Towards General Vision and Language Systems

Summary Report

By Chuyue Tang, Peking University

Motivation:
several open-ended questions

- What is the definition of ‘generality’?
- How much ‘generality’ can current models achieve?
- What should a general system be like?
- …

Contents

1. Multi-modal Pre-training

- Comparison among 5 models
- Generality
- Limitations of generality

Pre-training and Fine-tuning

Vision

ImageNet pre-trained CNN (Donahue et al., 2014)

Language

BERT (Devlin et al., 2019)

GPT-3 (Brown et al.,2020)

Vision and language?

Comparison: Pre-training

	Datasets (In-domain)	Datasets (Out-of-domain)	Amount (pairs)	Pre-training Tasks
UNITER [Y.-C. Chen et al., 2020]	• COCO-Cap • VG-Cap	• Conceptual Captions • SBU Captions	5.6M	• MLM • MVC+MVR+MVC-kl • SIA
LXMERT [Tan & Bansal, 2019]	• COCO-Cap • VG-Cap • VGQA+VQA +GQA	---	9.18M	• MLM • MVC+MVR • SIA • VQA
ViLBERT [Lu et al., 2019]	---	• Conceptual Captions	3.1M	• MLM • MVC • SIA
VisualBERT [Li et al., 2019b]	• COCO-Cap	---	3.3M (no direct mention)	• MLM • SIA
VL-BERT [Su et al., 2019]	---	• Conceptual Captions • BooksCorpus + Wikipedia	3.3M	• MLM • MVC

Comparison: Model

	Architecture	Visual Representation	Textual Representation	Para-meter
UNITER [Y.-C. Chen et al., 2020]	One cross-modal Transformer	RoI + location(7-D)	Token + position	86M
LXMERT [Tan & Bansal, 2019]	Two single-modal Transformer (respectively) + one co-attention Transformer	RoI + location	Token + position	183M
ViLBERT [Lu et al., 2019]	One single-modal Transformer (language) + one co-attention Transformer	RoI + location(5-D) + token[IMG]	Token + position + segment	221M
VisualBERT [Li et al., 2019b]	One cross-modal Transformer	RoI + location + segment	Token + position + segment	---
VL-BERT [Su et al., 2019]	One cross-modal Transformer	RoI + location(4-D) + token[IMG] + segment + position	Token + position + segment + RoI (whole image)	---

Comparison: Fine-tuning

	Downstream Tasks	**Epochs**	**Task-specific Pre-training**	**More Experiments**
UNITER [Y.-C. Chen et al., 2020]	VQA, VCR, NLVR2, VE, ITR, REC	5+	VCR	• Ablation study (pre-train tasks/datasets)
LXMERT [Tan & Bansal, 2019]	VQA, GQA, NLVR2	4	But early pre-train with VQA data	• Attention visualization • Ablation study (initialization, pre-train tasks/datasets, architecture)
ViLBERT [Lu et al., 2019]	VQA, VCR, REC, IR	20	---	• Ablation study (fusion, pre-train)
VisualBERT [Li et al., 2019b]	VQA, VCR, NLVR2, ITR(Flickr30K)	8~12	All tasks	• Attention visualization • Ablation study
VL-BERT [Su et al., 2019]	VQA, VCR, REC	20	---	• Ablation study (pre-train)

Similarities

- Extraction on **features**
  - Text: sub-word tokens -- WordPiece (Wu et al., 2016)
  - Image: region-of-interest (RoI) -- Faster R-CNN (Ren et al., 2015)
- Self-supervised pre-training **tasks**
  - inspired by BERT(Devlin et al., 2019)
- Transformer-based **architecture**
- **Large-scale pre-training** **dataset**
- …

Where is ‘Generality’?

Specific training	Task-specific architecture
	Task-specific representation
Large-scale pre-training	Task-agnostic + task-specific architecture
	Task-agnostic representation

