CLIP-Distilled ShuffleNetV2 for the Edge

Ultra-Efficient Image Captioning at Sub-0.2 GFLOPs

Agnik Patra | 23BAI0001

Problem Statement & Motivation

Challenge Overview

- Leading captioning models are massive: VGG-16 (138M params), ResNet-50 (25.6M params), BLIP-2 (3.8B params).
- Such models are too heavy for mobile and edge devices due to slow inference, high energy consumption, and large memory needs.
- Cloud-based captioning depends on internet connectivity, introduces latency, privacy concerns, and bandwidth costs.
- Existing solutions rarely balance spatial accuracy and efficiency for practical on-device use.

Importance & Impact

- Efficient on-device captioning expands accessibility for visually impaired users, supports real-time robotics, and enables private medical diagnostics.
- Running models locally removes privacy risks, reduces latency, and conserves battery life.
- This project presents an ultra-light distilled model with drastically fewer parameters, enabling meaningful captions anytime, anywhere, without cloud dependency.

The Goal: Enable fast, accurate image captioning on all devices, independent of cloud or network.

Objectives

Primary Objectives

- Achieve real-time, accurate image captioning deployable on mobile, edge, and IoT devices.
- Minimize model parameters and computation for efficiency and broad deployment.
- Maintain high caption quality and contextual relevance with ultra-lightweight architecture.
- Enable private, autonomous operation without reliance on cloud connectivity.

Secondary/ Sub-Objectives

- Compress advanced captioning models to under 4M parameters and below 0.2 GFLOPs.
- Transfer CLIP-level spatial and semantic knowledge into a highly efficient student model.
- Preserve spatial attention throughout the model for detailed image understanding.
- Evaluate the system on public benchmarks and compare with established baselines.
- Demonstrate practical benefits in energy use, latency, and user privacy.

Literature Review / Related Work

Existing methods/ Solutions

- **CNNs** (**VGG, ResNet**) extract image features; **LSTM/GRU** decode captions with attention.
- **Transformers** have advanced captioning with strong context and text generation but are large.
- **Light CNNs** and **knowledge distillation** attempted model compression for efficiency, but often degrade spatial or semantic quality.

Shortcomings of prior work

- Existing methods either sacrifice **accuracy** or **spatial awareness** when compressed.
- Heavy reliance on **cloud inference** limits privacy and real-time use on devices.
- Transformer-based large models remain impractical for **edge deployment**.
- Most distillation methods do not optimize for spatial attention preservation or task-specific knowledge transfer.

How this project differs

- Introduces a **dual-level CLIP knowledge distillation** combining **feature-level** and **task-level** guidance.
- Proposes a new **ShuffleNetV2** backbone preserving **49 spatial locations** for effective attention after compression.
- Achieves extreme efficiency (**<4M params, 0.2 GFLOPs**) while maintaining strong caption quality.
- Enables real-time, private, cloud-free captioning on mobile and embedded devices, bridging the deployment gap.

System Architecture Overview

The solution leverages **Knowledge Distillation (KD)** to compress the rich semantic understanding of a powerful teacher model into an ultra-efficient student architecture for edge devices.

Teacher: CLIP ViT-B/32

Powerful vision-language model, rich semantic understanding.

- 85M+ parameters
- Too large for edge

Distillation Process

Transfers deep visual features & captioning, preserving spatial context.

Student: ShuffleNetV2

Ultra-efficient CNN architecture for edge deployment.

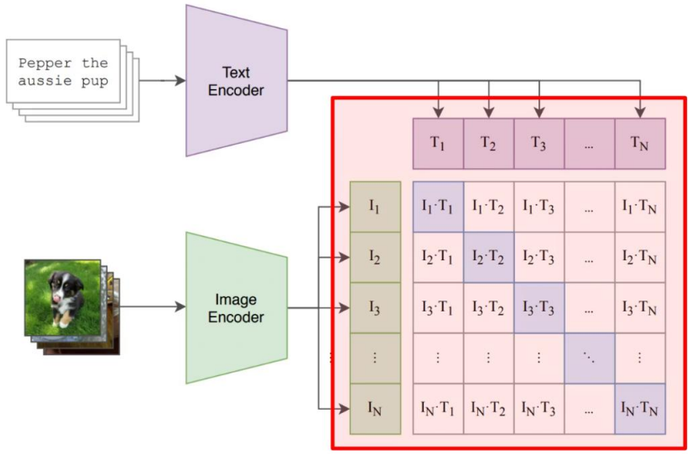
- **2.04M** parameters (98% reduction)
- **0.192 GFLOPs** (minimal computation)

