Uni-Perceiver v2: A Generalist Model for Large-Scale Vision and Vision-Language Tasks

Hao Li*, Jinguo Zhu*, Xiaohu Jiang*, Xizhou Zhu✉, Hongsheng Li, Chun Yuan, Xiaohua Wang, Yu Qiao, Xiaogang Wang, Wenhai Wang, Jifeng Dai
* Equal Contribution ✉ Corresponding Author

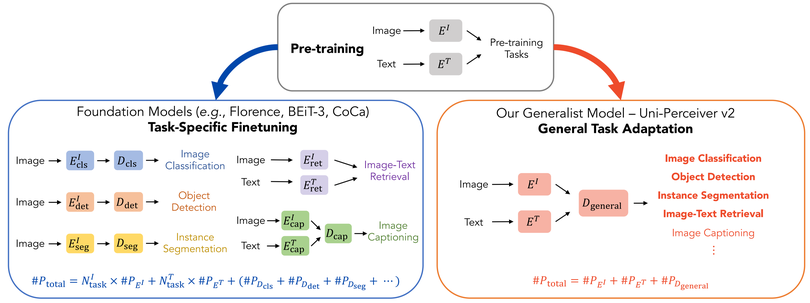
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Highlights

- Handling a broad range of vision / vision-language tasks without finetuning
- First generalist model that supports both localization and non-localization tasks with competitive performance
- Outperforming all existing generalist models in both versatility and performance
- Achieving competitive performance compared with commonly-recognized task-specific strong baselines
- Code and model will be released at `https://github.com/fundamentalvision/Uni-Perceiver`

arXiv GitHub Email: haoli@link.cuhk.edu.hk

Comparison of Foundation Models and Uni-Perceiver v2



Foundation models are costly to adapt to downstream tasks

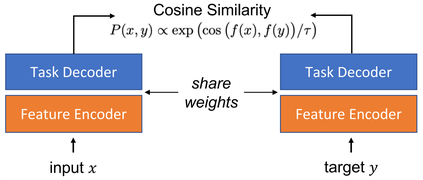
- Enough data needs to be collected and labeled for training on each downstream task
- Task modules (*e.g.*, detection heads) need to be designed and trained
- Thousands of models for thousands of tasks / real-world scenarios

How to build a generalist model capable of handling different tasks without finetuning?

- Requiring flexible image representation, unified task formulation, and effective multi-task training

Unified Task Formulation

Different tasks are identifies as different input set $\mathcal{X}$ and candidate output set $\mathcal{Y}$.
Given $x \in \mathcal{X}$, the task is formulated as finding $y \in \mathcal{Y}$ with the maximum likelihood for x.



Unified task formulation from Uni-Perceiver

- The likelihood between input x and target y is

$$P(x,y) \propto \exp\left(\cos\left(f(x), f(y)\right)/\tau\right)$$

- Given x, the target $\hat{y}$ with the maximum likelihood is

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} P(x,y)$$

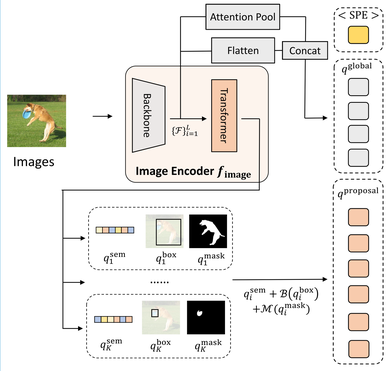
- Loss function for multi-task joint training is

$$L = \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\{x,y\} \in \{\mathcal{X}_i, \mathcal{Y}_i\}} \left[-\log \frac{P(x,y)}{\sum_{z \in \mathcal{Y}_i} P(x,z)} \right]$$

Encoding Images as General Region Proposals

What image representation is suitable for both localization and non-localization tasks?

