You Need Multiple Exiting: Dynamic Early Exiting for Accelerating Unified Vision Language Model

Shengkun Tang[1], Yaqing Wang[2], Zhenglun Kong[3],
Tianchi Zhang[4], Yao Li[5], Caiwen Ding[6], Yanzhi Wang[3], Yi Liang[2], Dongkuan Xu[1]

Paper Tag: WED-AM-243

Agenda

- ❑ Introduction
- ❑ Motivation
- ❑ Methods
- ❑ Results
- ❑ Takeaways

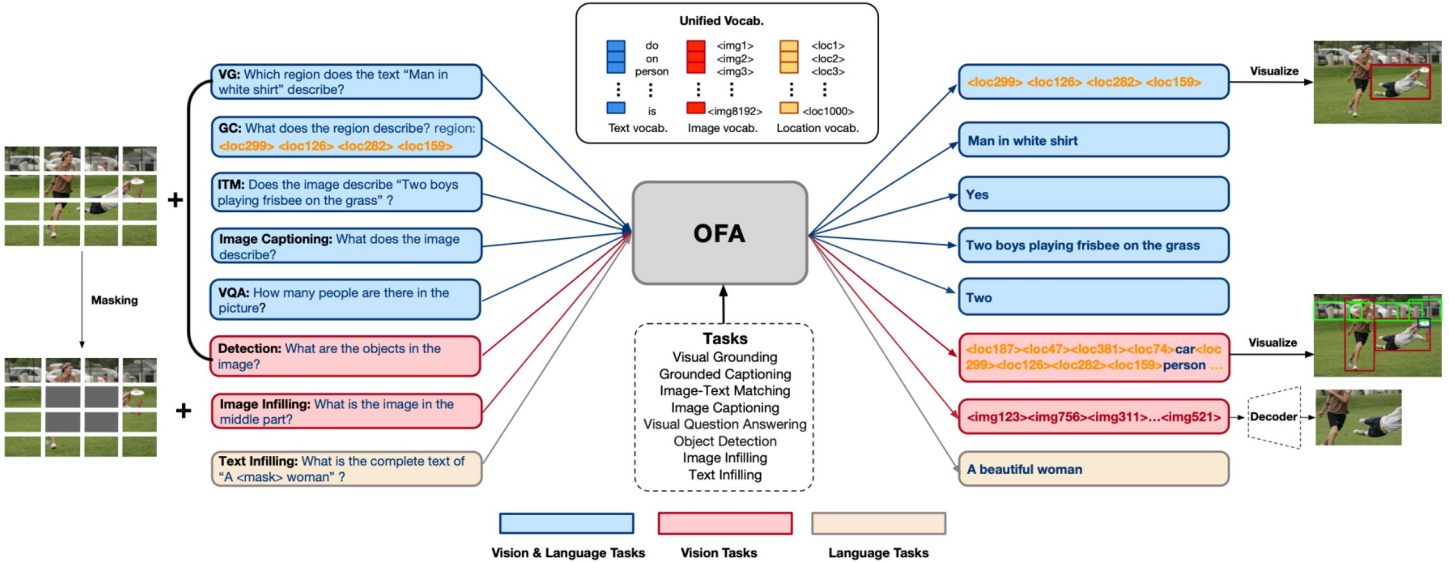
Agenda

- ❑ Introduction
- ❑ Motivation
- ❑ Methods
- ❑ Results
- ❑ Takeaways

Multi-Modal World

Modality: Text, Vision, Audio…

Unified Vision-Language Learning



Transformer-based Sequence-to-Sequence Framework

Wang P, Yang A, Men R, et al. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework[C]//International Conference on Machine Learning. PMLR, 2022: 23318-23340.

