Title : Design and Evaluation of Lightweight Architectures for Single Sentence Video Captioning

Under the Guidance of
Assistant Prof. Dr. Parijat Bhowmick

Kuldeep (234156024)
M.Tech CICPS

Outline

- Introduction
- Motivation
- Datasets Used
- Proposed Pipeline
- Decoder Models
- Key Frame Selection
- Feature Extractors Compared
- Results
- Conclusion
- Future Work

Introduction

Motivation for Lightweight Models

- Heavy models → slow + high compute
- Need real-time captioning
- Edge devices (mobile, CCTV, drones)
- Focus on CNNs like MobileNet, ShuffleNet

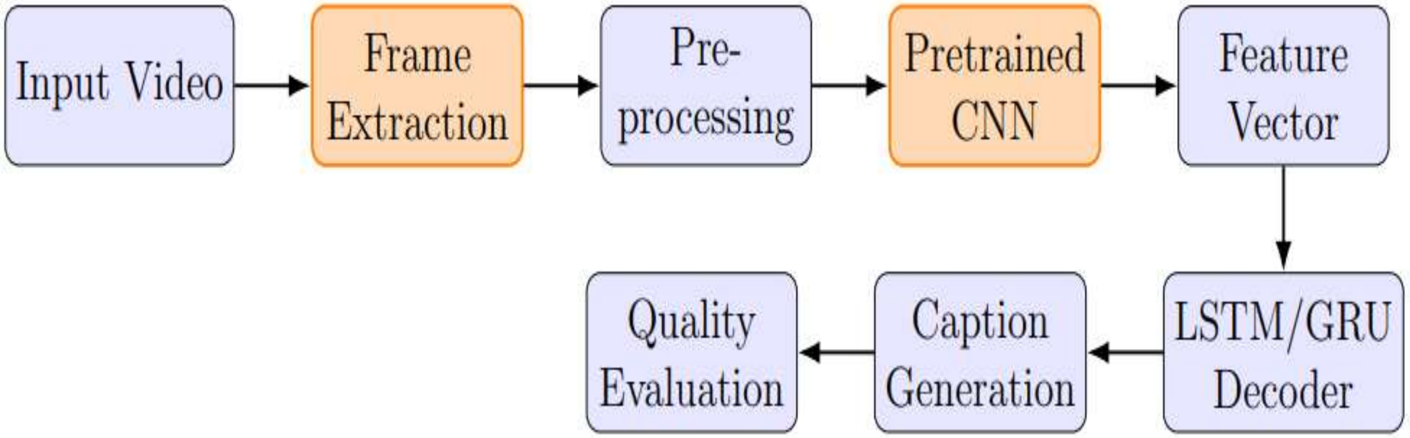
Challenges

High visual complexity
Temporal reasoning
Semantic ambiguity
Information bottleneck (1 sentence only)
Noisy captions in datasets

Datasets Used

- MSVD dataset(1970 Videos clip)
- MSR-VTT dataset(7010 Videos clip)
- Short videos + multiple captions

