UCF
CENTER FOR RESEARCH IN COMPUTER VISION

ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

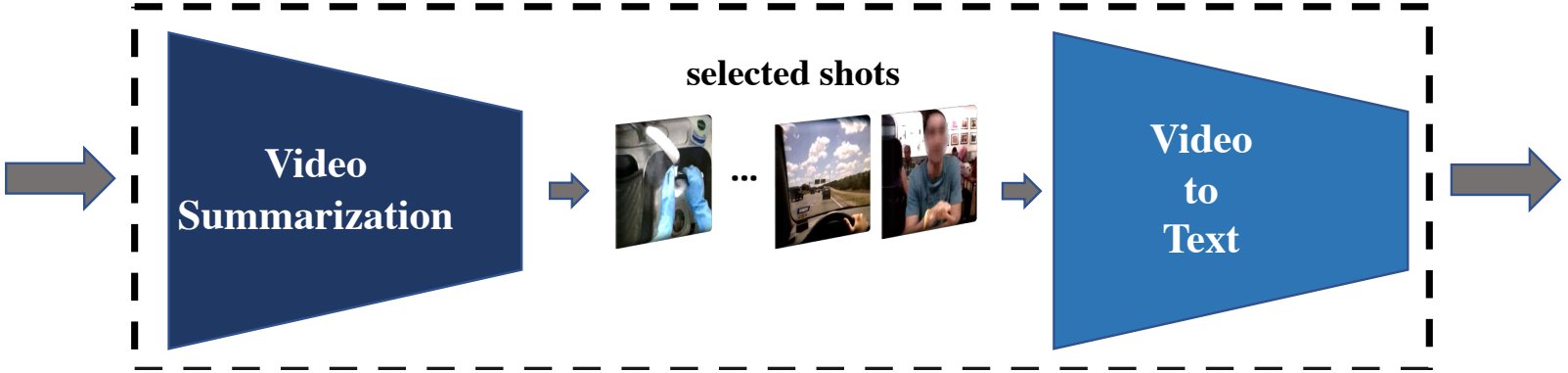
Text Synopsis Generation for Egocentric Videos

Aidean Sharghi, Niels da Vitoria Lobo, Mubarak Shah
University of Central Florida
Center for Research in Computer Vision

Video Summarization

- Objective: enable efficient browsing in large video datasets
- Conventional video summarization:
  - Select important segments from a given video
- Users still need to watch the summary videos
  - Browsing large datasets remains a challenge
- How to make summaries easier to browse?
- **Generate text summaries:**
  - Easy to read a short report
  - Search in text summary using natural language

Naïve Approach



My friend and I sat at the table and ate a meal together.
I drove in the car.

.
.
.

I washed the dishes.

Issues

- Imperfect summarization
  - Bad or repetitive selection
- Amplified through the text generation
  - Generate correct descriptions for egocentric videos is challenging

Alternative Approach

Video to Text

I looked at my friend.
I looked at my food.
I drove in the car.
.
.
.
I watched TV.

Text Summarization

My friend and I sat at the table and ate a meal together.
I drove in the car.

