Automatic Video Captioning

Pranav Prabhu

Background

- Automatic video captioning involves using multimodal media to generate captions
- **State of the Art Models:**
  - MaMMUT
  - VALOR
  - mPLUG-2
- Want to evaluate multiple video captioning models and compare them
- More comprehensive understanding of the benefits & drawbacks of different models
- Previously got some metrics for CNN-LSTM
- **Impact:** improve user experience, legal compliance, content indexing

Experiments (General Premise)

- Test multiple models
- Dataset: MSVD Dataset
- Run pre-trained models with testing data
- Adversarial examples
- Generate metrics
- Quantitative and qualitative analysis

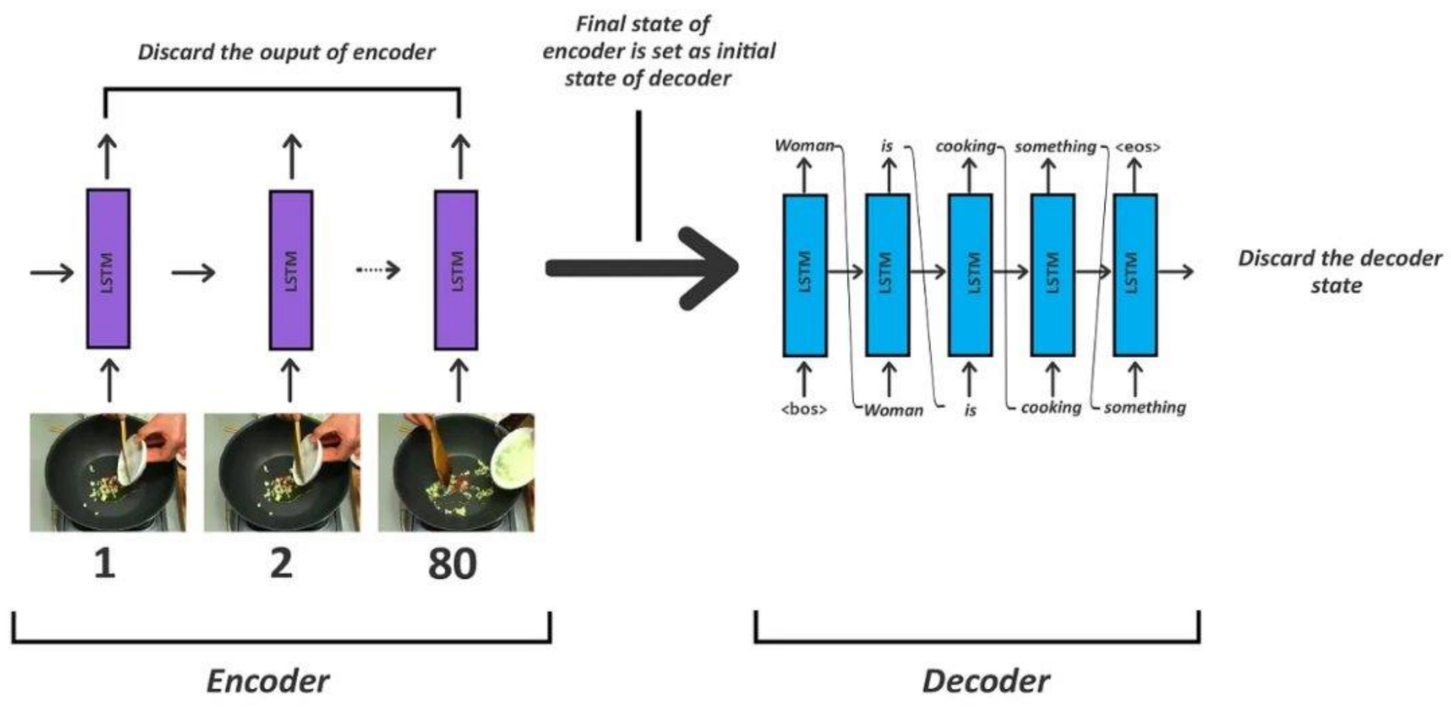
Dataset

- Microsoft Research Video Description **(MSVD)**
  - 1550 YouTube Clips
- Human-generated captions (avg 40-80 per video)
- Diverse dataset used to evaluate many state of the art video captioning models
- Standard split:
  - 1450 Training
  - 100 Testing