Proposed Pipeline



Input Video

Frame Extraction

Pre-processing

Pretrained CNN

Feature Vector

LSTM/GRU Decoder

Caption Generation

Quality Evaluation

Decoder Models

- LSTM
- GRU
- Decoder Base Transformer
- Compare performance

Experiment Frame Selection Strategy

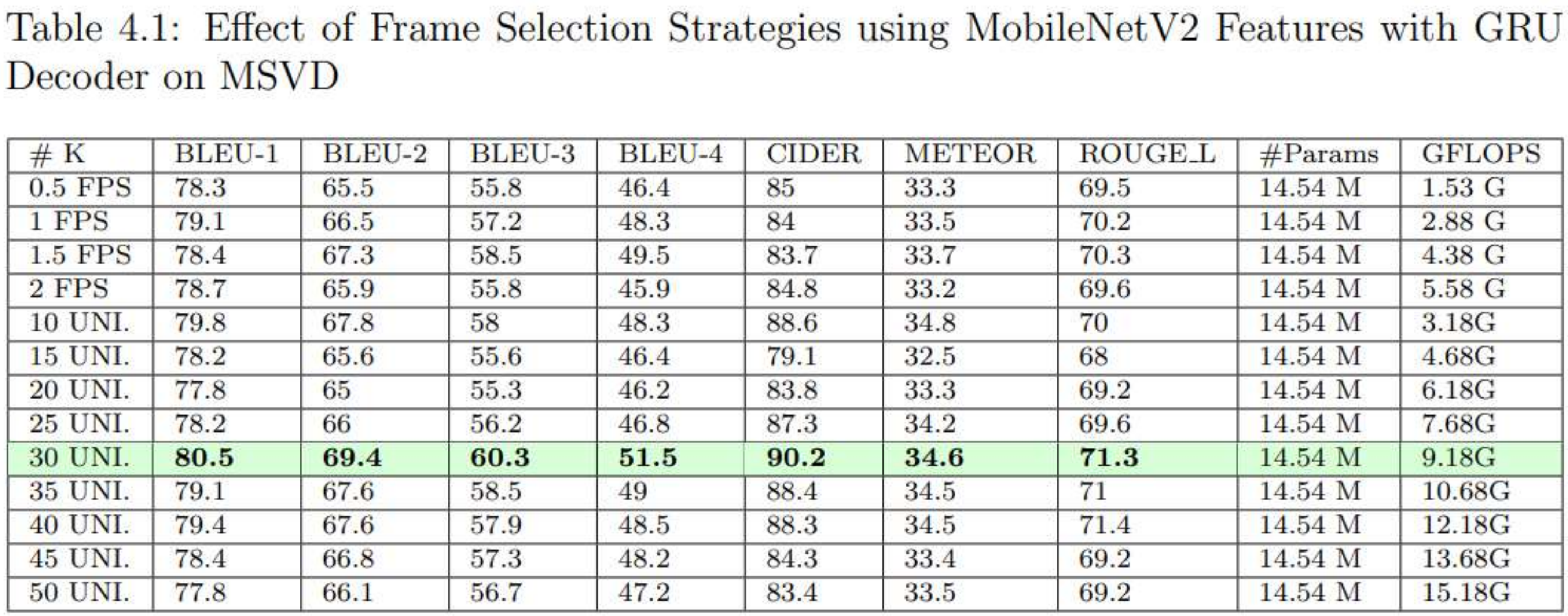


Table 4.1: Effect of Frame Selection Strategies using MobileNetV2 Features with GRU Decoder on MSVD

# K	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDER	METEOR	ROUGE_L	#Params	GFLOPS
0.5 FPS	78.3	65.5	55.8	46.4	85	33.3	69.5	14.54 M	1.53 G
1 FPS	79.1	66.5	57.2	48.3	84	33.5	70.2	14.54 M	2.88 G
1.5 FPS	78.4	67.3	58.5	49.5	83.7	33.7	70.3	14.54 M	4.38 G
2 FPS	78.7	65.9	55.8	45.9	84.8	33.2	69.6	14.54 M	5.58 G
10 UNI.	79.8	67.8	58	48.3	88.6	34.8	70	14.54 M	3.18G
15 UNI.	78.2	65.6	55.6	46.4	79.1	32.5	68	14.54 M	4.68G
20 UNI.	77.8	65	55.3	46.2	83.8	33.3	69.2	14.54 M	6.18G
25 UNI.	78.2	66	56.2	46.8	87.3	34.2	69.6	14.54 M	7.68G
30 UNI.	**80.5**	**69.4**	**60.3**	**51.5**	**90.2**	**34.6**	**71.3**	14.54 M	9.18G
35 UNI.	79.1	67.6	58.5	49	88.4	34.5	71	14.54 M	10.68G
40 UNI.	79.4	67.6	57.9	48.5	88.3	34.5	71.4	14.54 M	12.18G
45 UNI.	78.4	66.8	57.3	48.2	84.3	33.4	69.2	14.54 M	13.68G
50 UNI.	77.8	66.1	56.7	47.2	83.4	33.5	69.2	14.54 M	15.18G