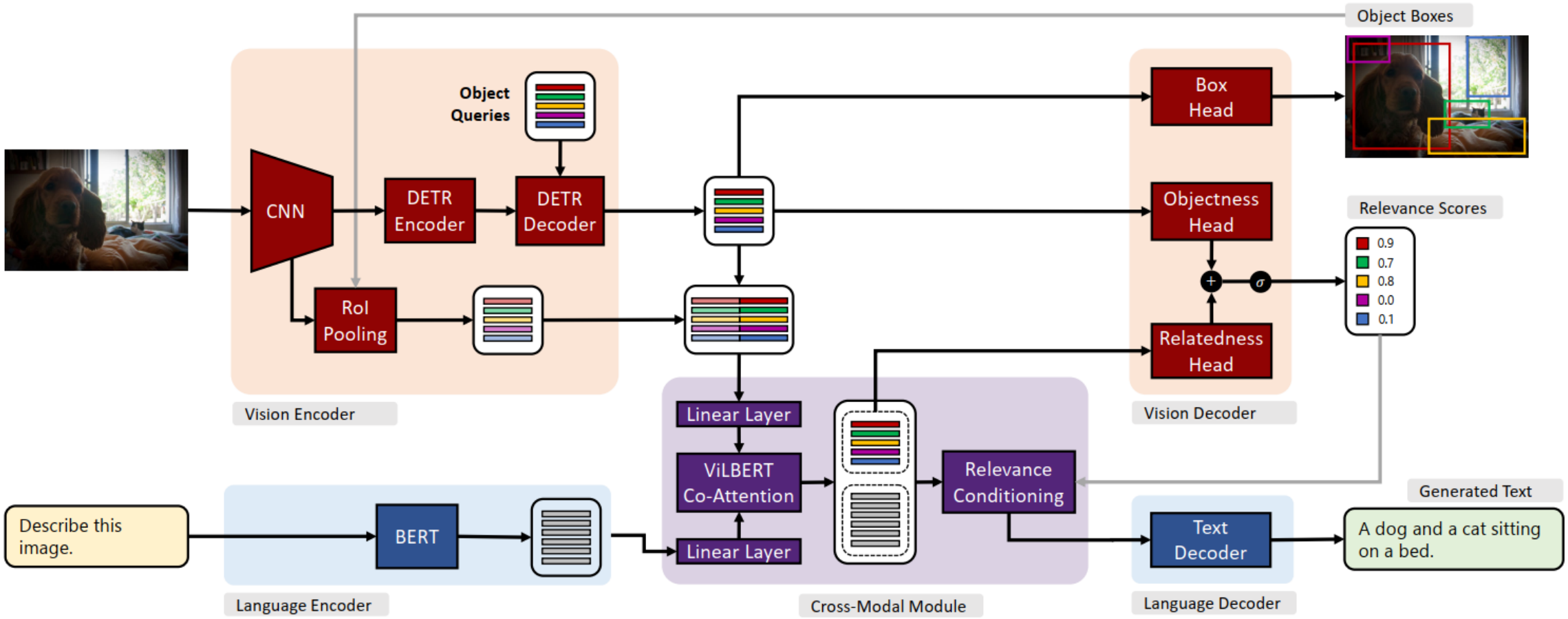
Limitations of ‘Generality’

- Pre-defined downstream tasks
- Task-specific output heads
- Task-specific parameters
- ‘Knowledge’ in the black box
- Data & Time consuming
- …

2. Multi-task Training

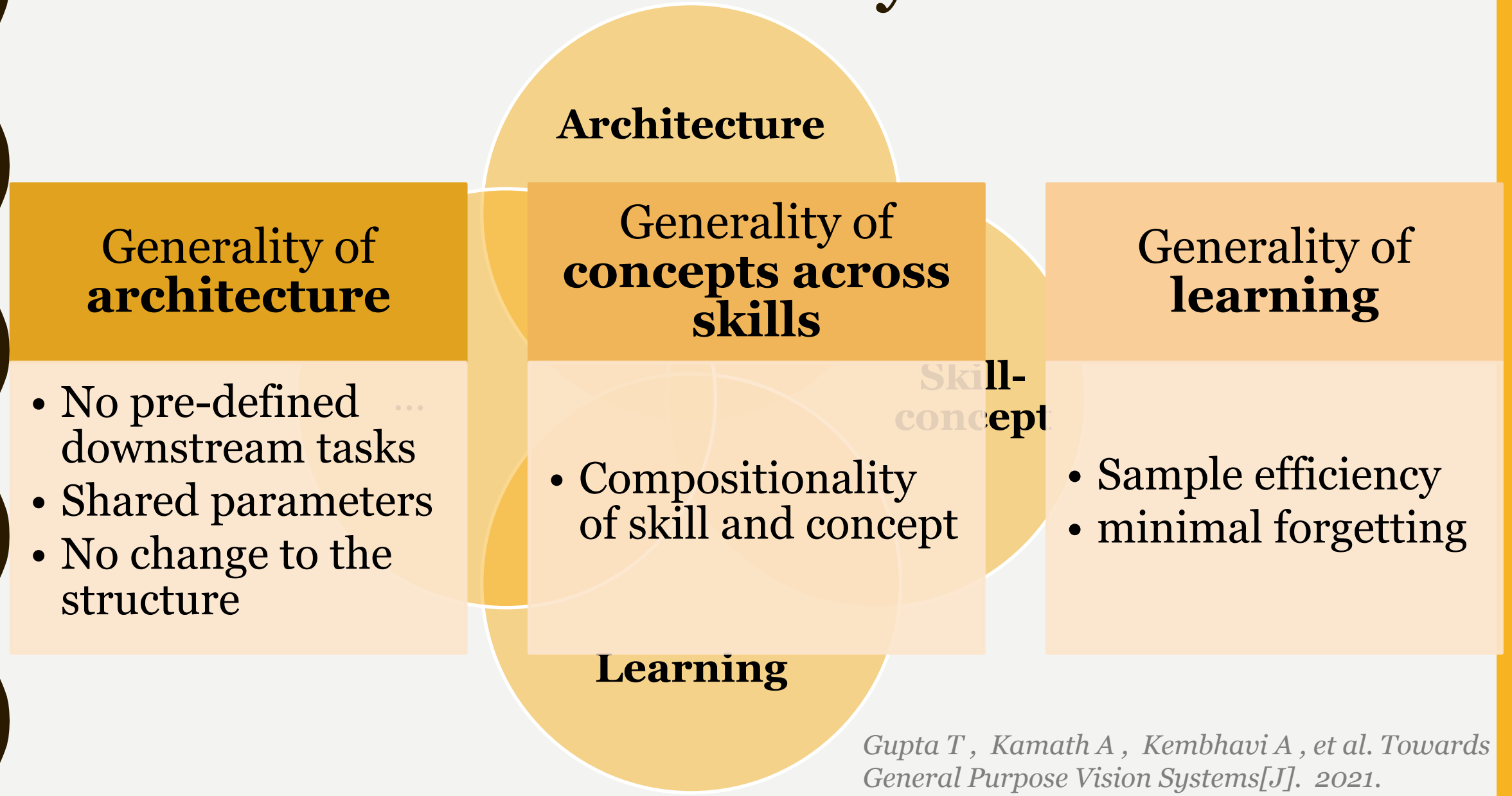
- GPV model
- Generality
- Limitations of generality