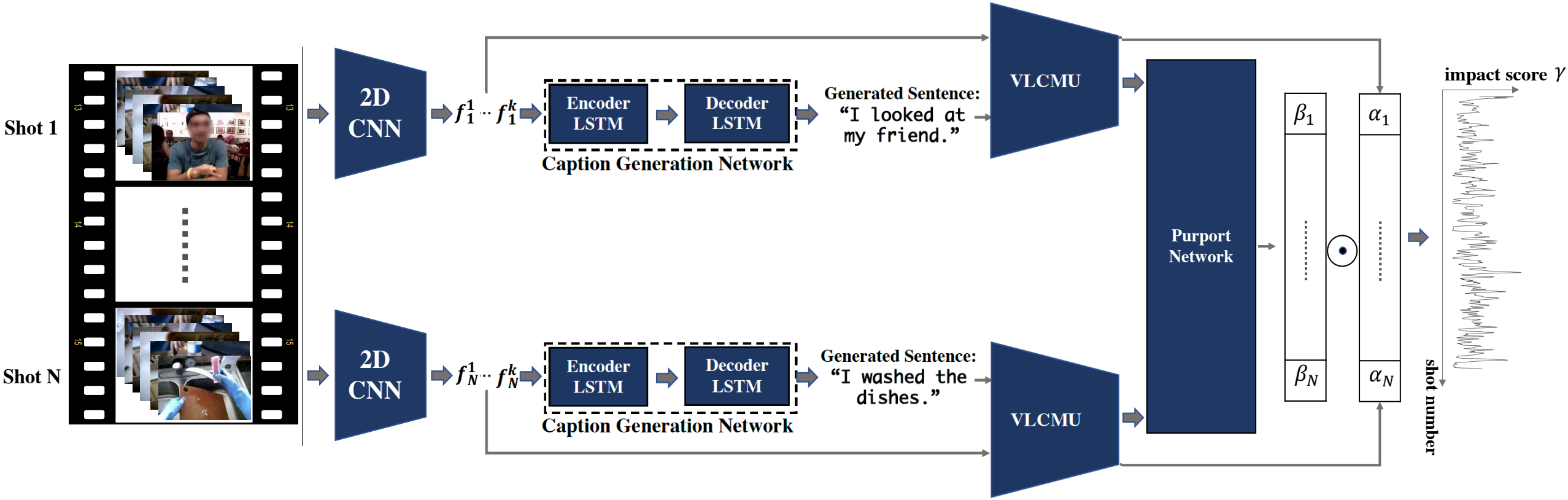
.
.
.

I washed the dishes.

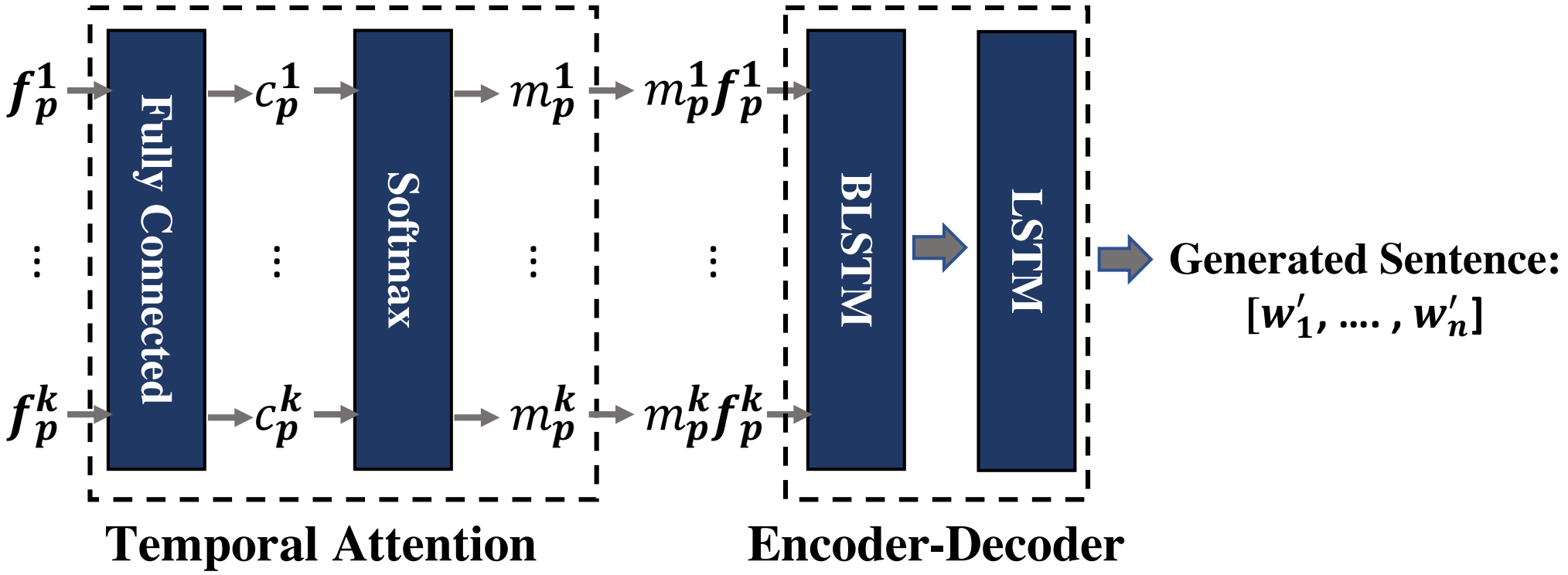
Advantages

- Generate as much text knowledge as possible
- Identify correct descriptions
- Select most representative ones to generate text summary