Experiment of Frame Selection

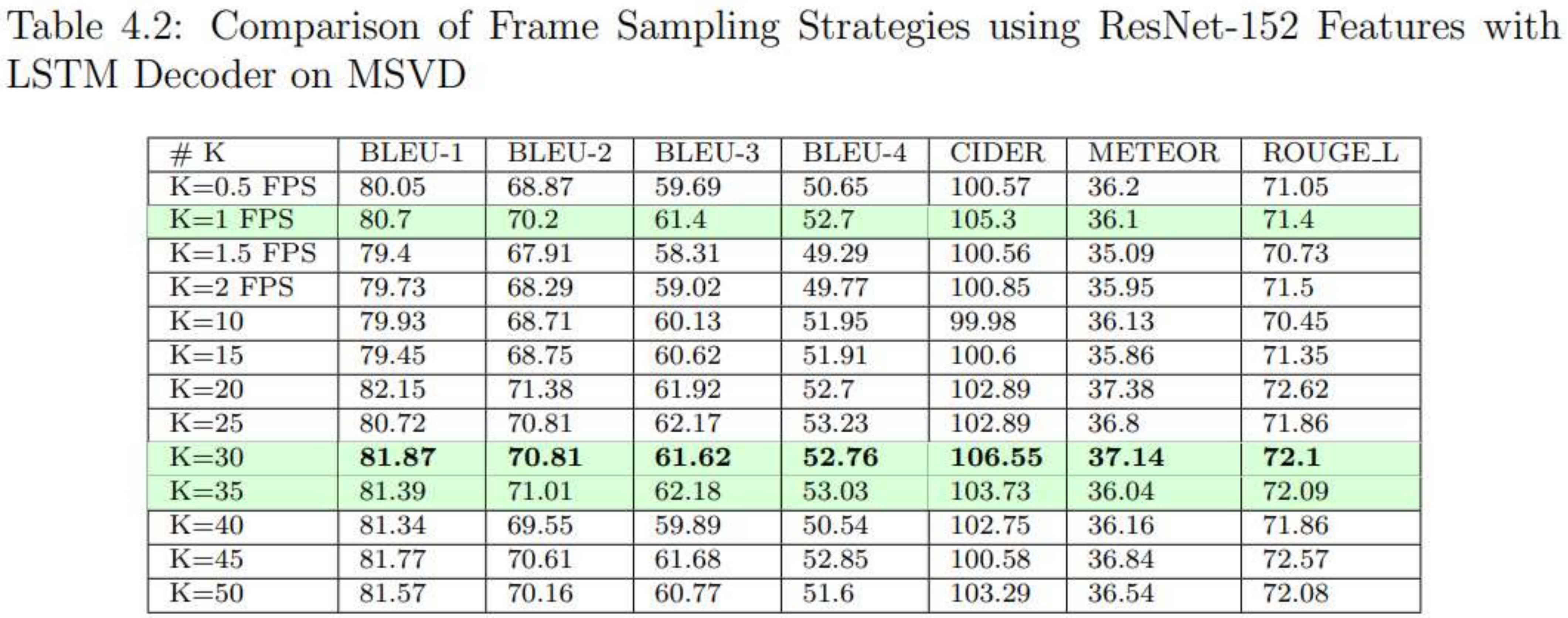


Table 4.2: Comparison of Frame Sampling Strategies using ResNet-152 Features with LSTM Decoder on MSVD