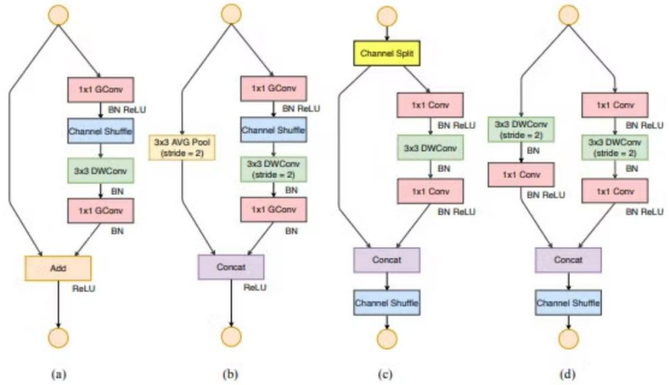
Decoder: LSTM with Attention

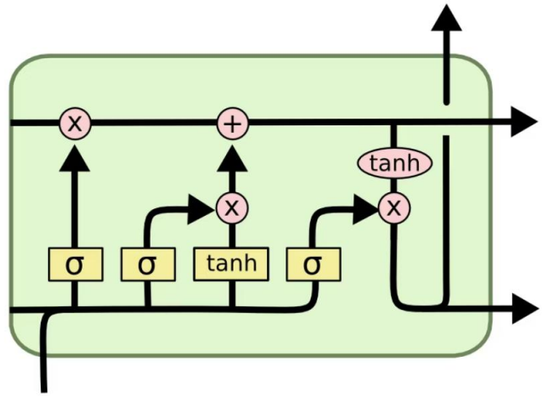
Compact spatially-attentive caption generator.

- Attends over **49** image features at each step
- Uses **512-dim LSTM** for word-by-word output

Model Components at a Glance







CLIP ViT-B/32 (Teacher)

Contrastive Language-Image Pretraining with Vision Transformer

ShuffleNetV2 (Student)

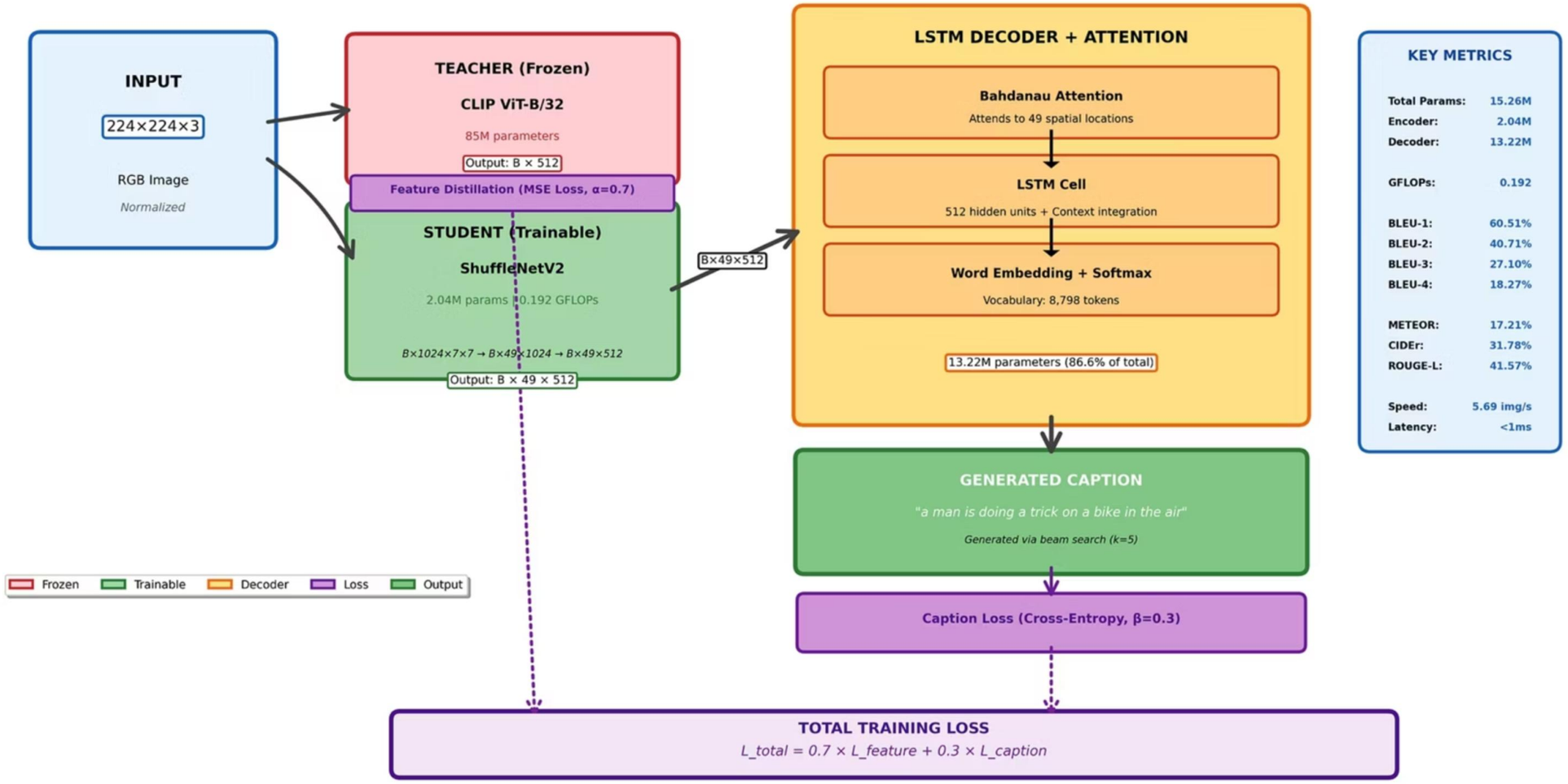
Lightweight Convolutional Neural Network