Model: GPV-I



Gupta T , Kamath A , Kembhavi A , et al. Towards General Purpose Vision Systems[J]. 2021.

Where is ‘Generality’?



Generality of **architecture**

- No pre-defined downstream tasks
- Shared parameters
- No change to the structure

Generality of **concepts across skills**

- Compositionality of skill and concept

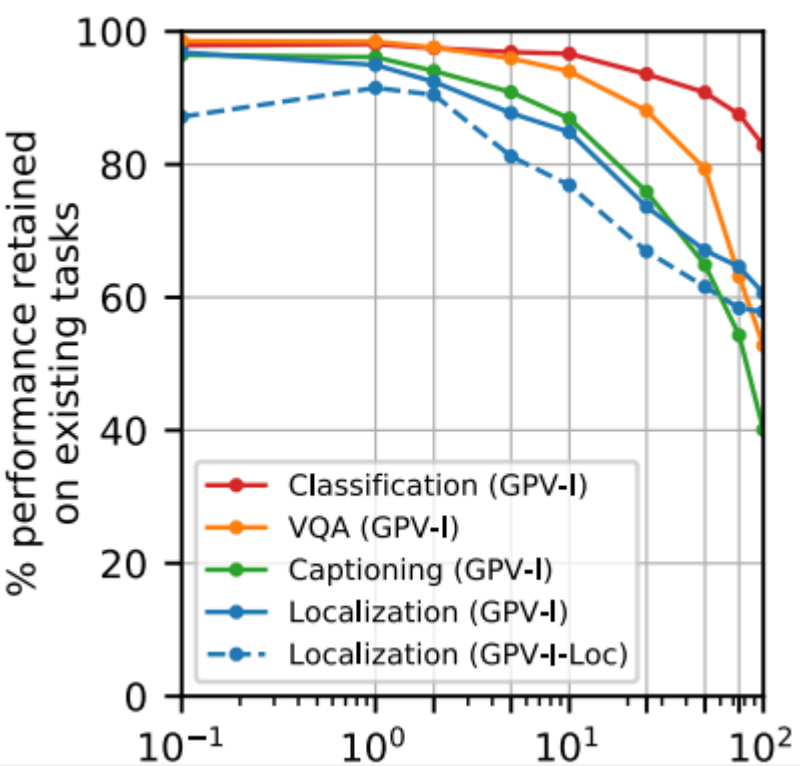
Generality of **learning**

- Sample efficiency
- minimal forgetting

Gupta T , Kamath A , Kembhavi A , et al. Towards General Purpose Vision Systems[J]. 2021.