Models Run

- Chose models that were simpler to run due to resource restrictions
    - Explores model complexity vs. accuracy
- CNN-LSTM
    - Naïve approach
- LLaVA
    - Slightly more complex
    - Utilized online UI, QA based
- Comparative Analysis

CNN-LSTM



Inference model

```
train_path = "data/training_data"
test_path = "data/testing_data"
batch_size = 320
learning_rate = 0.0007
epochs = 150
latent_dim = 512
num_encoder_tokens = 4096
num_decoder_tokens = 1500
time_steps_encoder = 80
max_probability = -1
save_model_path = 'model_final'
validation_split = 0.15
max_length = 10
search_type = 'greedy'
```

LLaVA

1. Pretraining

Hyperparameter	Global Batch Size	Learning rate	Epochs	Max length	Weight decay
LLaVA-v1.5-13B	256	1e-3	1	2048	0

2. Finetuning

Hyperparameter	Global Batch Size	Learning rate	Epochs	Max length	Weight decay
LLaVA-v1.5-13B	128	2e-5	1	2048	0

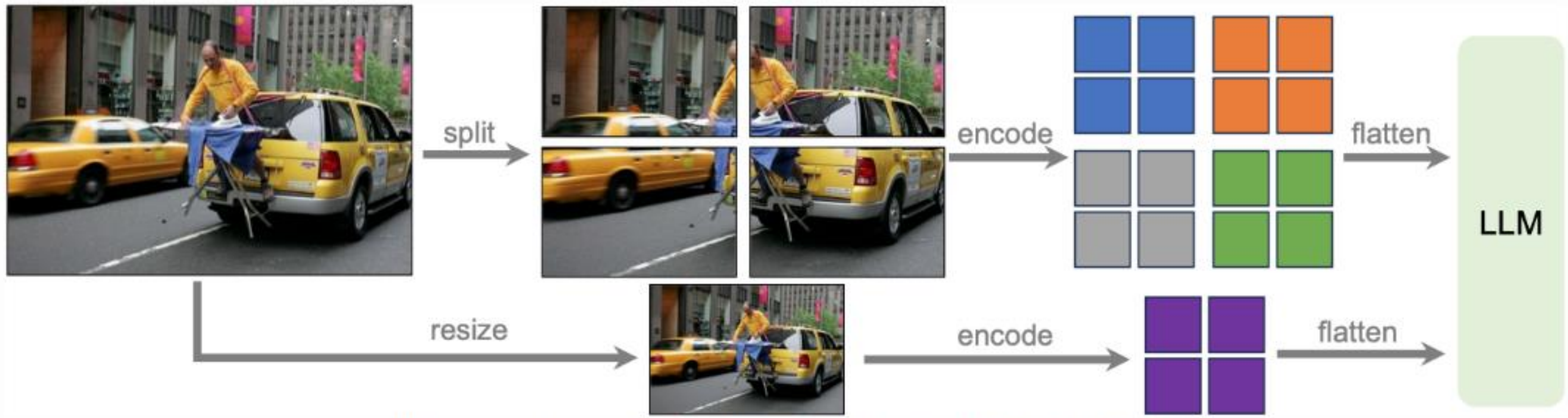


Illustration of dynamic high resolution scheme: a grid configuration of 2×2

Adversarial Examples

- **Method:** Removed a random frame(s) from each video (black screen)
- Re-tested both LLaVA model & CNN-LSTM
- **Goal:** To evaluate robustness of models with noisy examples
- Re-calculated metrics and compared them to original metrics

Sample Comparison of Models

MSVD	A man is putting salt on a chicken
CNN	A man is mixing a
CNN-A	A man is a a on a
LLaVA	A man in a red shirt is seen washing dishes in a kitchen.
LLaVA-A	A man in a red shirt is seen peeling potatoes in a kitchen.

Video ID: ScdUht-pM6s_53_63

Metrics

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR
CNN	75.09	35.56	19.58	6.47	30.31	48.03
CNN-A	74.60	33.05	15.98	3.80	30.23	46.19
LLaVA	55.93	25.74	11.02	5.07	24.26	60.01
LLaVA-A	57.17	26.52	11.16	4.27	23.94	60.29
VALOR	X	X	X	80.57	68.0	48.0
mPLUG-2	X	X	X	70.5	85.3	48.4

Analysis: LLaVA

- Performance on adversarial data set is only marginally worse
- LLaVA's higher METEOR score
  - Shows its strength in generating descriptions with better stemming and sentence structure.
- LLaVA's lower ROUGE-L score
  - Demonstrates limitations in creating longer video descriptions
- While some descriptions generated by LLaVA were very accurate, the most prevalent error was in object recognition
- Weaker performance than MPLUG-2 and VALOR except for the METEOR score.

