FUSECAP: Leveraging Large Language Models for Enriched Fused Image Captions

Noam Rotstein, David Bensaïd, Shaked Brody, Roy Ganz, Ron Kimmel
Technion - Israel Institute of Technology

Citations: 119
Accepted to WACV 2024

Presented by Group 2: Mahad Ali, Anthony Jackson, Rafeeq Shodeinde, Isaac Tuckey
Presentation outline

- Background and Motivation
- Caption Fusion
  - Leveraging visual experts
  - LLM fuser
  - Dataset generation
- Evaluation
  - Qualitative Evaluation
  - FuseCap-trained model performance
Motivation

- Captioning is a valuable capability of VLMs
- Capability to caption is largely dependant on data used
  - Large datasets reliant on scraping the web
  - Resultant model captions overlook details
- Generating high quality captions by hand, at scale, isn’t feasible
Contributions

- FuseCap: a Framework for improving caption quality
- A large scale dataset of image-text pairs
  - 12 million pairs from assorted other datasets
  - Captions have been enriched using FuseCap
Visualization

**GIT:** a red motorcycle parked on a road near a beach

**OFA:** motorcycle parked on the beach

**Prismer:** motorcycle parked on the beach

**BLIP2:** a red motorcycle parked in a parking lot next to a fence

**Ours:** a red motorcycle with a leather and black seat is parked on the side of the road, surrounded by a wood fence and tall palm trees the clear blue sky provides a serene backdrop
A woman wearing a yellow Nike shirt and black shorts holds a white and black tennis racket while playing against a blue wall with Emirates Airline branding. She accessorizes with a gold necklace and raises her hand in excitement.
Object Detector

- Faster-RCNN with ResNeXt-152 backbone
  - Pretrained on multiple detection datasets
  - Fine-tuned on Visual Genome
  - Knowledge on 1.6k classes
- Threshold of .7 applied to bounding boxes

Attribute Detector

- Faster-RCNN with ResNeXt-152 Backbone
- Attribute Classifier added to pretrained object detector
  - Descriptors for objects in the image
    - EX: Colors, size (small, large, etc), material (steel, wooden, etc)
- Detector fine-tuned using Visual Genome
  - 400 of the 2.8 million possible attributes
- Threshold of .2 used for attribute predictions
Text Detection Module

- Character Region Awareness for Text Detection (CRAFT)
  - CNN architecture
  - Produces bounding boxes for words or characters

- Scene Text Recognition with Permutated Autoregressive Sequence Models (Parseq)
  - Encoder-Decoder Architecture for OCR
  - ViT used to encode CRAFT bounding box contents

- Unique assignment of text to objects
  - Uses bounding boxes which contain text bounding box
  - Assigns to object w/ smallest bounding box

Source: Baek et al, “Character Region Awareness for Text Detection”
A woman wearing a yellow Nike shirt and black shorts holds a white and black tennis racket while playing against a blue wall with Emirates Airline branding. She accessorizes with a gold necklace and raises her hand in excitement.
LLM Fuser Training

20k “Fusing” Dataset Generation → Visual Experts → Detailed Captions

LLM Fuser Training

FuseCap 12M Dataset Creation

Visual Experts → Flan-T5-XL → Fused Captions

Visual Experts → Flan-T5-XL → Fused Captions
Fusing Caption Dataset Creation

Original Caption

OCR

Attribute Detector

Object Detector

Prompt Template

ChatGPT

Fusing Caption

20k pairs
”A caption of an image is given: original caption. The following objects are detected in the image from left to right:
A $a^1_1, ..., a^{k-1}_1$ and $a^k_1$ o_1 [with the following text: $t_1$].