Limitations of ‘Generality’

- **Catastrophic forgetting**

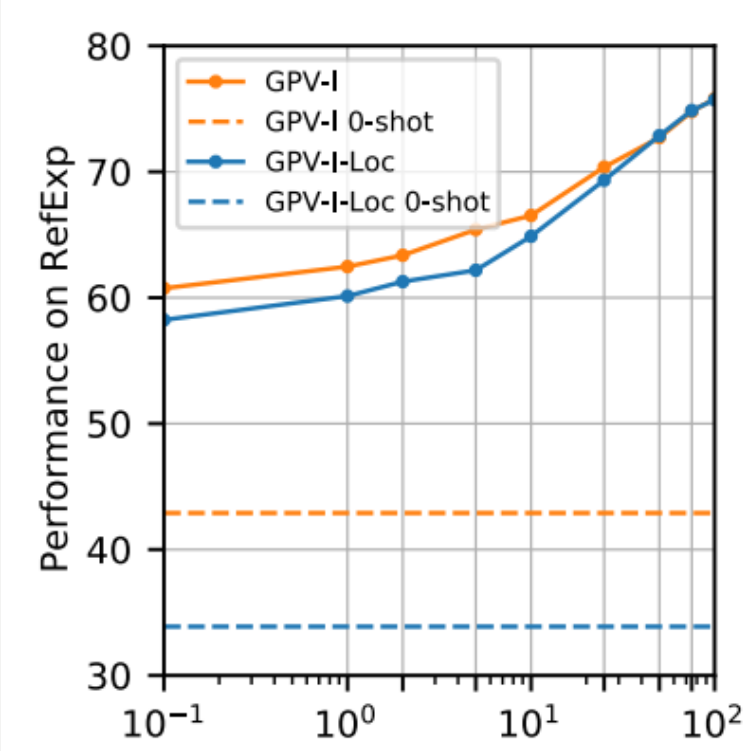


- Important weights are changed when new tasks introduced—protect such weights(synaptic consolidation)
- Implicit/explicit knowledge?
- Weight = knowledge?
- We learn and remember the principle, thus knowing how to solve a series of tasks.

Kirkpatrick J, Pascanu R, Rabinowitz N, et al. Overcoming catastrophic forgetting in neural networks[J]. Proceedings of the national academy of sciences, 2017, 114(13): 3521-3526.

Limitations of ‘Generality’

- Sample efficiency?—A rough estimation



Data usage	Performance
0% -> 0.1%	40+ -> 60+
0.1% -> 100%	60+ -> 80-

Gupta T , Kamath A , Kembhavi A , et al. Towards General Purpose Vision Systems[J]. 2021.

3. Towards General System

- Expectations

Toward General Purpose System

Special purpose

- Task-specific design
- Catastrophic forgetting
- Surplus data
- …

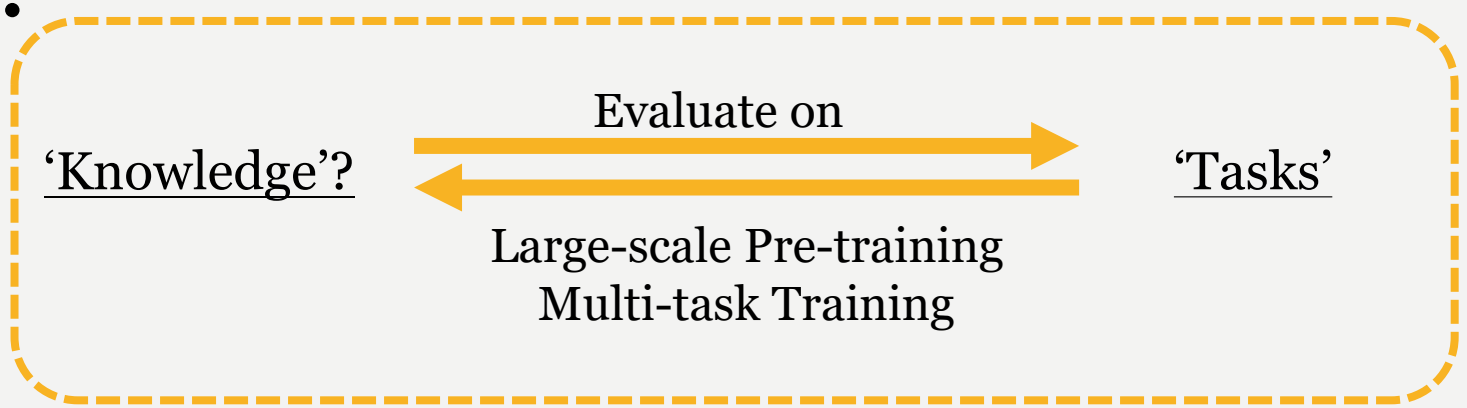
General purpose

- Domain/task-agnostic
- Comprehension and general knowledge
- Sample efficiency
- …

Toward General Purpose System

Humans:

Current models:



Final Words

- Does ‘generality’ only come from human intelligence?
- No absolute ‘better’ or ‘worse’?
- But always towards perfect systems! ☺

Thanks for your listening!