Analysis: CNN-LSTM

- Performance on adversarial data set is only marginally worse
- CNN’s lower METEOR score
  - Suggests challenges in generating descriptions with proper stemming and sentence structure
- CNN’s higher ROUGE-L score
  - Implies superiority over LLaVA in creating longer sentence descriptions.
- While CNN’s generated descriptions are accurate, they often lack details.
- Weaker performance than MPLUG-2 and VALOR

Overall Findings

- Generally, adversarial results were only **slightly worse**
  - Models appear to be fairly robust against noise
- Metrics do not illustrate the full picture
  - Qualitatively, LLaVA is much better than the CNN-LSTM
  - Complete sentences that are closer to original MSVD dataset with higher accuracy in common nouns and key actions
- Complexity does help
  - The tradeoffs made by adding additional modules enable stellar performance
  - High METEOR scores for LLaVA suggest better linguistic quality despite lower ROUGE-L.

Implications & Future Work

- Prompt-based models versus video captioning style models may have differing quality in automatic video captioning
- Prompt engineering: can explore how different prompts can generate better quality analysis
- Running mPLUG-2, VALOR, and MaMMUT with training data and adversarial examples can provide better understanding of the pros and cons of existing models for the video captioning task
- Develop methods specifically targeting object recognition and detail inclusion in descriptions.

References

- [1] “Papers with Code - mPLUG-2: A Modularized Multi-modal Foundation Model Across Text, Image and Video,” paperswithcode.com. https://paperswithcode.com/paper/mplug-2-a-modularized-multi-modal-foundation.
- [2] “Papers with Code - VALOR: Vision-Audio-Language Omni-Perception Pretraining Model and Dataset,” paperswithcode.com. https://paperswithcode.com/paper/valor-vision-audio-language-omni-perception.
- [3] J. Su, “Study of Video Captioning Problem.” Available: https://www.cs.princeton.edu/courses/archive/spring18/cos598B/public/projects/LiteratureReview/COS598B_spr2018_VideoCaptioning.pdf
- [4] M. Abdar et al., “A Review of Deep Learning for Video Captioning.” [Online]. Available: https://arxiv.org/pdf/2304.11431.pdf
- [5] X. Wang, W. Chen, J. Wu, Y.-F. Wang, and W. Wang, “Video Captioning via Hierarchical Reinforcement Learning.” [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/papers/Wang_Video_Captioning_via_CVPR_2018_paper.pdf
- [6] J. Wang et al., “GIT: A Generative Image-to-text Transformer for Vision and Language.” Available: https://arxiv.org/pdf/2205.14100v5.pdf
- [7] “Papers with Code - EnCLAP: Combining Neural Audio Codec and Audio-Text Joint Embedding for Automated Audio Captioning,” paperswithcode.com. https://paperswithcode.com/paper/enclap-combining-neural-audio-codec-and-audio.
- [8] W. Kuo et al., “MaMMUT: A Simple Architecture for Joint Learning for MultiModal Tasks.” Accessed: Feb. 20, 2024. [Online]. Available: https://arxiv.org/pdf/2303.16839v3.pdf
- [9] Shreya, “Shreyz-max/Video-Captioning,” GitHub, Apr. 01, 2024. https://github.com/Shreyz-max/Video-Captioning.
- [10] vsubhashini, “vsubhashini/caption-eval,” GitHub, May 28, 2023. https://github.com/vsubhashini/caption-eval.
- [11] D. Raj, “Metrics for NLG evaluation,” Explorations in Language and Learning, Sep. 16, 2017. https://medium.com/explorations-in-language-and-learning/metrics-for-nlg-evaluation-c89b6a781054.
- [12] “Consensus-based Image Description Evaluation (CIDEr),” oecd.ai. https://oecd.ai/en/catalogue/metrics/consensus-based-image-description-evaluation-%28cider%29.
- [13] “MSVD Dataset Corpus,” www.kaggle.com. https://www.kaggle.com/datasets/vtrnanh/msvd-dataset-corpus.