# K	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDER	METEOR	ROUGE_L
K=0.5 FPS	80.05	68.87	59.69	50.65	100.57	36.2	71.05
K=1 FPS	80.7	70.2	61.4	52.7	105.3	36.1	71.4
K=1.5 FPS	79.4	67.91	58.31	49.29	100.56	35.09	70.73
K=2 FPS	79.73	68.29	59.02	49.77	100.85	35.95	71.5
K=10	79.93	68.71	60.13	51.95	99.98	36.13	70.45
K=15	79.45	68.75	60.62	51.91	100.6	35.86	71.35
K=20	82.15	71.38	61.92	52.7	102.89	37.38	72.62
K=25	80.72	70.81	62.17	53.23	102.89	36.8	71.86
K=30	**81.87**	**70.81**	**61.62**	**52.76**	**106.55**	**37.14**	**72.1**
K=35	81.39	71.01	62.18	53.03	103.73	36.04	72.09
K=40	81.34	69.55	59.89	50.54	102.75	36.16	71.86
K=45	81.77	70.61	61.68	52.85	100.58	36.84	72.57
K=50	81.57	70.16	60.77	51.6	103.29	36.54	72.08

Experiment of Frame Selection

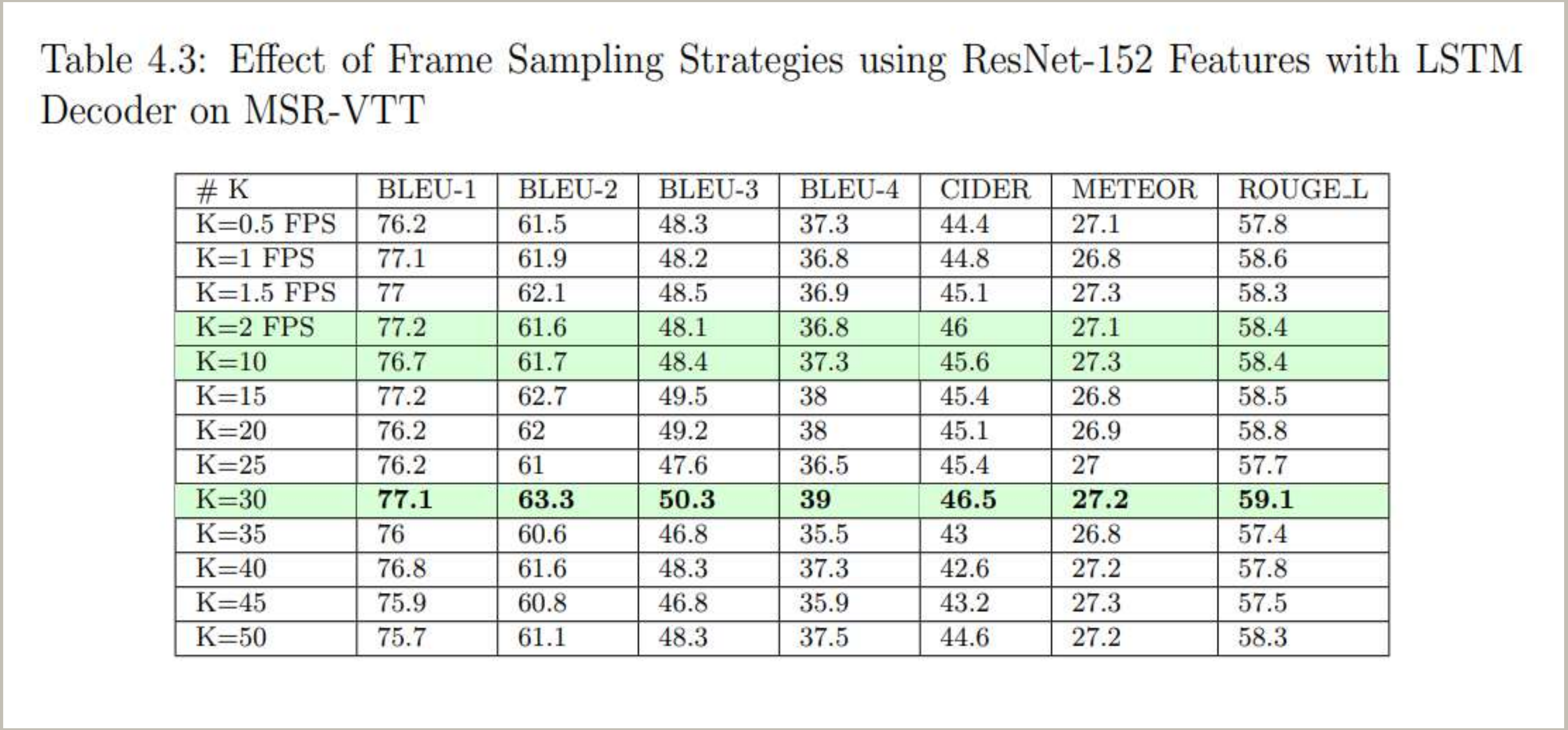
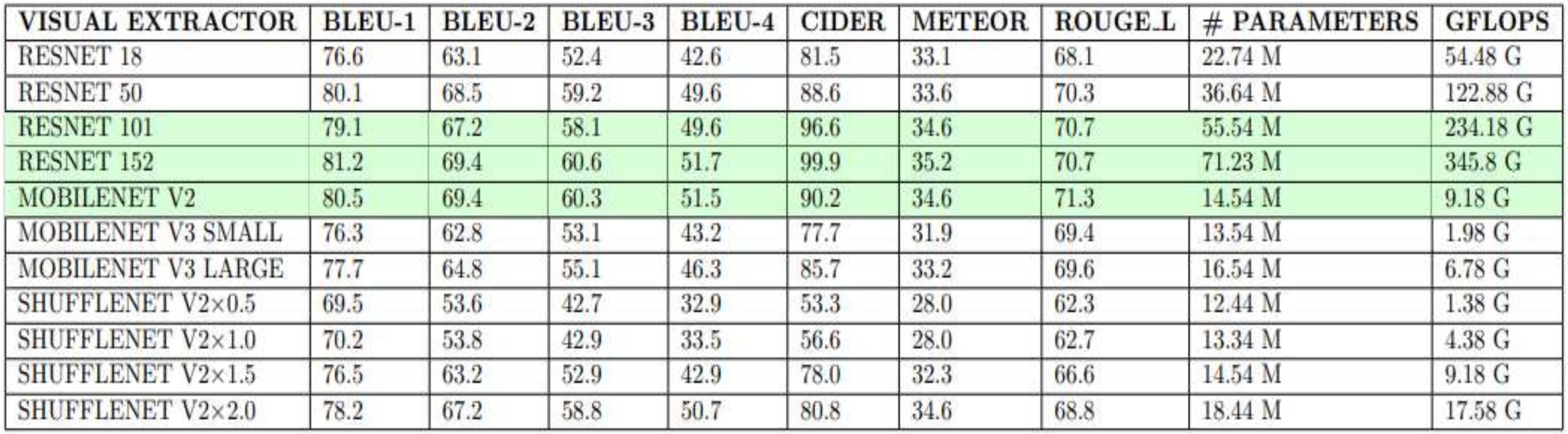


Table 4.3: Effect of Frame Sampling Strategies using ResNet-152 Features with LSTM Decoder on MSR-VTT