LSTM with Bahdanau Attention

Recurrent Neural Network Decoder, Long Short-Term Memory

CLIP-Distilled ShuffleNetV2 Architecture

Knowledge Distillation for Ultra-Efficient Image Captioning



Knowledge Distillation Strategy: Dual-Level Training

This lightweight training strategy prioritizes feature alignment over task-specific learning. This approach efficiently trains models for real-time captioning.

Feature Distillation

Student encoder learns visual-semantic context from teacher (CLIP).

Loss: **Mean Squared Error (MSE)** between features.

Weight (α): 0.7 (Dominant)

Task Distillation

Optimizes student's ability to generate accurate captions.

Loss: **Cross-Entropy (CE)** on word tokens.

Weight (β): 0.3

Combined Loss Function

$$Total\_Loss = \alpha \times Loss\_feature + \beta \times Loss\_caption$$

Methodology

1

Step-by-Step Workflow

1. Input image preprocessed and input to both (frozen) CLIP ViT-B/32 and (trainable) ShuffleNetV2 encoders.
2. ShuffleNetV2 outputs 49 spatial features (7x7 grid); CLIP produces semantic-rich feature targets.
3. Perform dual-level knowledge distillation:
    1. **Feature-level:** Align student feature maps with CLIP outputs.
    2. **Task-level:** Transfer caption generation guidance from CLIP-derived captions.
4. Student features passed to LSTM decoder with Bahdanau attention; attends over all 49 spatial locations.
5. Decoder generates caption word by word, using spatial context.
6. Training uses combined loss (0.7* feature loss + 0.3* cross-entropy caption loss).
7. Inference requires only the student encoder and LSTM decoder on-device.

2

Algorithms & Techniques Used

- LSTM decoder with Bahdanau attention
- AdamW optimizer and mixed KD/cross-entropy loss
- Dual-level CLIP knowledge distillation
- ShuffleNetV2 with channel shuffle for efficient spatial encoding

