the best is the large altarpiece in the church of San Bartolomeo in Bologna, to which this painting is closely related.
Other Datasets

- **Flickr8k** - 8091 images
  - Training - 6000
  - Validation - 1000
  - Test - 1000
  - Each image → 5 sentences that are relatively visual and unbiased

- **Flickr30k** - 30,783 images
  - Random train, val and test splits
  - 1000 testing and 1000 validation images
Method used

- Show and Tell [1] architecture
- Various embedding sizes
  - 512, 1024, 2048
- Various deep CNN architectures
Networks used

- **InceptionV3 [3]**
  - As it was proposed by Show and Tell paper
  - Acts as the baseline

- **ResNext101[4]**
  - Combination of Inception and Resnet block (Residual + Cardinality)
  - Resnets have been state of the art backbone networks

- **EfficientNet-B0 [5]**
  - SOTA in object detection and classification
Comparison Metrics

- Bilingual evaluation understudy (BLEU) scores
  - comparing a candidate translation of text to one or more reference translations
  - BLEU - i, where i=1, 2, 3, 4 → n-grams

What about if all words are different in the candidate?

```python
# all words different
from nltk.translate.bleu_score import sentence_bleu
reference = [['the', 'quick', 'brown', 'fox', 'jumped', 'over', 'the', 'lazy', 'dog']
candidate = ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i']
score = sentence_bleu(reference, candidate)
print(score)
```

We get the worse possible score.

```
0.0
```
## Results - Quantitative

- Flickr8k and Flickr30k for all networks

<table>
<thead>
<tr>
<th>Model</th>
<th>InceptionV3</th>
<th>ResNext101</th>
<th>EfficientNetB0</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Flickr8k</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLEU1</td>
<td>59.57</td>
<td>59.24</td>
<td><strong>60.03</strong></td>
</tr>
<tr>
<td>BLEU2</td>
<td>40.00</td>
<td>40.26</td>
<td><strong>40.98</strong></td>
</tr>
<tr>
<td>BLEU3</td>
<td>25.94</td>
<td>26.51</td>
<td><strong>27.08</strong></td>
</tr>
<tr>
<td>BLEU4</td>
<td>17.36</td>
<td>17.96</td>
<td><strong>18.20</strong></td>
</tr>
<tr>
<td><strong>Flickr30k</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLEU1</td>
<td>57.07</td>
<td>57.53</td>
<td><strong>57.99</strong></td>
</tr>
<tr>
<td>BLEU2</td>
<td>37.00</td>
<td><strong>37.66</strong></td>
<td>37.33</td>
</tr>
<tr>
<td>BLEU3</td>
<td>23.23</td>
<td><strong>23.94</strong></td>
<td>23.64</td>
</tr>
<tr>
<td>BLEU4</td>
<td>15.58</td>
<td><strong>16.26</strong></td>
<td>16.06</td>
</tr>
</tbody>
</table>
## Results - Quantitative

<table>
<thead>
<tr>
<th>Model</th>
<th>InceptionV3</th>
<th>ResNext101</th>
<th>EfficientNetB0</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Flickr8k</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLEU1</td>
<td>59.57</td>
<td>59.24</td>
<td>60.03</td>
</tr>
<tr>
<td>BLEU2</td>
<td>40.00</td>
<td>40.26</td>
<td>40.98</td>
</tr>
<tr>
<td>BLEU3</td>
<td>25.94</td>
<td>26.51</td>
<td>27.08</td>
</tr>
<tr>
<td>BLEU4</td>
<td>17.36</td>
<td>17.96</td>
<td>18.20</td>
</tr>
<tr>
<td><strong>Flickr30k</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLEU1</td>
<td>57.07</td>
<td>57.53</td>
<td>57.99</td>
</tr>
<tr>
<td>BLEU2</td>
<td>37.00</td>
<td><strong>37.66</strong></td>
<td>37.33</td>
</tr>
<tr>
<td>BLEU3</td>
<td>23.23</td>
<td><strong>23.94</strong></td>
<td>23.64</td>
</tr>
<tr>
<td>BLEU4</td>
<td>15.58</td>
<td><strong>16.26</strong></td>
<td>16.06</td>
</tr>
<tr>
<td><strong>Art History</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLEU1</td>
<td>20.69</td>
<td>20.12</td>
<td><strong>22.47</strong></td>
</tr>
<tr>
<td>BLEU2</td>
<td>10.72</td>
<td>10.42</td>
<td><strong>12.02</strong></td>
</tr>
<tr>
<td>BLEU3</td>
<td>06.56</td>
<td>06.46</td>
<td><strong>07.95</strong></td>
</tr>
<tr>
<td>BLEU4</td>
<td>04.94</td>
<td>04.78</td>
<td><strong>06.27</strong></td>
</tr>
</tbody>
</table>
## Results - Flickr 8k