# K	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDER	METEOR	ROUGE_L
K=0.5 FPS	76.2	61.5	48.3	37.3	44.4	27.1	57.8
K=1 FPS	77.1	61.9	48.2	36.8	44.8	26.8	58.6
K=1.5 FPS	77	62.1	48.5	36.9	45.1	27.3	58.3
K=2 FPS	77.2	61.6	48.1	36.8	46	27.1	58.4
K=10	76.7	61.7	48.4	37.3	45.6	27.3	58.4
K=15	77.2	62.7	49.5	38	45.4	26.8	58.5
K=20	76.2	62	49.2	38	45.1	26.9	58.8
K=25	76.2	61	47.6	36.5	45.4	27	57.7
K=30	**77.1**	**63.3**	**50.3**	**39**	**46.5**	**27.2**	**59.1**
K=35	76	60.6	46.8	35.5	43	26.8	57.4
K=40	76.8	61.6	48.3	37.3	42.6	27.2	57.8
K=45	75.9	60.8	46.8	35.9	43.2	27.3	57.5
K=50	75.7	61.1	48.3	37.5	44.6	27.2	58.3

Feature Extractors Compared

Table 4.7: Comparison of Visual Extractors using GRU Decoder ($K = 30$ Uniformly Sampled Frames)



VISUAL EXTRACTOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDER	METEOR	ROUGE_L	# PARAMETERS	GFLOPS
RESNET 18	76.6	63.1	52.4	42.6	81.5	33.1	68.1	22.74 M	54.48 G
RESNET 50	80.1	68.5	59.2	49.6	88.6	33.6	70.3	36.64 M	122.88 G
RESNET 101	79.1	67.2	58.1	49.6	96.6	34.6	70.7	55.54 M	234.18 G
RESNET 152	81.2	69.4	60.6	51.7	99.9	35.2	70.7	71.23 M	345.8 G
MOBILENET V2	80.5	69.4	60.3	51.5	90.2	34.6	71.3	14.54 M	9.18 G
MOBILENET V3 SMALL	76.3	62.8	53.1	43.2	77.7	31.9	69.4	13.54 M	1.98 G
MOBILENET V3 LARGE	77.7	64.8	55.1	46.3	85.7	33.2	69.6	16.54 M	6.78 G
SHUFFLENET V2×0.5	69.5	53.6	42.7	32.9	53.3	28.0	62.3	12.44 M	1.38 G
SHUFFLENET V2×1.0	70.2	53.8	42.9	33.5	56.6	28.0	62.7	13.34 M	4.38 G
SHUFFLENET V2×1.5	76.5	63.2	52.9	42.9	78.0	32.3	66.6	14.54 M	9.18 G
SHUFFLENET V2×2.0	78.2	67.2	58.8	50.7	80.8	34.6	68.8	18.44 M	17.58 G