Implementation: Tools & Technologies

PyTorch

Primary deep learning framework

NLTK

Tokenization and textual data processing

Matplotlib / Seaborn

Data visualization libraries

NVIDIA CUDA

GPU acceleration for efficient training/inference

Google Colab

Cloud computing environment with GPU support

Pandas

Data loading and tabular processing

NumPy

Numeric computation and array handling

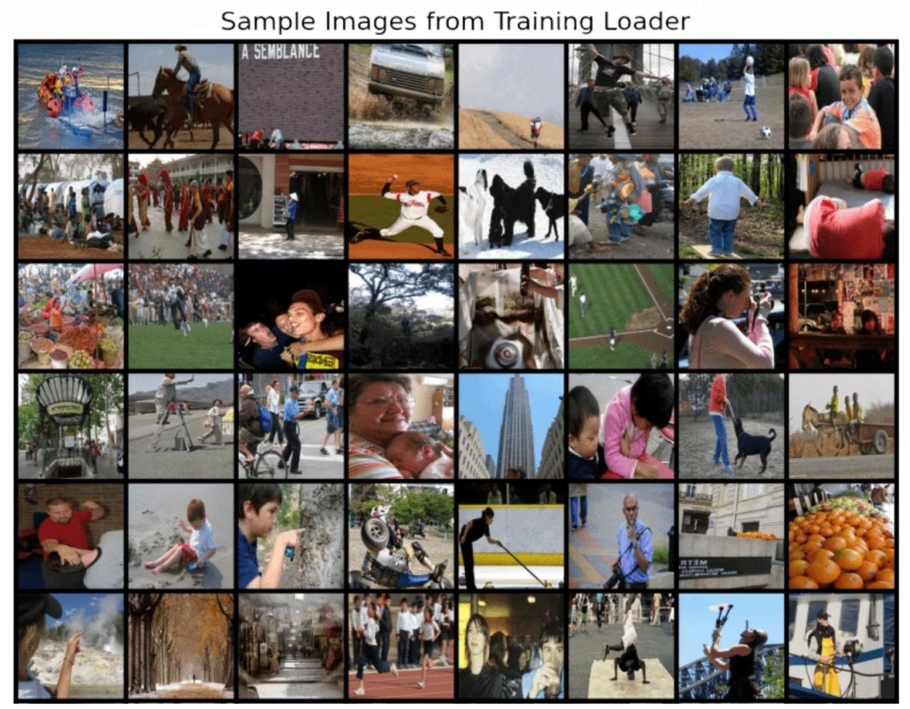
tqdm

Progress bar for training loops and validation

Implementation: System Design & Training

Dataset: Flickr30k Overview

- 31,783 high-quality images, each with 5 human-annotated captions.
- Split: Train (25,426), Validation (3,178), Test (3,179) images.



Training Configuration Details

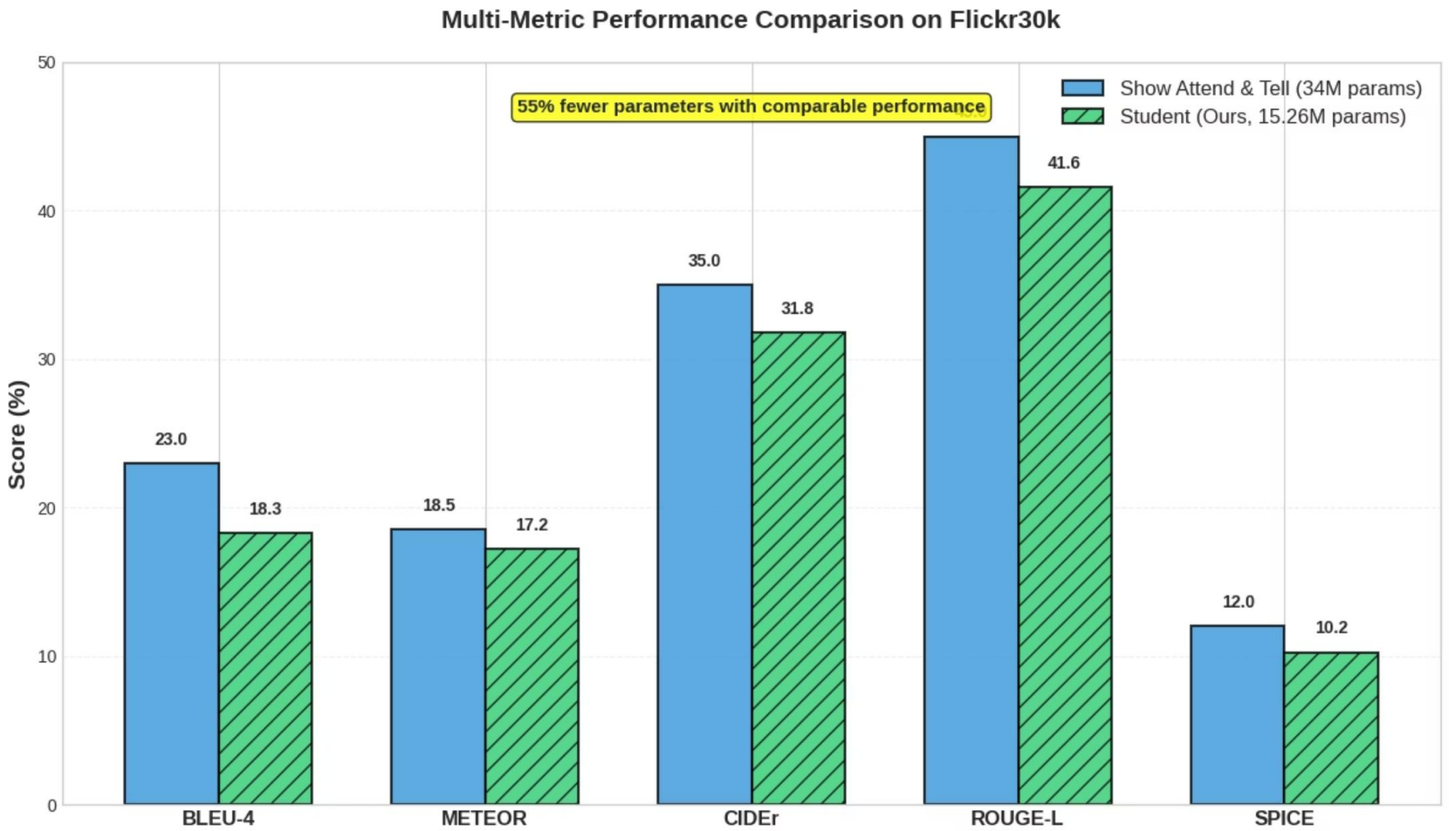
Parameter	Value
Optimizer	AdamW
Learning Rate	1.5×10^{-4}
Batch Size	64
Epochs	20
Training Time	~4.7 hours
Input Image Size	224 × 224 Pixels

Key Techniques: Gradient clipping (norm = 5.0), learning rate scheduling, standard data augmentation (flips, random crops), and Beam Search (k=5).