- Non-overlapping patches with fixed sizes is not flexible enough for modeling objects of varying sizes and shapes
- We propose to encode images as general region proposals with a transformer-based region proposal network
- A sequence of regional proposals is extracted from the image, and complemented with global representations



Transformer-based region proposal network

- The image is encoded as the concatenation of global and regional representations

$$f_{\text{image}}(x) = \text{Concat}\left(\{q_i^{\text{global}}\}_{i=1}^{M}, \{q_j^{\text{proposal}}\}_{j=1}^{N}\right)$$

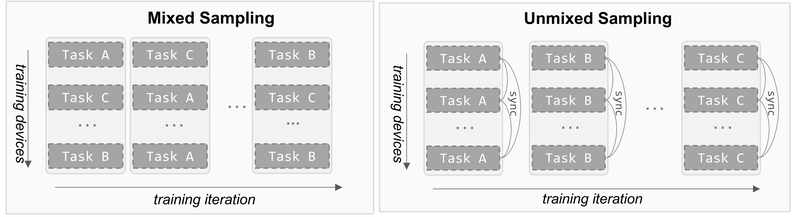
- The global representation is encoded as

$$q^{\text{global}} = \text{Concat}\left(\{\text{AttnPool}_i(\mathcal{F}_L)\}_{i=1}^{M'}, \text{Flatten}(\mathcal{F}_L)\right)$$

- The regional representations are encoded from the semantic, bounding box and segmentation mask of each general region proposal

$$q_j^{\text{proposal}} = q_j^{\text{sem}} + \mathcal{B}(q_j^{\text{box}}) + \mathcal{M}(q_j^{\text{mask}})$$

Task-Balanced Gradient Normalization for Multi-task Learning



Unmixed sampling strategy: sampling only one tasks across all GPUs in one iteration

- Increasing effective batch-size for each task, which improves efficiency and performance
- Reducing the synchronization cost caused by difference task time on different GPUs
- **Problem:** the gradients differ significantly between iterations, causing training instability

Task-Balanced Gradient Normalization

$$\begin{cases} \mathbf{g}_t \leftarrow \nabla L_{t,k}(\theta_{t-1}) \\ \mathbf{m}_t = (1-\beta_1)\mathbf{m}_{t-1} + \beta_1 \mathbf{g}_t \\ \mathbf{n}_t = (1-\beta_2)\mathbf{n}_{t-1} + \beta_2 \mathbf{g}_t^2 \\ \theta_t = \theta_{t-1} - \alpha \frac{\mathbf{m}_t}{\sqrt{\mathbf{n}_t + \varepsilon}} \end{cases} \Rightarrow \begin{cases} \mathbf{g}_t \leftarrow \omega_k \frac{\nabla L_{t,k}(\theta_{t-1})}{\|\nabla L_{t,k}(\theta_{t-1})\|} \\ \mathbf{m}_t = (1-\beta_1)\mathbf{m}_{t-1} + \frac{\beta_1}{s_k} \mathbf{g}_t \\ \mathbf{n}_t = (1-\beta_2)\mathbf{n}_{t-1} + \frac{\beta_2}{s_k} \mathbf{g}_t^2 \\ \theta_t = \theta_{t-1} - \alpha \frac{\mathbf{m}_t}{\sqrt{\mathbf{n}_t + \varepsilon}} \end{cases}$$

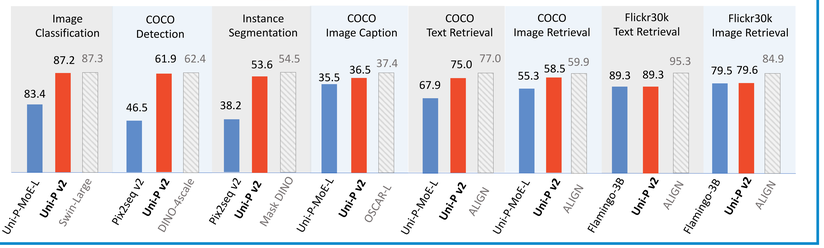
- Adaptively normalizing the gradients of each task
- Stabilizing the multi-task joint training with unmixed sampling strategy

