USC Viterbi
School of Engineering

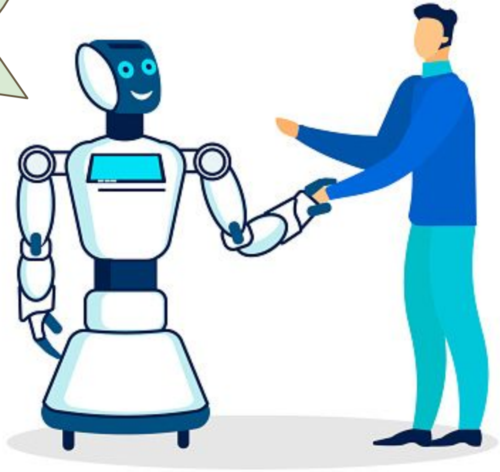
CLiMB
A Continual Learning Benchmark for Vision-and-Language Tasks

Tejas Srinivasan, Ting-Yun Chang, Leticia Pinto-Alva, Georgios Chochlakis, Mohammad Rostami, Jesse Thomason

Multimodal Agents that can be Deployed

Responses/actions by reasoning over inputs from multiple modalities

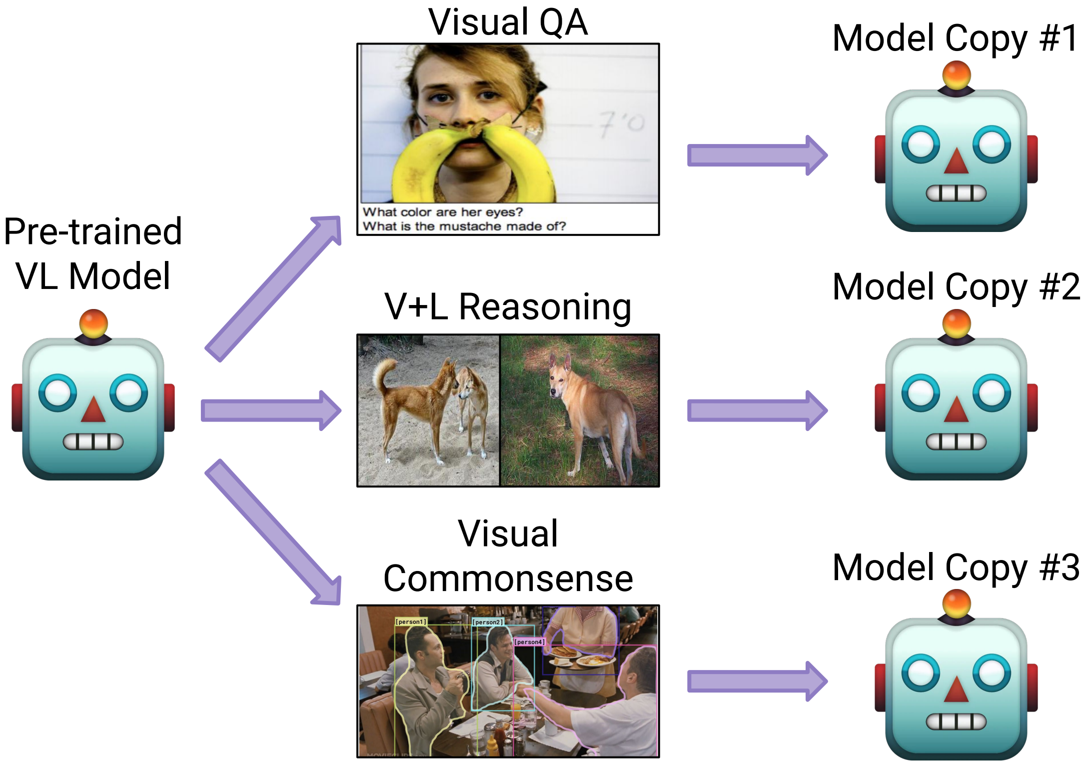
Visual QA + Dialog



Image/Scene Classification

Knowledge-based QA

Paradigms of VL Deployment: Single-Task Finetuning



Need to store a copy of the model for each task!

Paradigms of VL Deployment: Multi-Task Learning

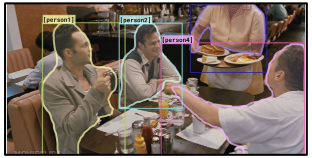
Pre-trained
VL Model

Visual QA

What color are her eyes?
What is the mustache made of?