Framework

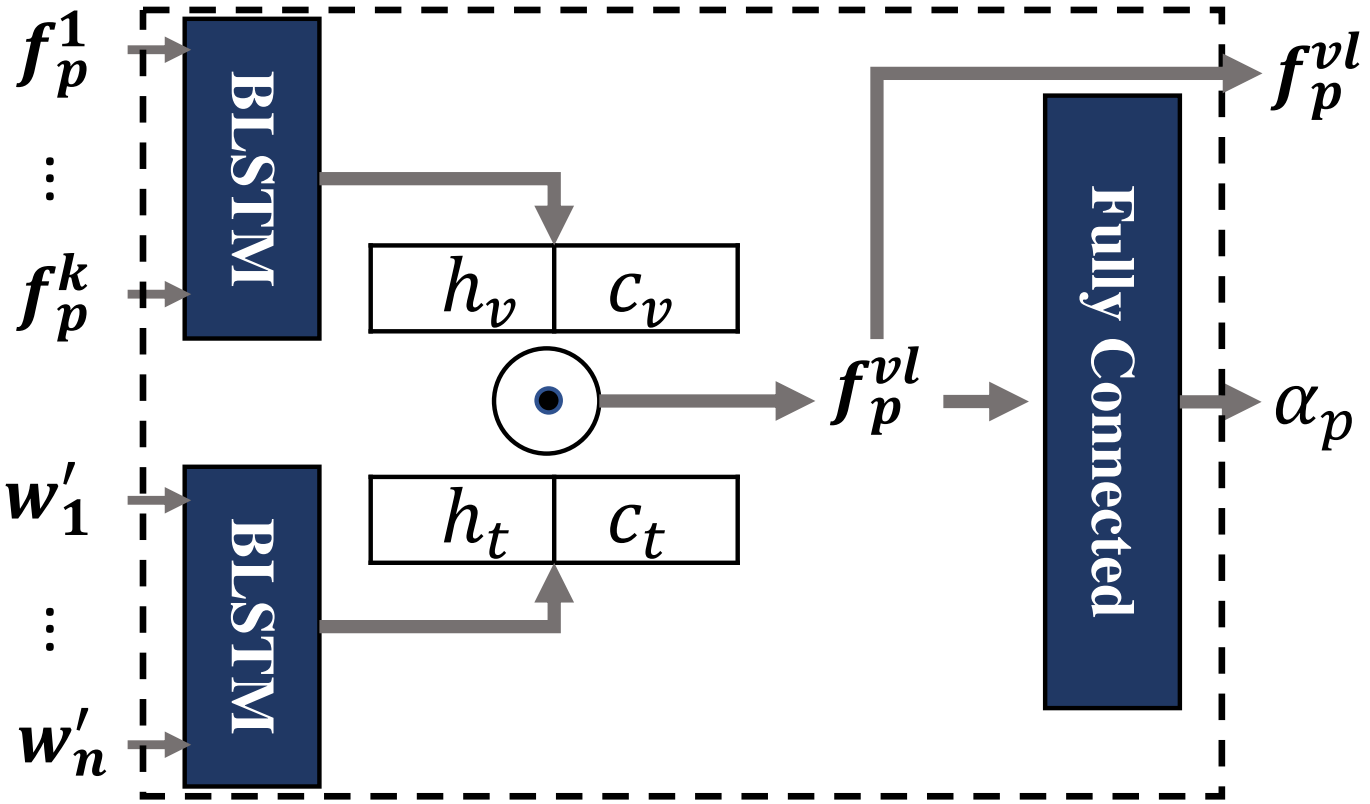

Shot 1
Shot N
2D CNN
2D CNN
$f_1^1 \cdots f_1^k$
$f_N^1 \cdots f_N^k$
Encoder LSTM
Decoder LSTM
Caption Generation Network
Encoder LSTM
Decoder LSTM
Caption Generation Network
Generated Sentence: "I looked at my friend."
Generated Sentence: "I washed the dishes."
VLCMU
VLCMU
Purport Network
β_1
β_N
α_1
α_N
impact score γ
shot number