\vdots

A $a^1_n, ..., a^{k_n-1}_n$ and $a^{k_n}_n$ o_n [with the following text: $t_n$].
Write a comprehensive and concise caption of the scene using the objects detected.”
LLM Fine-tuning

- Flan-T5-XL Checkpoint
- Input: Original captions + Visual Expert output
- Target output: Enriched captions from ChatGPT
FuseCap Dataset

20k “Fusing” Dataset Generation → Visual Experts → Detailed Captions

LLM Fuser Training → Visual Experts → Flan-T5-XL → Fused Captions

FuseCap 12M Dataset Creation → Visual Experts → Flan-T5-XL → Fused Captions
BLIP Captioning Model

- FUSECAP dataset used to optimize ITC, ITM, and LM Loss in BLIP
- Fine tuned on the COCO dataset with the LM loss
- Context length increased from 30 -> 60 tokens to improve comprehensive caption generation
Experiments
Qualitative Evaluation

- Human study
- “Does caption 2 provide an additional meaningful and truthful description of the image compared to caption 1?”

Original: Mhmm, some clouds in the sky

Ours: A woman wearing dark sunglasses stands next to a red car with a black license plate reading 166882, PRI. The car has off round headlights, a chrome and silver bumper, a black tire, and a red door. The cloudy and white sky is visible in the background.
CLIPScore comparison

- CLIPScore: Cosine similarity between image and text features
  - Mean: mean score
  - Voting: choose between captions

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Captions</th>
<th>Mean</th>
<th>Voting</th>
</tr>
</thead>
<tbody>
<tr>
<td>COCO</td>
<td>Original</td>
<td>76.7</td>
<td>31.7%</td>
</tr>
<tr>
<td></td>
<td>FUSECap</td>
<td>80.3</td>
<td>67.6%</td>
</tr>
<tr>
<td>SBU</td>
<td>Original</td>
<td>71.9</td>
<td>32.1%</td>
</tr>
<tr>
<td></td>
<td>FUSECap</td>
<td>75.5</td>
<td>60.2%</td>
</tr>
<tr>
<td>CC</td>
<td>Original</td>
<td>72.6</td>
<td>34.7%</td>
</tr>
<tr>
<td></td>
<td>FUSECap</td>
<td>75.4</td>
<td>59.7%</td>
</tr>
</tbody>
</table>
## Image-Text Retrieval Task

<table>
<thead>
<tr>
<th>Model</th>
<th>img → text</th>
<th></th>
<th></th>
<th></th>
<th>text → img</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@1</td>
<td>R@5</td>
<td>R@10</td>
<td></td>
<td>R@1</td>
<td>R@5</td>
<td>R@10</td>
</tr>
<tr>
<td>BLIP†</td>
<td>75.1</td>
<td>92.7</td>
<td>96.4</td>
<td></td>
<td>58.2</td>
<td>82.4</td>
<td>89.2</td>
</tr>
<tr>
<td>BLIP-L</td>
<td>82.4</td>
<td>95.4</td>
<td>97.9</td>
<td></td>
<td>65.2</td>
<td>86.3</td>
<td>91.8</td>
</tr>
<tr>
<td>BLIP2</td>
<td>85.4</td>
<td>97.0</td>
<td>98.5</td>
<td></td>
<td>68.3</td>
<td>87.7</td>
<td>92.6</td>
</tr>
<tr>
<td>BLIP\textsuperscript{fusecap}</td>
<td>97.2</td>
<td>99.5</td>
<td>99.9</td>
<td></td>
<td>93.0</td>
<td>97.4</td>
<td>98.3</td>
</tr>
</tbody>
</table>
ITR with Generated Captions

- Captions generated by BLIP models

<table>
<thead>
<tr>
<th>Model</th>
<th>img → text</th>
<th>text → img</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLIP†</td>
<td>R@1 75.1</td>
<td>R@1 58.2</td>
</tr>
<tr>
<td></td>
<td>R@5 92.7</td>
<td>R@5 82.4</td>
</tr>
<tr>
<td></td>
<td>R@10 96.4</td>
<td>R@10 89.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>img → text</th>
<th>text → img</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLIP†</td>
<td>R@1 56.3</td>
<td>R@1 54.5</td>
</tr>
<tr>
<td></td>
<td>R@5 83.0</td>
<td>R@5 81.2</td>
</tr>
<tr>
<td></td>
<td>R@10 90.3</td>
<td>R@10 88.7</td>
</tr>
</tbody>
</table>