V+L Reasoning

Visual
Commonsense



Deployed
Multi-task
VL Model

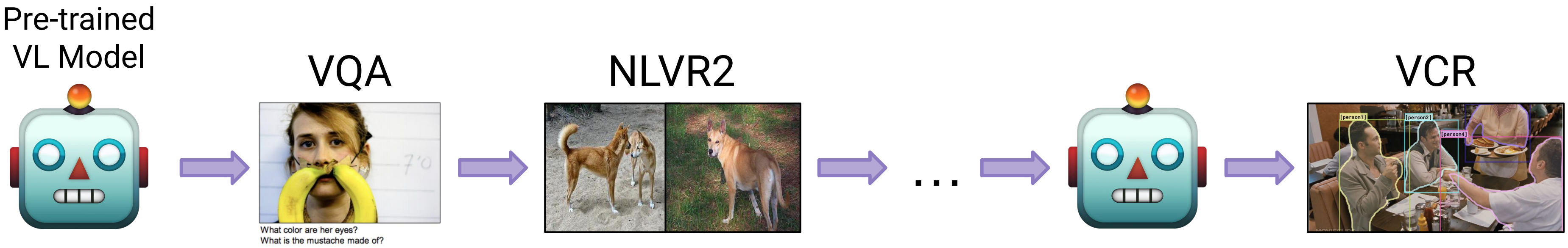
New Task
Visual Entailment

A man shoots a basketball
Contradiction

???

Static; cannot learn new tasks that it encounters!

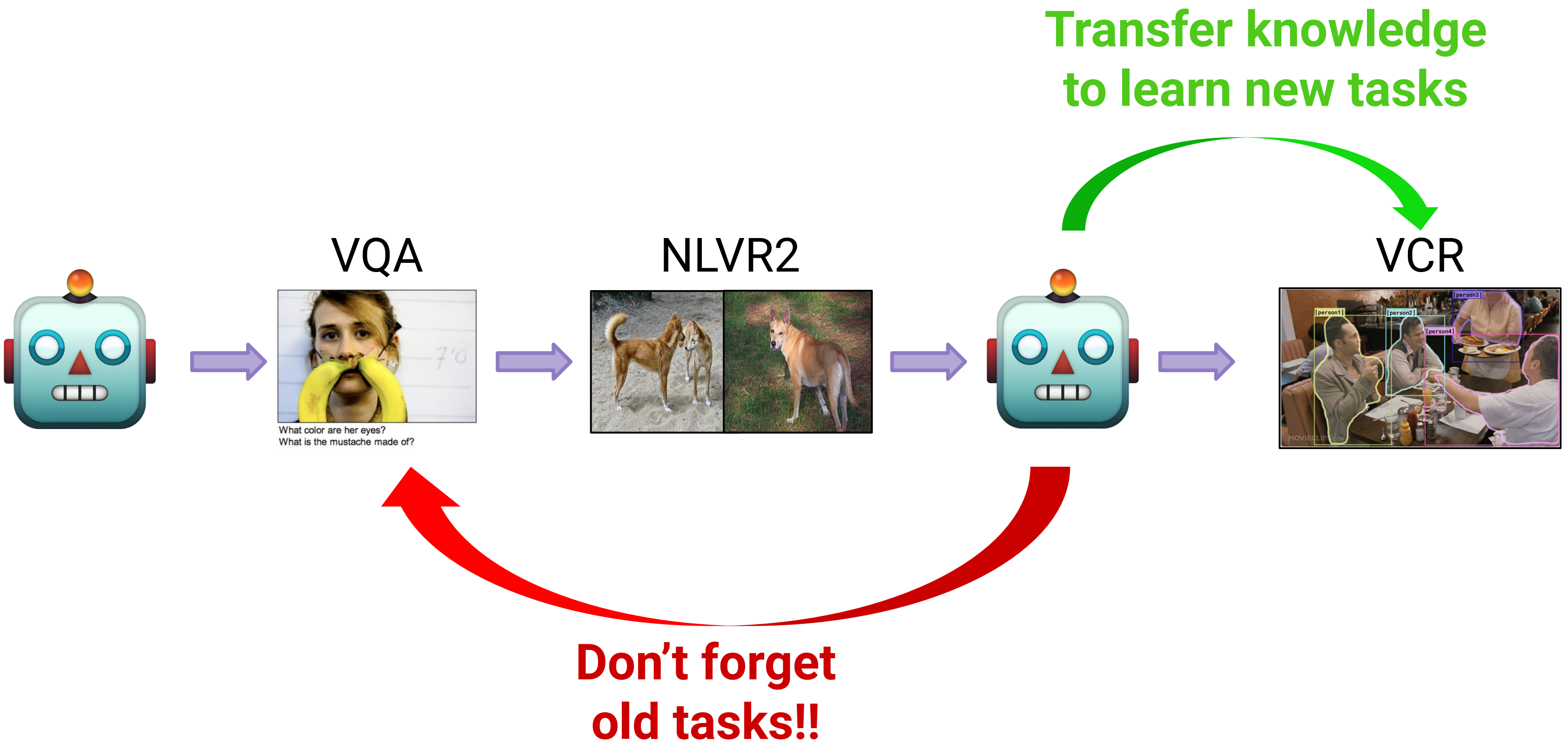
Paradigms of VL Deployment: Continual Learning