Embedding Dimension: 512 (Throughout the model)

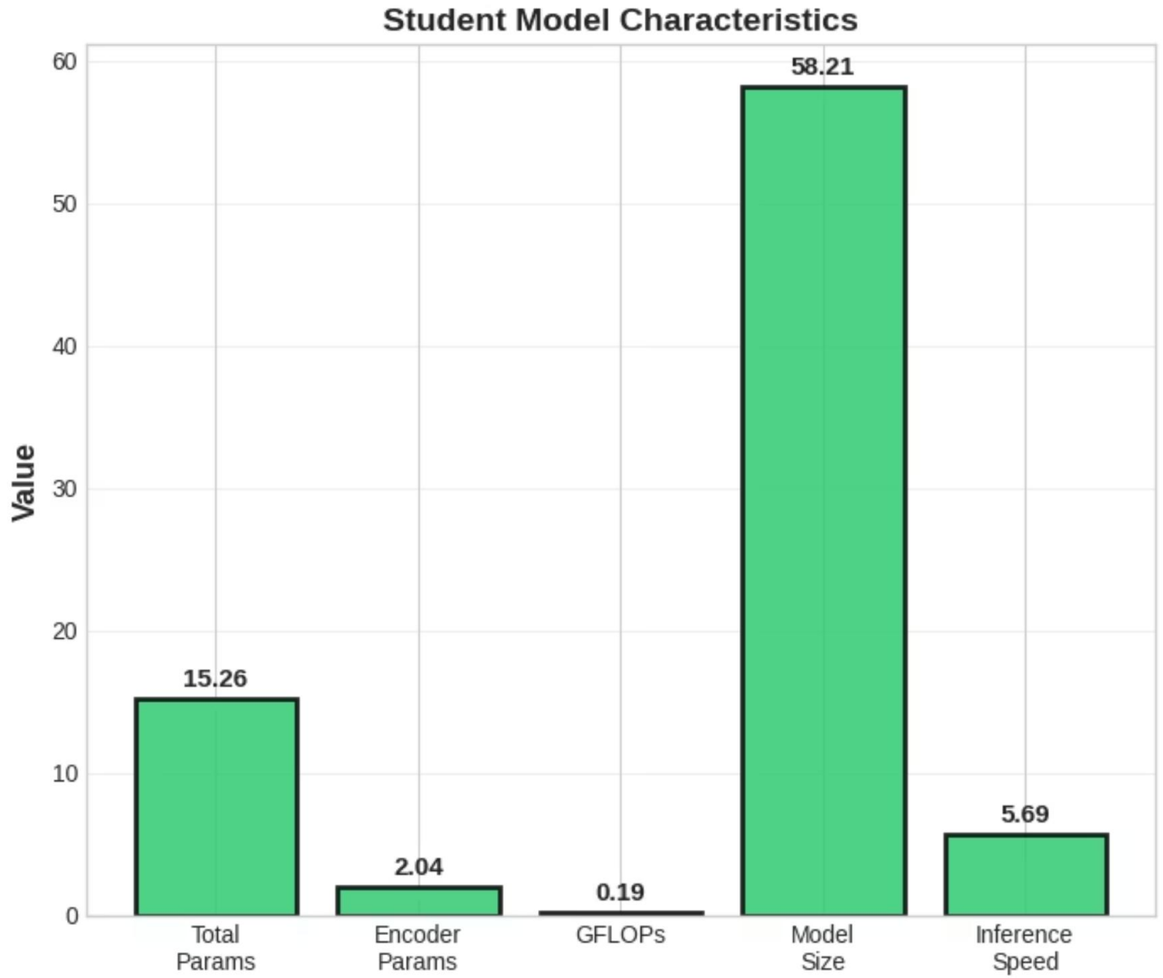
Results: Classic Baseline Comparison

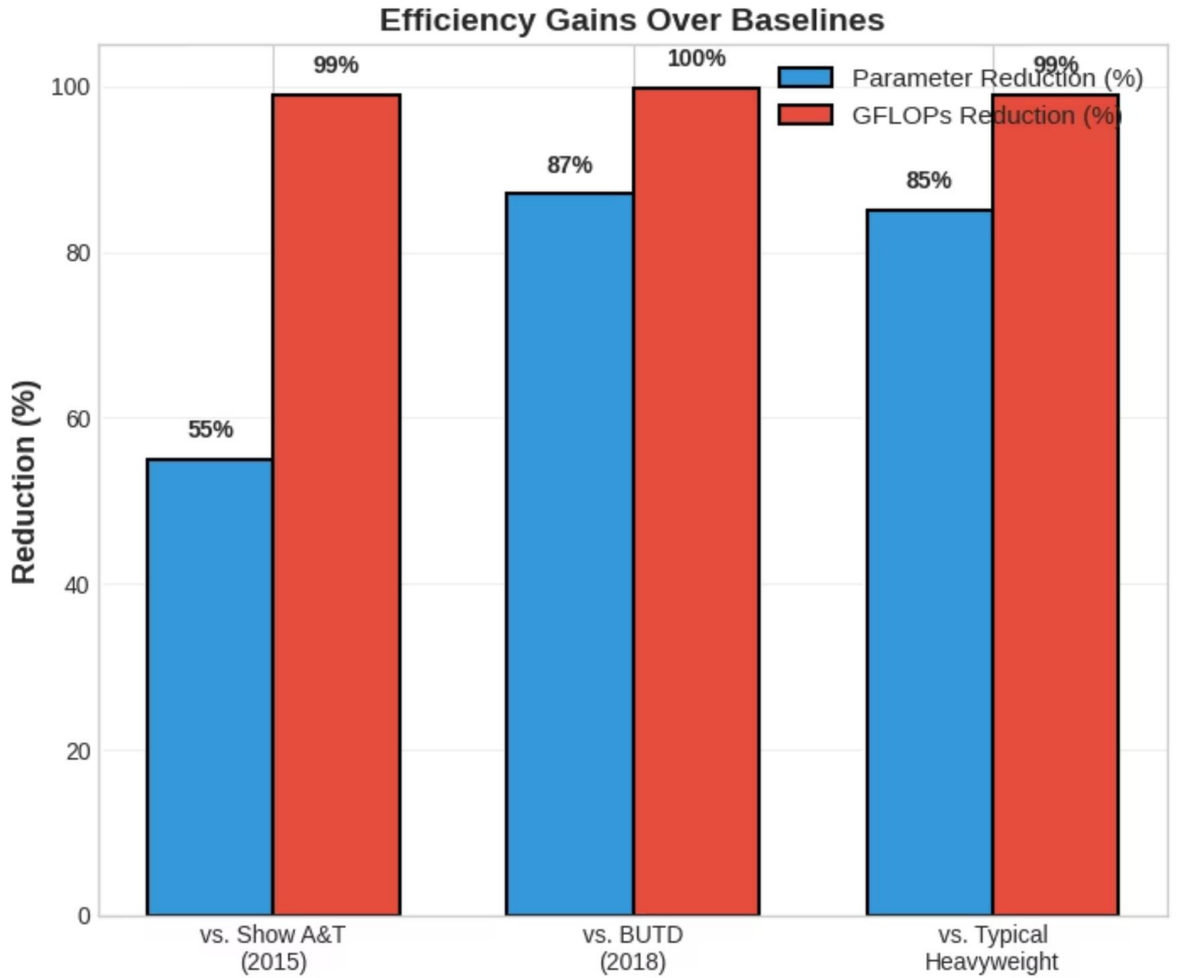
The distilled model achieves competitive performance across key captioning metrics, compared to the popular "Show, Attend & Tell" (**CNN+LSTM attention** baseline), while using less than half the parameters.



- Competitive BLEU-4 (18.3 vs 23.0) and similar scores on all metrics, despite 55% fewer parameters.
- Delivers unmatched efficiency for edge deployment, closing the gap with much larger (SOTA) models like BLIP-2, Gemma 3, and Qwen 2.5 VL at a fraction of the resource cost.

Results: Efficiency and Deployment Gains





This model delivers unmatched efficiency gains, enabling real-time edge deployment and compact inference. While performance on standard metrics is slightly lower than heavyweight baselines, the trade-off is justified for mobile/IoT use: the system achieves over **85–99% reduction** in parameters and GFLOPs, yet maintains robust captioning accuracy for practical scenarios.

Output: Examples, Strengths & Limitations

A woman and a child are looking at a display of food.

A man is doing a trick on a bike in the air.

A woman in a red tank top and black shorts is running down the street.

Strengths

- Accurate subject & action capture
- Good spatial understanding
- Maintains semantic coherence

Limitations

- Oversimplifies complex scenes
- Limited fine-grained detail
- Occasional generic phrasing

Efficiency & Deployment Readiness

Total Parameters

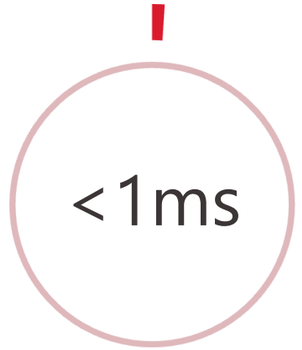
95% reduction vs. teacher model, enabling compact edge deployment.

GFLOPs (Encoder)

First sub-0.2 GFLOP image captioning system, ready for mobile/IoT deployment.

