Diversifying Joint Vision-Language Tokenization Learning

Vardaan Pahuja[1*], AJ Piergiovanni[2], Anelia Angelova[2]
[1]The Ohio State University [2]Google DeepMind
*Work done while at Google

Motivation

- **Advantages of disentangled representations**
  - ➔ Superior out-of-domain (OOD) generalization
  - ➔ Better interpretability
  - ➔ Better sample efficiency
  - ➔ Better transfer learning capabilities
- Use a diversity-enforcing loss to encourage disentangled representations.

Background

- **Tokenlearner**
  - Adaptively learn a fixed set of token representations across one or more modalities.
  - Select a series of informative combinations of spatial locations in the image conditioned on all modalities.
  - For the i^{th} token z_i, it learns a spatial attention map $\alpha_i(X)$ which is multiplied with the input X to generate a token output $A_i(X)$,

$$z_i = A_i(X) = \rho(X \odot \gamma(\alpha_i(X)))$$

- **Co-tokenization**
  - Cross-modality interaction during the visual feature extraction process by learning token representations, rather than rather considering them as an afterthought after feature extraction.
  - Multiple streams of video at different spatio-temporal scales for multimodal representation learning.

Overall VideoQA Results

Model	MSRVTT-QA	MSVD-QA	GFLOPs
Co-tokenization	33.7	32.5	67
Ours	33.1	30.1	41

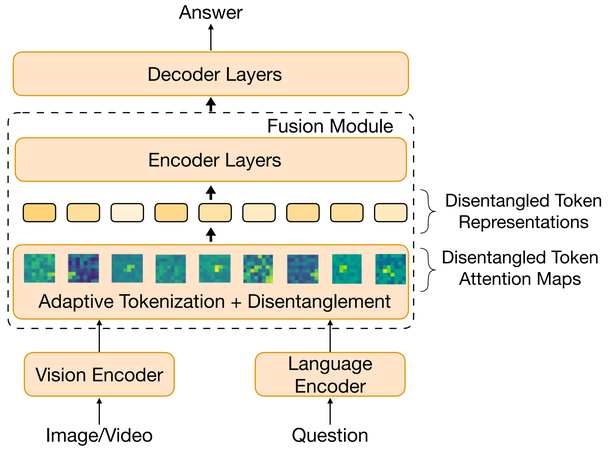
Table 1: Comparison to state-of-the-art approaches for VideoQA (open-vocabulary). We pretrain on 10% subset of the HowTo69MVQA dataset, whereas Co-tokenization pretrained on the full HowTo100M dataset. We demonstrate competitive performance despite having a smaller model capacity.

Proposed Approach

- Add ***diveristy loss*** as an auxiliary objective.

$$\mathcal{L}_{div} = \sum_{k=1}^{N} \sum_{j=1, j\neq i}^{M} < \alpha_i^k, \alpha_j^k >^2$$

α_i^k denotes the spatial attention weights for the i^{th} token in example k.



Overall VQA Results

Model	GQA	SNLI-VE
SimVLM (Huge)	–	**86.32**
UNITER	–	79.38
VinVL	65.05	–
LXMERT	60.0	–
Ours	**76.79**	80.15

Table 2: Comparison to state-of-the-art approaches (VQA)

VideoQA Results

Dataset	Pre-training	Model	Accuracy
MSRVTT-QA		Baseline	31.06
		Ours	**31.37**
	✓	Baseline	31.78
	✓	**Ours**	**33.05**
MSVD-QA		Baseline	27.98
		Ours	**28.22**
	✓	Baseline	28.08
	✓	**Ours**	**30.11**
IVQA		Baseline	9.48
		Ours	**9.96**
	✓	Baseline	8.86
	✓	**Ours**	**9.97**

Table 3: Video QA results in the open vocabulary setting (val. set). The baseline is a similar capacity Co-tokenization model.

VQA Results

		Val. set		Test set	
Dataset	**Model**	**E.M.**	**F1**	**E.M.**	**F1**
SNLI-VE	Baseline	76.70	76.70	76.59	76.59
	Ours	**78.06**	**78.06**	**77.36**	**77.37**
GQA	Baseline	73.48	73.56	**73.5**	**73.57**
	Ours	**75.02**	**75.11**	75.01	75.1

Table 4: VQA results in the pre-training setting.

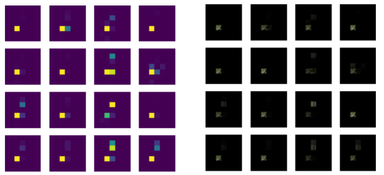
		Val. set		Test set	
Dataset	**Model**	**E.M.**	**F1**	**E.M.**	**F1**
SNLI-VE	Baseline	73.08	73.08	72.5	72.5
	Ours	**73.15**	**73.15**	**72.69**	**72.69**
GQA	Baseline	**68.08**	**68.13**	**68.14**	**68.2**
	Ours	67.98	68.02	67.98	68.02

Table 5: VQA results in the no pre-training setting.

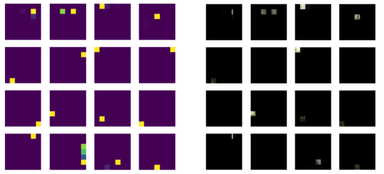
Visualizations

- Localize attention to salient areas of the image, that are vital for answering the question.

(a) Question Image



(b) Token visualization (Baseline)



(c) Token visualization (Ours)

Figure 1: (a) *Question*: what kind of climbing vine or plant is this? *Baseline*: tombppry, *Ours*: ivy, *Ground truth answers* = ['fern', 'grape', 'vine', 'ivy', 'unanswerable', 'creeping fig', 'unanswerable', 'unanswerable', 'ivy', 'green']; *Bottom left*: Weights assigned to each image patch for every token, lighter shades like yellow correspond to higher weights; *Bottom right*: Token attention masks grounded to the input image.