Dynamic, continually evolving paradigm

Unexplored in multimodal domain!

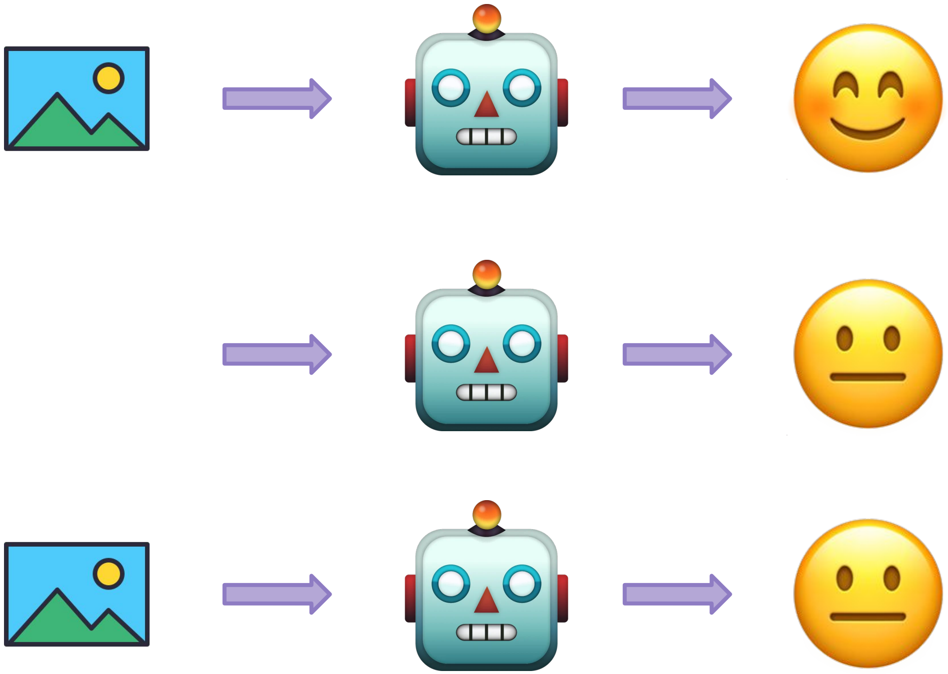
Challenges of Multimodal Continual Learning Deployment



USC Viterbi
School of Engineering

Challenges of Multimodal Continual Learning Deployment

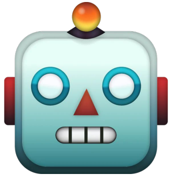
Not guaranteed to have all modalities when encountering new tasks!