Drawback

Large computation resource requirement during inference

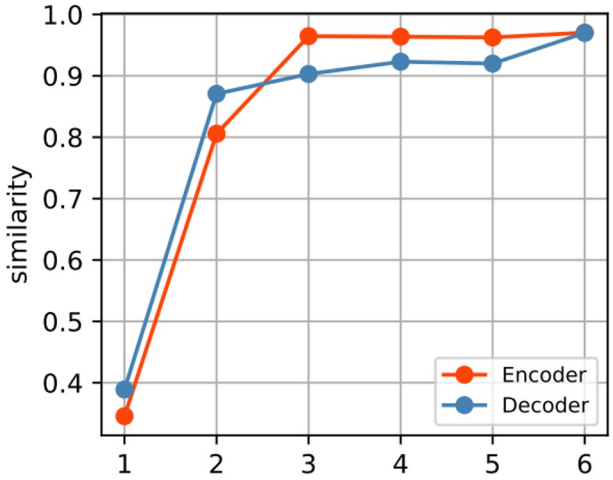
Agenda

- ❑ Introduction
- ❑ Motivation
- ❑ Methods
- ❑ Results
- ❑ Takeaways

Motivation

Saturation status [1]

Redundant layer removal

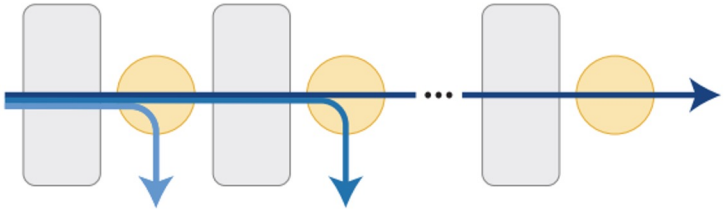


[1] Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. arXiv preprint arXiv:2203.14680, 2022.

Agenda

- ❑ Introduction
- ❑ Motivation
- ❑ Methods
- ❑ Results
- ❑ Takeaways

Early Exiting



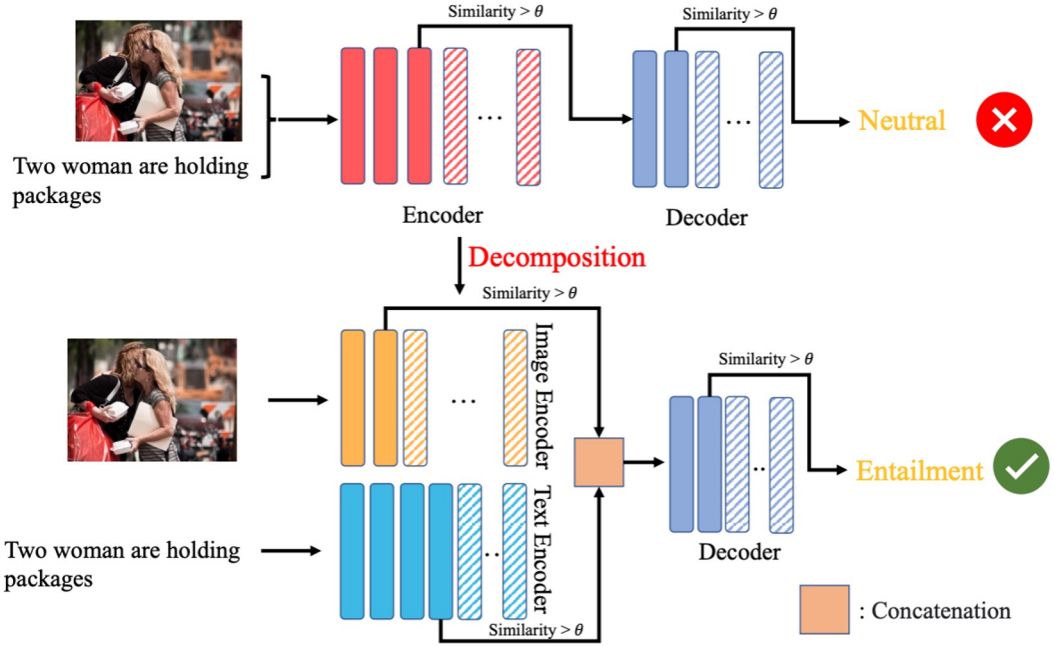
A classifier to make exiting decision based on confidence or entropy

Xin J, Tang R, Lee J, et al. DeeBERT: Dynamic early exiting for accelerating BERT inference[J]. arXiv preprint arXiv:2004.12993, 2020.