Experiments

Comparison of Uni-Perceiver v2 with existing generalist models

Methods	#params	Image Classification	Object Detection	Instance Segmentation	Image Captioning		Text Retrieval		Image Retrieval	
		ImageNet-1k	COCO	COCO	COCO		COCO	Flickr30k	COCO	Flickr30k
		Acc	mAP	mAP	B@4	CIDEr	R@1	R@1	R@1	R@1
Pix2Seq v2	132M	-	46.5	38.2	34.9	-	-	-	-	-
UniTab	185M	-	-	-	-	115.8	-	-	-	-
Unified-IO $_{\text{LARGE}}$	776M	71.8	-	-	-	-	-	-	-	-
Unified-IO $_{\text{XL}}$	2.9B	79.1	-	-		122.3	-	-	-	-
Flamingo-3B	3.2B	-	-	-	-	-	65.9	89.3	48.0	79.5
Uni-Perceiver $_{\text{BASE}}$	124M	79.2	-	-	32.0	-	64.9	82.3	50.7	71.1
Uni-Perceiver $_{\text{LARGE}}$	354M	82.7	-	-	35.3	-	67.8	83.7	54.1	74.2
Uni-Perceiver-MoE $_{\text{BASE}}$	167M	80.3	-	-	33.2	-	64.6	82.1	51.6	72.4
Uni-Perceiver-MoE $_{\text{LARGE}}$	505M	83.4	-	-	35.5	-	67.9	83.6	55.3	75.9
Uni-Perceiver-v2 $_{\text{BASE}}$	308M	86.3	58.6	50.6	35.4	116.9	71.8	88.1	55.6	73.8
Uni-Perceiver-v2 $_{\text{LARGE}}$	446M	**87.2** (+3.8)	**61.9** (+15.4)	**53.6** (+15.4)	**36.5** (+1.6)	**122.5** (+0.2)	**75.0** (+7.1)	**89.3** (+0.0)	**58.5** (+3.2)	**79.6** (+0.1)

Comparison of Uni-Perceiver v2 with generalist models and commonly-recognized strong task-specific models



Ablation Studies

Ablation of task collaboration and interference

Tasks	COCO Detection	ImageNet-1k Classification	COCO Retrieval		COCO Caption
Single Task	50.1	76.1	50.0	37.6	30.2
All Tasks	49.8	76.3	46.0	34.7	28.9
w/o Detection	-	76.6 (+0.3)	47.0 (+1.0)	34.6 (−0.1)	30.4 (+0.5)
w/o Classification	50.1 (+0.3)	-	51.6 (+5.6)	38.6 (+3.9)	25.9 (−3.0)
w/o Retrieval	49.5 (−0.3)	76.3 (+0.0)	-	-	27.4 (−1.5)
w/o Captioning	49.7 (−0.1)	76.3 (+0.0)	51.2 (+5.2)	38.3 (+3.6)	-
All Tasks w/ MoE	49.9 (+0.1)	76.9 (+0.6)	51.3 (+5.3)	38.8 (+4.1)	30.6 (+0.7)

Ablation of different pre-trained image encoders

Pretrained Method	Pretrained Data	COCO Detection	ImageNet-1k Classification	COCO Retrieval		COCO Caption
Supervised	IN-1k	45.7	76.8	51.2	38.9	27.3
Supervised	IN-21k	48.3	**80.1**	55.1	41.2	30.2
Supervised	IN-1k & COCO	**49.9**	76.9	51.3	38.8	30.6
MoCo v2	IN-1k	48.3	75.0	54.8	40.5	29.6
CLIP	CLIP data	47.2	73.8	**55.3**	**41.3**	**32.0**

Ablation of different representation types

Representation Types	COCO Detection	ImageNet-1k Classification	COCO Retrieval		COCO Caption
Global	-	76.8	46.3	34.6	28.8
Regional	48.2	75.9	**52.3**	**39.2**	**31.2**
Global + Regional	**49.9**	**76.9**	51.3	38.8	30.6

Ablation of sampling and optimization strategies

Task Sampling	Gather Feature	TBGN	COCO Detection	ImageNet-1k Classification	COCO Retrieval		COCO Caption
mixed			49.6	76.7	40.1	31.9	27.6
unmixed			49.2	76.6	39.8	30.9	27.5
unmixed	✓		49.3	76.8	50.4	37.3	27.6
unmixed	✓	✓	**49.9**	**76.9**	**51.3**	**38.8**	**30.6**