Caption Generation Network



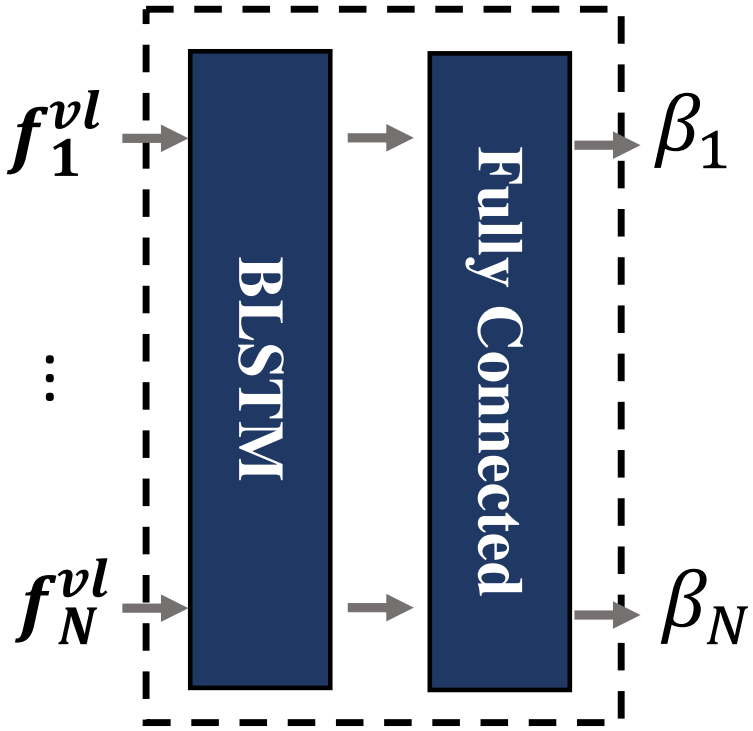
$$\mathcal{L}^c = -\sum \log pr(w_i)$$

Visual-Language Content Matching Network

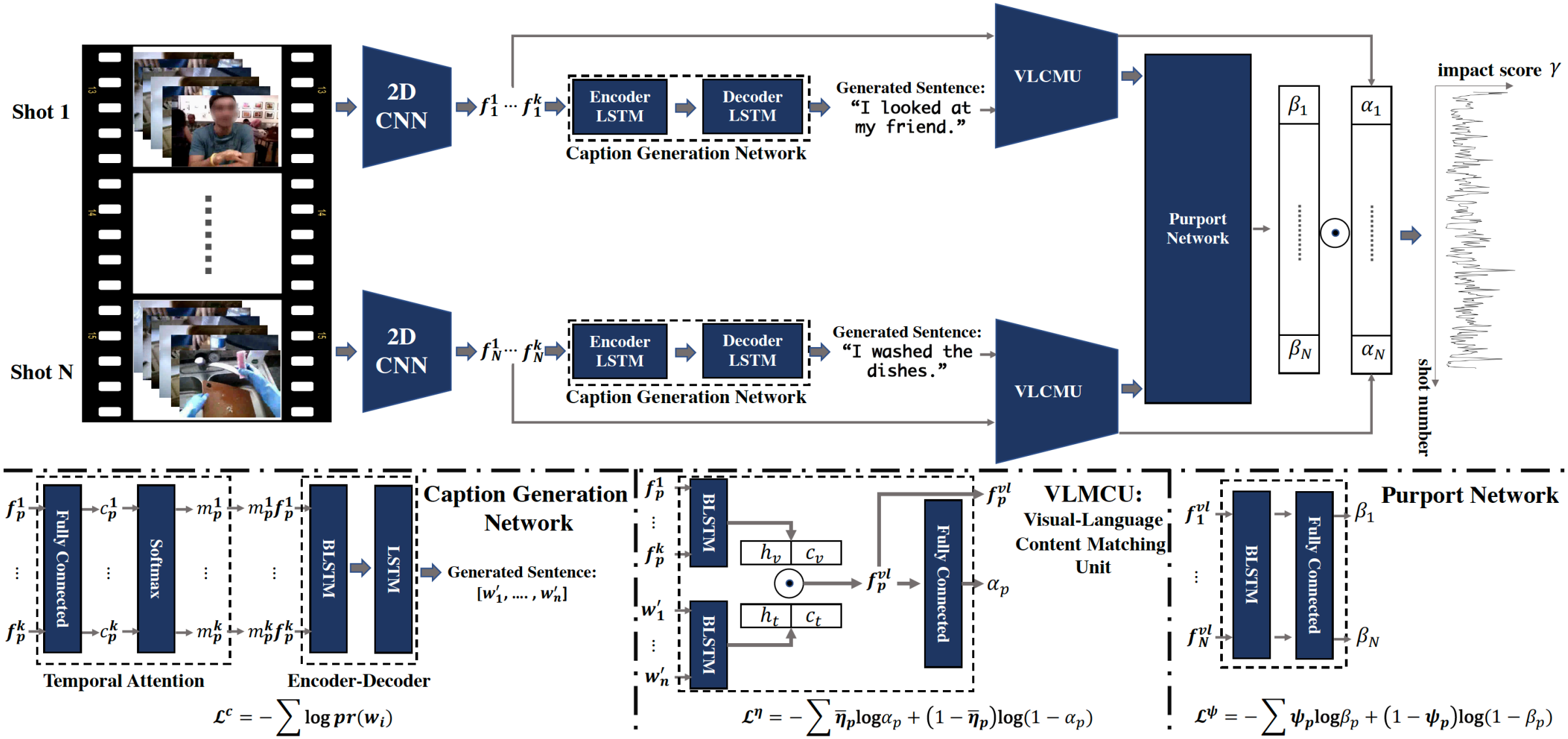


$$\mathcal{L}^{\eta} = -\sum \bar{\eta}_p \log \alpha_p + (1 - \bar{\eta}_p) \log(1 - \alpha_p)$$

Purport Network



$$\mathcal{L}^{\psi} = -\sum \boldsymbol{\psi_p} \mathbf{log} \beta_p + \left(1 - \boldsymbol{\psi_p}\right) \mathbf{log}(1 - \beta_p)$$



$$\mathcal{L}^c = -\sum \log pr(w_i)$$

$$\mathcal{L}^\eta = -\sum \bar{\eta}_p \log \alpha_p + (1-\bar{\eta}_p)\log(1-\alpha_p)$$

$$\mathcal{L}^\psi = -\sum \psi_p \log \beta_p + (1-\psi_p)\log(1-\beta_p)$$

$$\mathcal{L} = \sum_{v=1}^{V} \mathcal{L}_v^c + \lambda_1 \mathcal{L}_v^\eta + \lambda_2 \mathcal{L}_v^\varphi$$

Experiments

- Dataset: UT-Egocentric[1]
- Evaluation:
  - Visual Domain
    - AUC of IOU[2]
  - Text Domain
    - ROUGE
    - BLEU
    - METEOR