Challenges

1. Dependencies in making decisions for exiting decisions in the encoder and decoder
2. Difficulty to apply confidence classifiers to skip the encoder layers

Methods



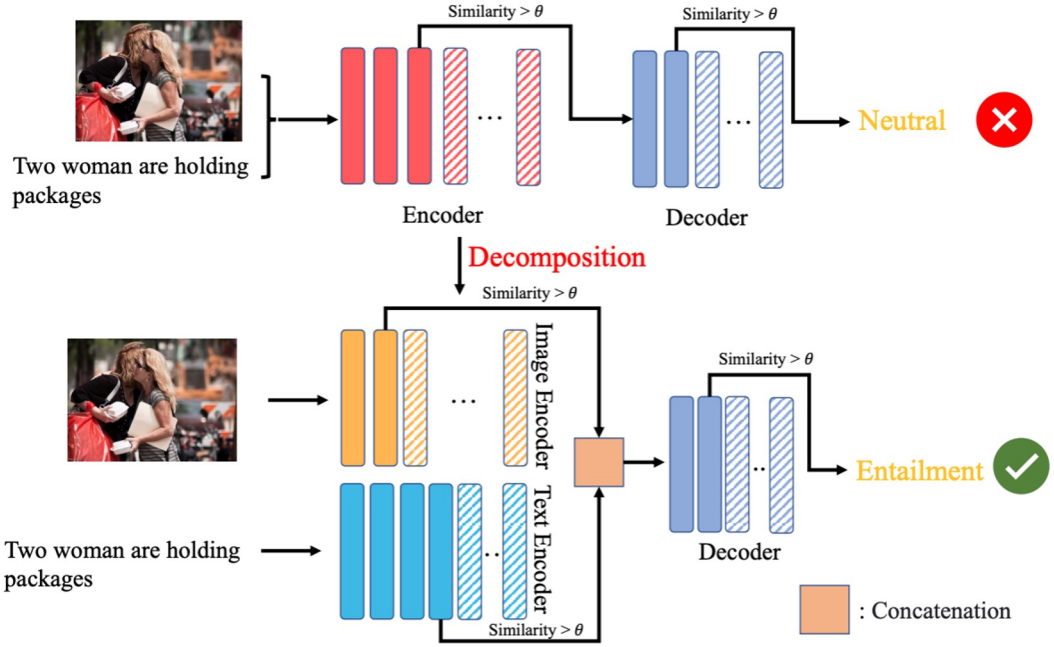
Modality Decomposition

$$[\mathbf{I}_n; \mathbf{T}_n] = E_{1:n}([\mathbf{I}_0; \mathbf{T}_0]).$$

$$[\mathbf{I}_p; \mathbf{T}_q] = [E_{1:p}(\mathbf{I}_0); E_{1:q}(\mathbf{T}_0)].$$

T_0 : Text input
I_0 : Image input
$E_{1:p}$: Encoder

Methods



Similarities for exiting decision

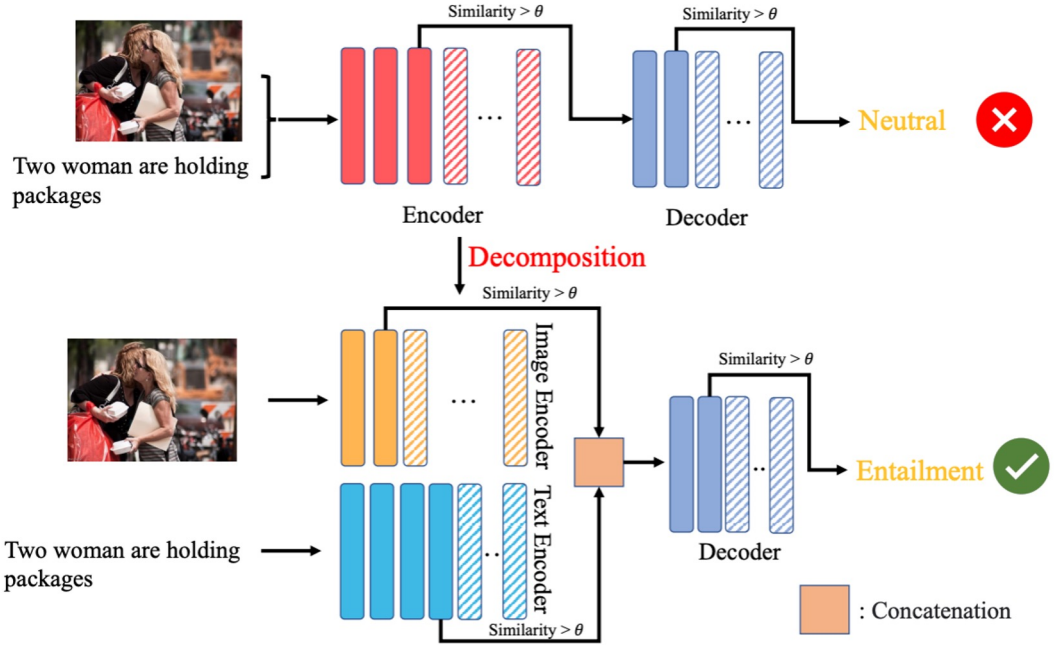
$$\text{ImgSim}_i = \text{CosSim}(E_i(\mathbf{I}_i), E_{i-1}(\mathbf{I}_{i-1})),$$
$$\text{TxtSim}_i = \text{CosSim}(E_i(\mathbf{T}_i), E_{i-1}(\mathbf{T}_{i-1})),$$
$$\text{DecSim}_{i,s} = \text{CosSim}(D_i(Td_{i,s}), D_{i-1}(Td_{i-1,s})).$$

