Leveraging Local Temporal Information for Multimodal Scene Classification

Saurabh Sahu[1], Palash Goyal[1]

[1]Samsung Research America

April 18, 2022

Outline

Motivation

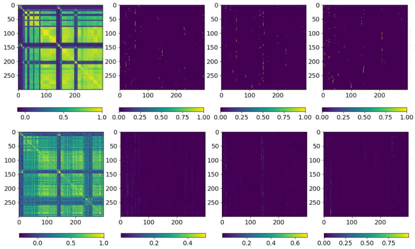

Figure: Inter-frame similarity for visual (leftmost top) and audio features (leftmost bottom). Attention maps from three heads of the baseline self-attention models with visual (top right) and audio (bottom right) features as input.

- Similarity matrices capture high inter-frame correlation present in the video
- Attention maps are quite sparse with vertical lines indicating that only a few selective frames are being attended to get the output
- This can lead to erroneous predictions

GMSA - Formulation

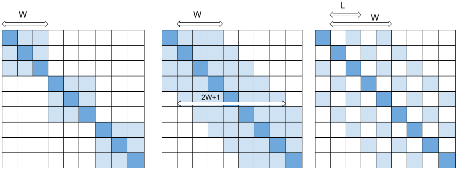

Figure: Block diagonal (left), Toeplitz (center) and Toeplitz-Dilated (right) masks.

- **Leveraging local attention maps**

$$A_m = \text{softmax}\left(\frac{Q_m K_m^T}{\sqrt{d_m}} \odot mask_m\right) \tag{1}$$

$$mask_m[i,j] = \begin{cases} 1, & \text{if } j \in N_i \\ -\infty, & \text{otherwise} \end{cases} \tag{2}$$

$mask_m$ enforces a frame i to attend to only the frames j in its neighborhood N_i by masking out other time-steps

GMSA - Formulation

- **Sharing attention maps**
  Half of the attentions heads compute global attention maps (as defined in original Transformer architecture) while the other half compute local attention maps as defined in equations 1 and 2.
- **Gating attention maps**

$$R^g, R^l = softmax([A_m^g W_g^g, A_m^l W_g^l]) \tag{3}$$
$$A_m = R^g \odot A_m^g + R^l \odot A_m^l \tag{4}$$

  where $W_g^g, W_g^l \in \mathbb{R}^{T \times T}$ are learnable layers shared across the heads
- **Gating contextual representations**

$$O^g = \text{concat}(A_1^g V_1, \ldots, A_M^g V_M) \tag{5}$$
$$O^l = \text{concat}(A_1^l V_1, \ldots, A_M^l V_M) \tag{6}$$
$$Y^g = O^g W^{og}, \quad Y^l = O^l W^{ol} \tag{7}$$
$$R^g, R^l = softmax([O^g W_g^g, O^l W_g^l]) \tag{8}$$
$$Y = R^g \odot Y^g + R^l \odot Y^l \tag{9}$$

  where $W^{og}, W^{ol}, W_g^g, W_g^l \in \mathbb{R}^{D \times D}$ are learnable layers

Results

	Baseline	ShareAtt			GateAtt			GateOp		
		BD	*TP*	*TD*	BD_{10}	TP_{10}	TD^{5}_{80}	BD_{20}	TP_{30}	TD^{4}_{60}
GAP	85.07	85.18	85.25	85.16	85.21	85.30	85.30	**86.03**	**86.13**	**85.99**
MAP	44.61	44.81	45.04	44.69	44.89	45.28	45.31	**47.03**	**47.49**	**46.98**
PERR	78.97	79.05	79.22	79.03	79.13	79.27	79.21	**79.96**	**80.10**	**79.88**
Hit@1	87.75	87.79	87.92	87.78	87.85	87.96	87.91	**88.40**	**88.49**	**88.33**
Train Time/epoch	43 min	50 min			60 min			49 min		
Disk Size	48.9 MB	48.9 MB			50.2 MB			61 MB		

Qualitative analysis

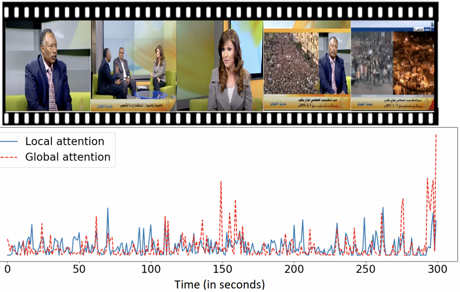

Figure: Local and global attention profiles obtained for a video in test set. Link: *youtube.com/watch?v=ygORXiV2Zpw*

- Global attention profile is more uneven and puts most attention towards the end of the video showing a crowd of people. Hence, the baseline model predicts the video incorrectly ('Association Football')
- In contrast, local attention profile is more uniform temporally. Using information from both local and global contexts, we get the correct prediction using GateOp model with BD_{20} mask ('News Program')

Importance of local information

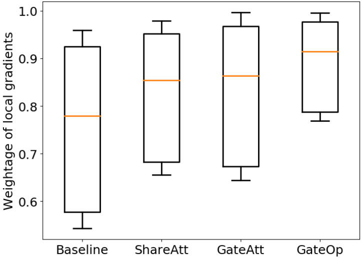

Figure: Ratio of local to non-local gradients for visual features

- We measure the sensitivity S_i of an output frame to the local input frames in its neighborhood N_i by computing $G[i,j]$ which denotes the norm of the gradient of output frame $Y[i]$ with respect to an input frame $X[j]$

$$S_i = \frac{avg_{j \in N_i} G[i,j]}{avg_{j \in N_i} G[i,j] + avg_{j \notin N_i} G[i,j]} \tag{10}$$

- We observe a positive correlation between model performance and sensitivity to local context.

Summary and Future Work

- **Summary**
  - Enforcing Transformer to learn local and global level attention separately improves model's generalizability
  - Outputs of the better performing GMSA models were found to be more sensitive to local input frames than the baseline mode thereby.
  - Adding local context could mitigate the effects of incorrect global attention maps
- **Future Work**
  - Using self-supervision to reduce dependency on labeled data

Summary and Future Work

- **Summary**
  - Enforcing Transformer to learn local and global level attention separately improves model's generalizability
  - Outputs of the better performing GMSA models were found to be more sensitive to local input frames than the baseline mode thereby.
  - Adding local context could mitigate the effects of incorrect global attention maps
- **Future Work**
  - Using self-supervision to reduce dependency on labeled data
  - Training models on raw videos with new base models such as Multiscale Vision Transformers

Summary and Future Work

- **Summary**
    - Enforcing Transformer to learn local and global level attention separately improves model's generalizability
    - Outputs of the better performing GMSA models were found to be more sensitive to local input frames than the baseline mode thereby.
    - Adding local context could mitigate the effects of incorrect global attention maps
- **Future Work**
    - Using self-supervision to reduce dependency on labeled data
    - Training models on raw videos with new base models such as Multiscale Vision Transformers
    - Adding local attention maps at various hierarchies.

Summary and Future Work

- **Summary**
  - Enforcing Transformer to learn local and global level attention separately improves model's generalizability
  - Outputs of the better performing GMSA models were found to be more sensitive to local input frames than the baseline mode thereby.
  - Adding local context could mitigate the effects of incorrect global attention maps
- **Future Work**
  - Using self-supervision to reduce dependency on labeled data
  - Training models on raw videos with new base models such as Multiscale Vision Transformers
  - Adding local attention maps at various hierarchies.
  - Explore video segmentation techniques to define better neighborhoods for computing local attention maps rather than a fixed-length mask based approach.

Questions?

THANK YOU