1. Yeung, Serena, Alireza Fathi, and Li Fei-Fei. "Videoset: Video summary evaluation through text." *arXiv preprint arXiv:1406.5824* (2014).

2. Aidean Sharghi, Ali Borji, Chengtao Li, Tianbao Yang, Boqing Gong. “Improving Sequential Determinantal Point Processes for Supervised Video Summarization” In European Conference on Computer Vision (ECCV 2018)

Quantitative Results – Text Domain

- Baselines:
  - State-of-the-art Methods
  - Video Caption Generation

	SeqDPP	Superframes	SubMod	$\text{SubMod}_{\text{vis+lang}}$	LM-SeqGDPP	LSTMDPP	**TSG**
ROUGESU4	11.95	15.59	16.19	14.10	16.24	13.98	**17.33**
ROUGE-L	11.41	13.88	16.24	13.95	16.99	12.84	**24.98**
METEOR	19.02	18.36	18.90	18.02	**19.90**	17.32	19.27
BLEU2	16.69	34.68	35.28	35.38	35.11	31.72	**46.90**

Quantitative Results – Visual Domain

	SeqDPP	Superframes	SubMod	$\text{SubMod}_{\text{vis+lang}}$	LM-SeqGDPP	LSTMDPP	**TSG**
AUC	12.91	9.85	11.53	12.26	13.14	7.57	**13.19**

Thank you!