Decay threshold for Decoder

$$\Theta(t) = \beta\theta + (1-\beta)e^{-\tau t/N}$$

t : Timestep
N : All steps
D_i : Decoder

Methods



Task layer-wise loss

$$\mathcal{L} = \frac{1}{N}\sum_{i}^{N}\mathcal{L}_{CE},$$

$$\mathcal{L}_{CE} = -\sum_{i=1}^{|y|}\log P_\theta(y_i|y_{<i}, \mathbf{I}, \mathbf{T}),$$

Agenda

- ❑ Introduction
- ❑ Motivation
- ❑ Methods
- ❑ Results
- ❑ Takeaways

Results

Models	SNLI-VE			Image Captioning				
	Dev	Test	Time	BLEU-4	METEOR	CIDEr	SPICE	Time
OFA_{Base}	89.3	89.2	1	42.8	31.7	146.7	25.8	1
OFA_{Tiny}	85.3	85.2	-33%	38.1	29.2	128.7	23.1	-33%
DeeBERT	78.9	78.8	-15%	30.1	26.3	102.1	20.5	-15.5%
PABEE	85.3	85.2	-15.3%	31.4	26.8	105.8	21	-16.3%
DeeCap	-	-	-	38.7	29.1	129	22.5	-38%
Ours	**88.7**	**88.5**	**-50%**	**41.6**	**30.6**	**137**	**24.4**	**-40.2%**

Image Captioning on MS-COCO [1]

Visual Entailment on SNLI-VE [2]

[1] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014: 740-755.
[2] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. arXiv:1901.06706, 2019.

Findings

Task	Image Layer	Text Layer	BLEU-4	METEOR	CIDERr	SPICE
Image Captioning	6.0	6.0	42.4	31.2	143.9	25.1
	3.1	2.0	32.8	27.4	112.1	20.8
	3.1	6.0	33.1	27.4	112.2	20.7
	6.0	2.0	**42.0**	**31.2**	**143.6**	**25.1**

Task	Image Layer	Text Layer	Dev	Test
Visual Entailment	6	6	88.6	88.7
	2.03	2.9	76.1	75.6
	6	2.9	79.1	79.5
	2.03	6	**88.4**	**88.6**

Computation requirement varies for different modalities on different tasks

Agenda

- ❑ Introduction
- ❑ Motivation
- ❑ Methods
- ❑ Results
- ❑ Takeaways

Takeways

Unified vision language models require more efficiency

Early Exiting accelerates vision language models inference by:

1. Modality Decomposition
2. Similarity for exiting decision
3. Task layer-wide loss

Open issues:

1. Performance drops when efficiency > 60%
2. More modality

Thank you!

You Need Multiple Exiting: Dynamic Early Exiting for Accelerating Unified Vision Language Model

Shengkun Tang[1] (@Shengkun Tang), Yaqing Wang[2], Zhenglun Kong[3], Tianchi Zhang[4], Yao Li[5], Caiwen Ding[6], Yanzhi Wang[3], Yi Liang[2], Dongkuan Xu[1]

North Carolina State University

Web: https://tangshengku.github.io/

05-31-2023