- COCO Retrieval

<table>
<thead>
<tr>
<th>Model</th>
<th>img → text</th>
<th>text → img</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLIP†</td>
<td>R@1 56.3</td>
<td>R@1 54.5</td>
</tr>
<tr>
<td></td>
<td>R@5 83.0</td>
<td>R@5 81.2</td>
</tr>
<tr>
<td></td>
<td>R@10 90.3</td>
<td>R@10 88.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>img → text</th>
<th>text → img</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLIP†</td>
<td>R@1 56.3</td>
<td>R@1 54.5</td>
</tr>
<tr>
<td></td>
<td>R@5 83.0</td>
<td>R@5 81.2</td>
</tr>
<tr>
<td></td>
<td>R@10 90.3</td>
<td>R@10 88.7</td>
</tr>
</tbody>
</table>

- BLIP†: BLIP with captions generated by BLIP models
### Image Captioning

- Metric: CLIPScore

<table>
<thead>
<tr>
<th>Model</th>
<th>Images</th>
<th>Parameters</th>
<th>Val</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLIP†</td>
<td>12M</td>
<td>247M</td>
<td>75.2</td>
<td>75.3</td>
</tr>
<tr>
<td>BLIP-L</td>
<td>129M</td>
<td>470M</td>
<td>76.1</td>
<td>76.0</td>
</tr>
<tr>
<td>OFA</td>
<td>20M</td>
<td>470M</td>
<td>76.6</td>
<td>76.4</td>
</tr>
<tr>
<td>GIT</td>
<td>800M</td>
<td>700M</td>
<td>77.1</td>
<td>77.0</td>
</tr>
<tr>
<td>BLIP2-G-OPT_2.7</td>
<td>129M</td>
<td>3.8B</td>
<td>77.8</td>
<td>77.5</td>
</tr>
<tr>
<td>Prismer</td>
<td>13M</td>
<td>1.6B</td>
<td>76.7</td>
<td>76.7</td>
</tr>
</tbody>
</table>
Git: a man riding a horse with a dog in the background.

OFA: a man riding on the back of a white horse

Prism: A man riding a horse next to a small dog.

BLIP2: a man riding a horse with a dog in the field

Ours: a man wearing a red hat and blue jeans rides a white horse with a long tail, while a small white dog follows closely behind

Git: a man riding a small motorcycle in a parking lot.

OFA: a man riding a motorcycle in a parking lot

Prism: A man riding a motorcycle in a parking lot.

BLIP2: a man riding a motorcycle in a parking lot with tents

Ours: a man wearing a white shirt and blue jeans rides a motorbike in a parking lot surrounded by white and yellow tents, with a white line marking the edge of the parking
## Large-Scale Dataset Influence

<table>
<thead>
<tr>
<th>Pre-training Data</th>
<th>Fine-tune + Test Data</th>
<th>B@4</th>
<th>CIDEr</th>
<th>SPICE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>Standard</td>
<td>37.8</td>
<td>126.5</td>
<td>22.9</td>
</tr>
<tr>
<td>FUSECAP</td>
<td>Standard</td>
<td>38.4</td>
<td>128.7</td>
<td>23.0</td>
</tr>
<tr>
<td>Standard</td>
<td>FUSECAP</td>
<td>35.4</td>
<td>111.4</td>
<td>25.0</td>
</tr>
<tr>
<td>FUSECAP</td>
<td>FUSECAP</td>
<td>37.3</td>
<td>123.1</td>
<td>26.8</td>
</tr>
</tbody>
</table>
Conclusion

- FUSECAP utilizes visual experts to extract meaningful information from images.
- LLM fuses the data into the existing captions to yield better captions.