SJSU SAN JOSÉ STATE UNIVERSITY

Show and Tell: A Neural Image Caption Generator

Sai Anoushka K

OBJECTIVE

- Automatically generate natural language descriptions from images.
- Bridges **Computer Vision** (image understanding) and **Natural Language Processing** (caption generation).

Architecture Overview

Architecture:

- **CNN (Convolutional Neural Network)**: Extracts visual features from images.
- **LSTM (Long Short-Term Memory)**: Generates a sequence of words based on image features.

How It Works

- The CNN encodes the image into a feature vector.
- The LSTM takes the feature vector and generates a sentence word-by-word.
- The model is trained end-to-end to maximize the likelihood of the correct caption.

Key Results

- State-of-the-art performance on the **COCO dataset** with a **BLEU-4 score of 27.7**.
- Generates grammatically correct, contextually accurate captions.
- Outperforms previous methods on automatic metrics (e.g., BLEU, METEOR).

Applications

- Assistive technologies for the visually impaired (image description).
- Automated image captioning for social media and content management systems.

Conclusion

- The "Show and Tell" model effectively combines **visual understanding** and **language generation** in a single, end-to-end architecture.
- Has broad applications in AI, from assistive tools to content automation.


Thank You!