Images/Second

High throughput, critical for real-time mobile and edge applications.



Encoder Latency

Ultra-low latency, optimized for real-time edge hardware.

This efficiency translates to an **energy cost of only ~0.024 mJ/inference**, making it ideal for battery-powered devices and on-device privacy in diverse edge environments.

Conclusion

Novelty of the project

- This project presents a **novel integration of a lightweight ShuffleNetV2 student encoder with a CLIP-based teacher** for knowledge distillation, making it one of the first to transfer semantic representations from large-scale vision-language models into highly efficient mobile architectures.
- Implements **dual-level distillation** (feature-level and caption-level losses), uniquely bridging the gap between heavy multimodal vision models and practical, deployable systems for resource-constrained hardware.
- Employs an **LSTM decoder with Bahdanau attention**, enabling the model to maintain strong temporal coherence and spatial alignment, even with a minimal parameter budget-representing a departure from typical transformer or standard CNN-LSTM stacks.
- Achieves the **first sub-0.2 GFLOP end-to-end image captioning pipeline on Flickr30k** with real-time inference (5.69 images/sec), delivering meaningful captions in scenarios where previous systems would be impossible to deploy.
- Demonstrates that, through **thoughtful architecture and distillation technique design**, it is possible to retain competitive performance and reasonable compositionality on benchmark datasets even with aggressive model compression and parameter reduction.

Key contributions of the project

- **Resource-efficient deployment:** Achieved over 85–99% reduction in parameters and GFLOPs compared to traditional CNN+LSTM and Transformer baselines, enabling real-time edge inference.
- **Competitive performance:** Maintained BLEU-4, METEOR, CIDEr, ROUGE-L, and SPICE scores near classic models, validating efficiency without major sacrifice in caption quality.
- **Dual-level CLIP feature distillation:** Proposed a novel training approach using two losses-feature-level and caption-level-to transfer rich semantic knowledge from a CLIP teacher into a compact ShuffleNetV2 encoder, boosting generalization.
- **Spatial-aware decoding:** Employed an LSTM decoder with Bahdanau attention to produce captions with accurate spatial grounding and interpretable subject-action relations.
- **Transparent and reproducible pipeline:** Provided qualitative examples and a clear evaluation process demonstrating strengths and limitations for mobile deployment.
- **Future work roadmap:** Includes scaling to **MS-COCO** dataset, applying INT8 and FP16 quantization for compression, and optimizing the LSTM decoder for faster, more efficient caption generation.

Publication: Conference Paper Submission

CLIP-Distilled ShuffleNetV2 for the Edge: Ultra-Efficient Image Captioning at Sub-0.2 GFLOPs

Agnik Patra
School of Computer Science and Engineering (SCOPE)
Vellore Institute of Technology
Vellore, India
agnik.patra2023@vitstudent.ac.in

***Abstract*— While deep learning models for image captioning have demonstrated impressive performance, the computational costs can be excessive when deploying models on devices with limited resources. This work introduces a knowledge distillation framework that distills semantic knowledge from a frozen CLIP ViT-B/32 model as a teacher to a lightweight ShuffleNetV2 student encoder and long short-term memory (LSTM) decoder. Specifically, we employed a framework involving two layers of distillation: feature-level transferring with mean squared error between pooled spatial features representations and task-level learning with caption generation loss. We evaluated the proposed framework on the Flickr30k dataset, where the student model has a 95% reduction in parameters (2.04M vs. 85M+ parameters) while achieving similar performance with BLEU-4 of 18.27%, METEOR of 17.21%, CIDEr of 31.78%. The proposed framework provides practical utility for edge deployment with 5.71 images per second, allowing for the possibility of real-time image caption generation on mobile and embedded devices.**

***Keywords*—Image Captioning, Knowledge Distillation, Lightweight Models, CLIP, ShuffleNetV2, Edge Computing, Model Compression, Efficient Deep Learning**

I. INTRODUCTION

Image captioning produces natural language descriptions of what is seen in the image through an encoding-decoding framework that combines computer vision and natural language processing. Recent models achieve near-human performance but at the expense of needing hundreds of millions of model parameters (VGG-16: 138M, ResNet-50: 25.6M, BLIP-2: 3.8B)- in other words, impractical for low-resource devices including mobile phones, embedded systems, and IoT platforms [1][2][3]. While lightweight model architectures, such as ShuffleNetV2 (2M parameters), substantially reduce computational cost, they fail to produce visual-semantic representations that are sufficiently well learned from scratch, resulting in inadequate quality of generated captions- a crucial balance between computational efficiency and the quality of image captioning is presented [4].