Feature Extractors Compared

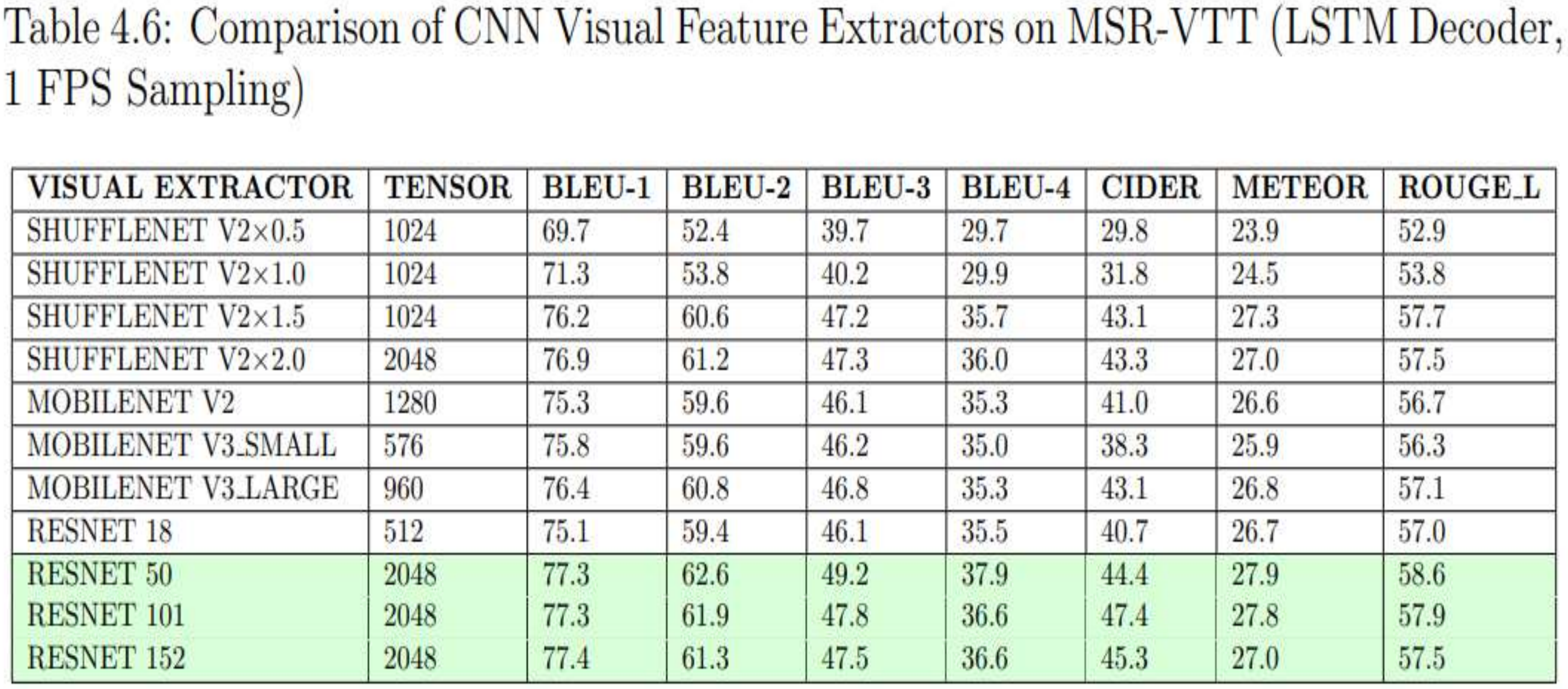
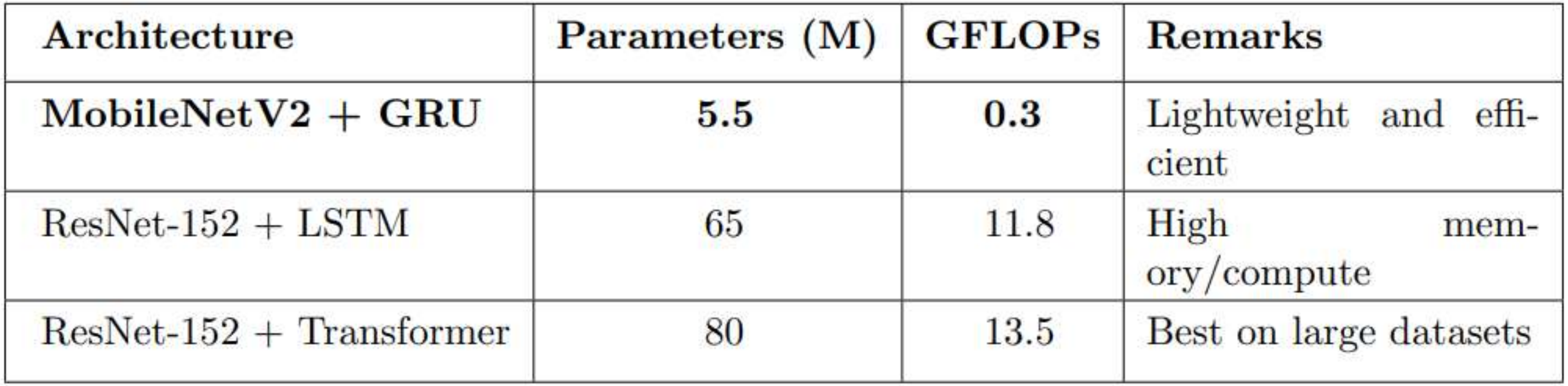