<table>
<thead>
<tr>
<th>Training set</th>
<th>Test set</th>
<th>Validation test</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Dog jumping" /></td>
<td><img src="image2.png" alt="Boy in car" /></td>
<td><img src="image3.png" alt="Girl jumping" /></td>
</tr>
<tr>
<td>Original: A black and white dog is jumping up to catch a green Frisbee.</td>
<td>Original: A boy is sitting in a car seat in the backseat of a car.</td>
<td>Original: A girl jumping from a diving board into a pool.</td>
</tr>
<tr>
<td>EfficientNet: a black and white dog is jumping over a hurdle</td>
<td>EfficientNet: a little boy in a red shirt is looking at the camera</td>
<td>EfficientNet: a boy jumps into a swimming pool</td>
</tr>
<tr>
<td>Inception: a black and white dog is jumping over a hurdle</td>
<td>Inception: a little boy in a red shirt is looking at the camera</td>
<td>Inception: a young boy jumps into a swimming pool</td>
</tr>
<tr>
<td>ResNext101: a black and white dog jumping over a hurdle</td>
<td>ResNext101: a little girl in a blue shirt is sitting on a couch</td>
<td>ResNext101: a young boy jumps into a swimming pool</td>
</tr>
</tbody>
</table>
## Results - Flickr 30k

<table>
<thead>
<tr>
<th>Training set</th>
<th>Test set</th>
<th>Validation set</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.jpg" alt="Image" /></td>
<td><img src="image2.jpg" alt="Image" /></td>
<td><img src="image3.jpg" alt="Image" /></td>
</tr>
</tbody>
</table>

**Original:** A large crowd is tossing someone dressed in black and red up in the air using a large cloth.

**EfficientNet:** a man in a red shirt and black pants is jumping in the air in front of a crowd of people

**Inception:** a man in a red shirt is standing in front of a building

**ResNext101:** a man in a red shirt is jumping into the air

**Original:** A man stands in front of a board covered in papers while giving a speech to a crowd.

**EfficientNet:** a group of people are sitting around a table in a room

**Inception:** a group of people are sitting at a table in a room

**ResNext101:** a group of people are sitting at a table in a restaurant

**Original:** A group of people on a riverbank some sitting on the grass some standing.

**EfficientNet:** a group of people are standing in the grass

**Inception:** a group of people are standing in front of a tree

**ResNext101:** a group of children are playing in the field
Results - Art history (no fine-tuning)

A man in a white shirt is standing in front of a building

A group of people are standing in front of a building in front of a building
Fig. 1. Still-life iconography example image from test set. **Original caption**: This still-life of plums and apples, a blue and white bowl of Strawberries, grapes, gooseberries, a Roemer of wine and a pewter vessel on a table. **Show-and-Tell**: this picture shows a still life with a roemer a silver glit columbine cup a jar a porcelain dish with an orange and peaches. **EfficientNet**: the painting represents a still life with peaches plums and grapes in a blue and white wan li porcelain dish with an orange and peaches.
Fig. 2. Rape iconography example image from validation set. **Original caption**: The Rape of the Sabines marked the climax of Giambologna’s career as an official Medici sculptor. This great marble was unveiled in the Loggia dei Lanzi in January 1583 in place of Donatello’s Judith. **Show-and-Tell**: the virgin is clothed in a drapery drawn out in very narrow folds with sharp ridges the sculpture comes as an exception from a time. **EfficientNet**: the rape of the sabines marked the climax of giambologna’s career as an official medici sculptor this great marble was unveiled in the loggia.
Discussion

- EfficientNet has the compound scaling architecture which helps in generating very good context features for *art history data*

- As we saw that our caption data is not tested for quality, and hence the metrics are low.

- Surprisingly, the best embedding size was 512, increasing it worsens the performance
Conclusion

- Initial work and progress towards coherent source agnostic image captioning in art history domain
- One step towards context based scene understanding and describing
- SOTA discriminative architectures are not always the best, however EfficientNet features improve the results by a huge amount
Future research directions

- Graphically show objects in images, postures and composition lines → captions

- Generate captions → which regions are highlighted/CAMs in the images

- Clean and Improve dataset from various sources
  - Overlapping iconographies due to unclear captions
  - Open access and academic copyrights across these 9 iconographies for narrative understanding of scenes

- Use of the generated captions for retrieval and being inclusive for vision impaired community with respect to the context and contents of artworks
References


Thank you
Virgin and Child
Rape
Still-life