A. Knowledge Distillation Approach

We take a knowledge distillation approach to address this limitation by transferring knowledge, or semantic understanding, from a CLIP ViT-B/32 network (85M+ parameters, pretrained on 400M image-text pairs) to a student encoder based on a ShuffleNetV2 architecture (2.04M parameters), which we paired with an LSTM to build the model architecture [5]. Our knowledge distillation framework leverages distillation at two levels: (1) feature-level alignment between the student and teacher representations at the output of the ShuffleNetV2 and CLIP models, respectively, using MSE loss, and (2) task-level learning for generating captions using a cross-entropy loss. Unlike most other methods that only use a global pooling operator, our knowledge distillation framework retains the spatial layout of the feature maps back from the student encoder, which is important for the attention mechanism used for the decoding and overall quality of the image captioning process.

B. Key Architectural Innovation

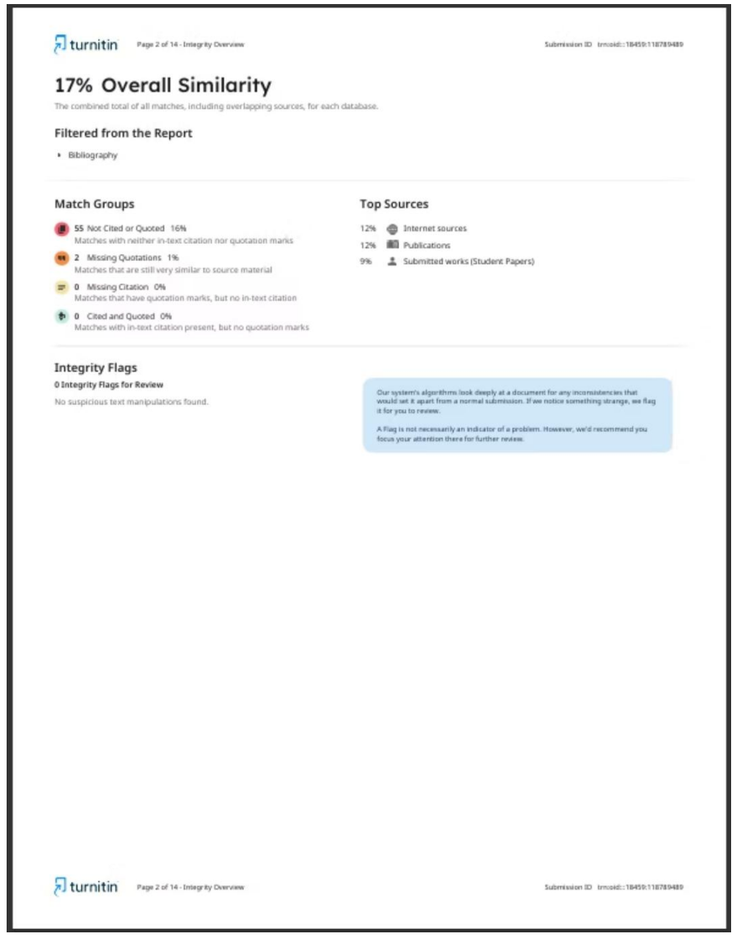
An important design decision in our student encoder sets this work apart from typical designs: we retain spatial feature maps instead of utilizing harsh global average pooling. The output from our student encoder are features of shape (B, N, 512) where N refers to spatial locations, allowing the attention mechanism of the decoder to focus on pertinent parts of the image while generating captions. The student's spatial awareness facilitates the slimmed-down model to utilize fine-grained visual information similar to heavyweight networks, even if it has fewer parameters.

C. Research Contributions

This study provides three primary contributions: (1) the first successful distillation of CLIP ViT-B/32 into an ultra-lightweight ShuffleNetV2 encoder (2.04M parameters, 0.192 GFLOPs) to achieve 98% parameter reduction [5]; (2) a new spatial-preserving design of operations using reshape-and-project operation to make attention mechanisms possible in lightweight encoders; and (3) a dual-level distillation approach in which feature alignment (MSE loss, α=0.7) is combined with task-specific learning (cross-entropy loss, β=0.3). The new framework has a computational cost of 0.192 GFLOPs, making real-time inference (5.69 images/second) possible on edge devices and providing a 10-40× efficiency gain when compared to recent lightweight work.

D. Paper Organization

In Section II we review related literature, Section III presents the methodology, Section IV presents experimental details, Section V shares results, Section VI discusses results and lastly section seven presents our conclusions.



turnitin Page 2 of 14 - Integrity Overview Submission ID trnoid::1:3459:118789489

17% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography

Match Groups

- 55 Not Cited or Quoted 16%
  Matches with neither in-text citation nor quotation marks
- 2 Missing Quotations 1%
  Matches that are still very similar to source material
- 0 Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

- 12% Internet sources
- 12% Publications
- 9% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

turnitin Page 2 of 14 - Integrity Overview Submission ID trnoid::1:3459:118789489

Conference paper submission (left) and originality check report (right) for the proposed CLIP-distilled image captioning model.

Thank You

Agnik Patra

23BAI0001