Table 4.6: Comparison of CNN Visual Feature Extractors on MSR-VTT (LSTM Decoder, 1 FPS Sampling)

VISUAL EXTRACTOR	TENSOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDER	METEOR	ROUGE_L
SHUFFLENET V2×0.5	1024	69.7	52.4	39.7	29.7	29.8	23.9	52.9
SHUFFLENET V2×1.0	1024	71.3	53.8	40.2	29.9	31.8	24.5	53.8
SHUFFLENET V2×1.5	1024	76.2	60.6	47.2	35.7	43.1	27.3	57.7
SHUFFLENET V2×2.0	2048	76.9	61.2	47.3	36.0	43.3	27.0	57.5
MOBILENET V2	1280	75.3	59.6	46.1	35.3	41.0	26.6	56.7
MOBILENET V3_SMALL	576	75.8	59.6	46.2	35.0	38.3	25.9	56.3
MOBILENET V3_LARGE	960	76.4	60.8	46.8	35.3	43.1	26.8	57.1
RESNET 18	512	75.1	59.4	46.1	35.5	40.7	26.7	57.0
RESNET 50	2048	77.3	62.6	49.2	37.9	44.4	27.9	58.6
RESNET 101	2048	77.3	61.9	47.8	36.6	47.4	27.8	57.9
RESNET 152	2048	77.4	61.3	47.5	36.6	45.3	27.0	57.5

Comparison of Model Architectures Based on Parameters and GFLOPs



Architecture	Parameters (M)	GFLOPs	Remarks
MobileNetV2 + GRU	**5.5**	**0.3**	Lightweight and efficient
ResNet-152 + LSTM	65	11.8	High memory/compute
ResNet-152 + Transformer	80	13.5	Best on large datasets

Results

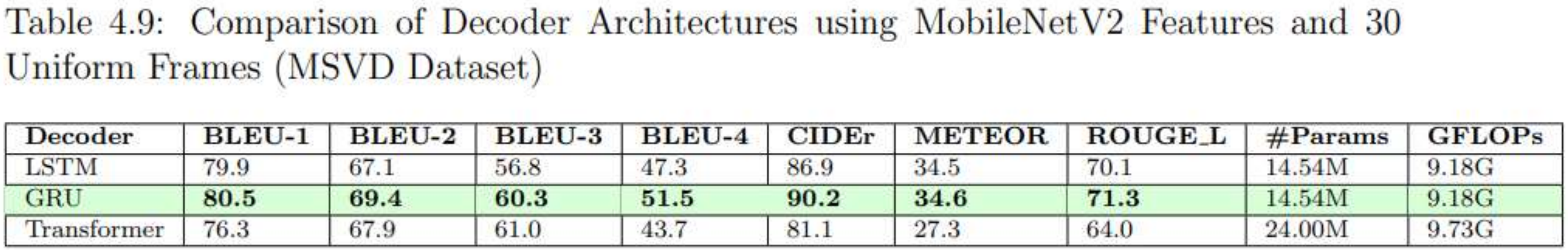


Table 4.9: Comparison of Decoder Architectures using MobileNetV2 Features and 30 Uniform Frames (MSVD Dataset)

Decoder	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**CIDEr**	**METEOR**	**ROUGE_L**	**#Params**	**GFLOPs**
LSTM	79.9	67.1	56.8	47.3	86.9	34.5	70.1	14.54M	9.18G
GRU	**80.5**	**69.4**	**60.3**	**51.5**	**90.2**	**34.6**	**71.3**	14.54M	9.18G
Transformer	76.3	67.9	61.0	43.7	81.1	27.3	64.0	24.00M	9.73G

- Lightweight models = efficient + accurate
- Suitable for real-time systems
- Good trade-off achieved
- Deployed on low compute devices

Future Works

- Learning-Based Frame Selection
- Multimodal Fusion
- Multilingual & Domain-Specific Expansion

Thank You