CLiMB: The Continual Learning in Multimodality Benchmark

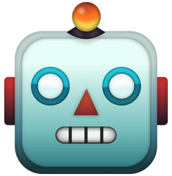
Phase I. Upstream Continual Learning

Pre-trained
VL Model

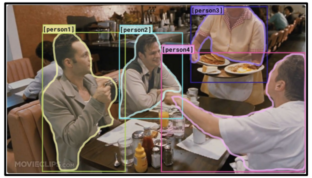


VQA

NLVR2

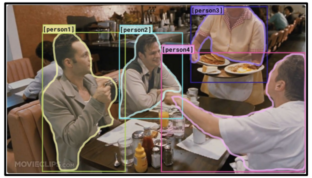


VCR



Phase II. Downstream Low-Shot Transfer

Unseen V+L Tasks



Language Tasks

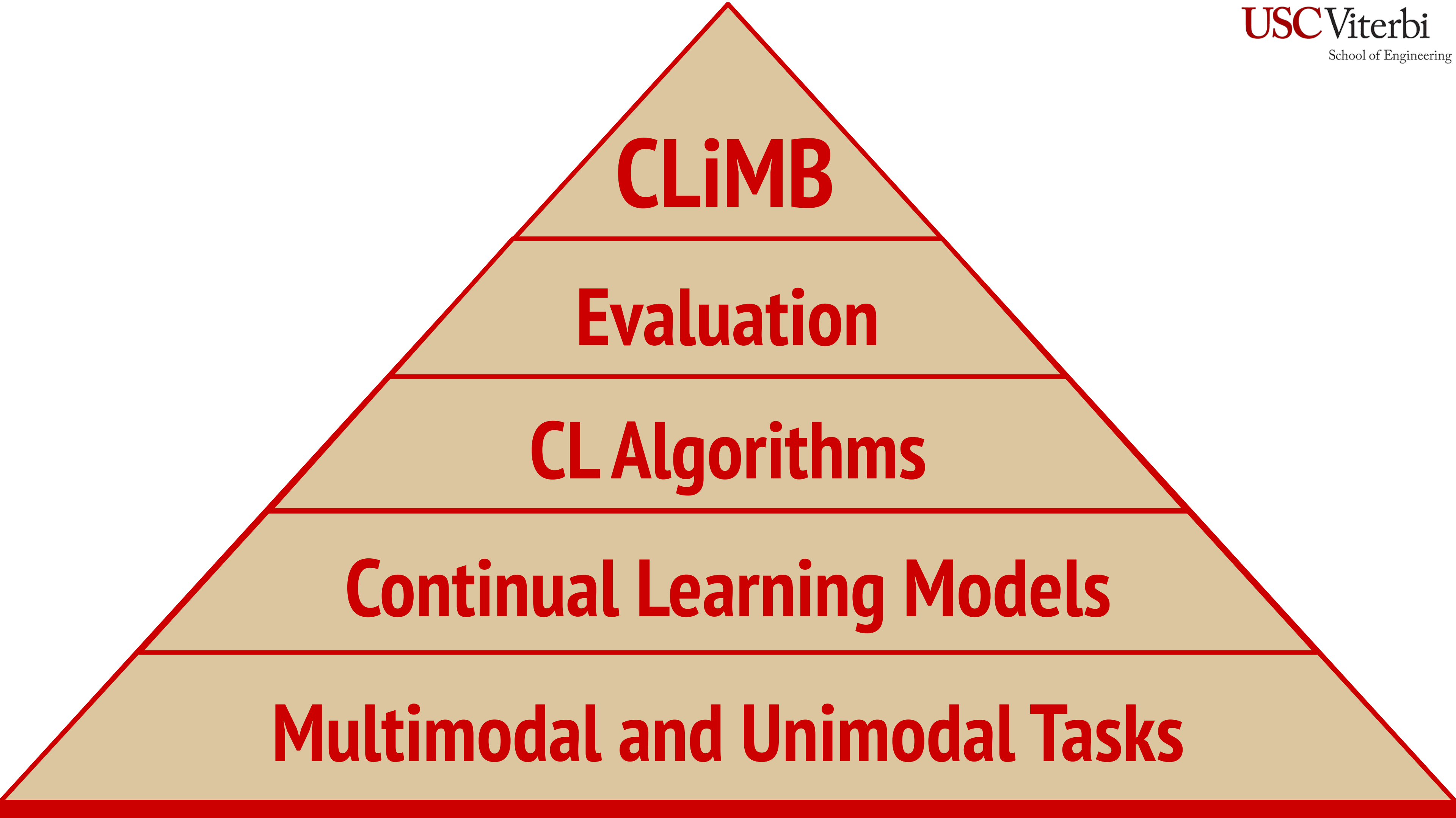
GLUE

Vision Tasks

IMAGENET

COCO
Common Objects in Context

USC Viterbi
School of Engineering

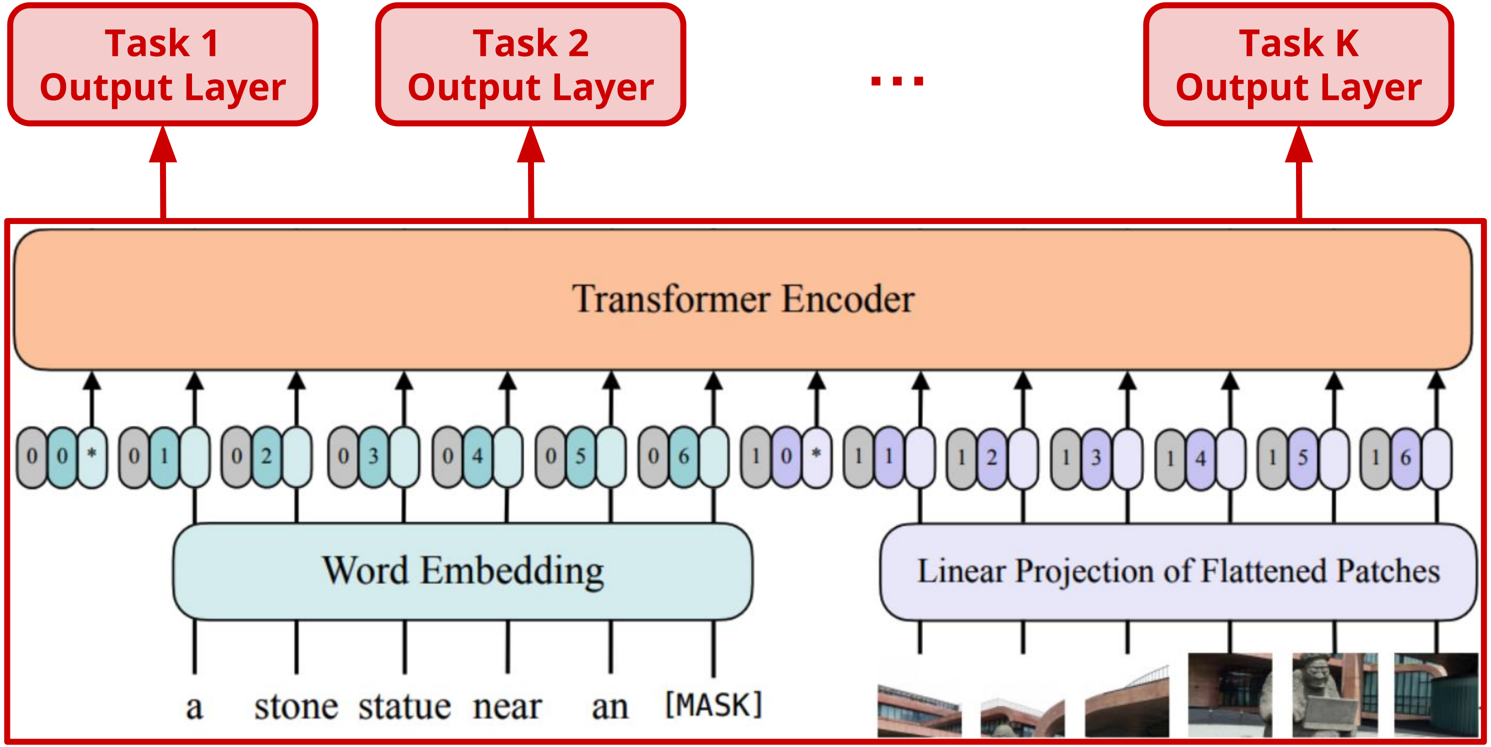


I. Multimodal and Unimodal Tasks

Vision-and-Language Tasks	• Visual Question Answering (VQAv2) • Natural Language Visual Reasoning (NLVR2) • Visual Entailment (SNLI-VE) • Visual Commonsense Reasoning (VCR)
Language-Only Tasks	• IMDb, SST-2 Sentiment Classification • HellaSwag • CommonsenseQA • Physical Interaction QA (PIQA)
Vision-Only Tasks	• ImageNet-1K Image Classification • iNaturalist2019 Image Classification • Places365 Image Classification • MS-COCO Object Detection

CLiMB can be easily extended to include new multimodal and unimodal tasks!

II. Continual Learning Models

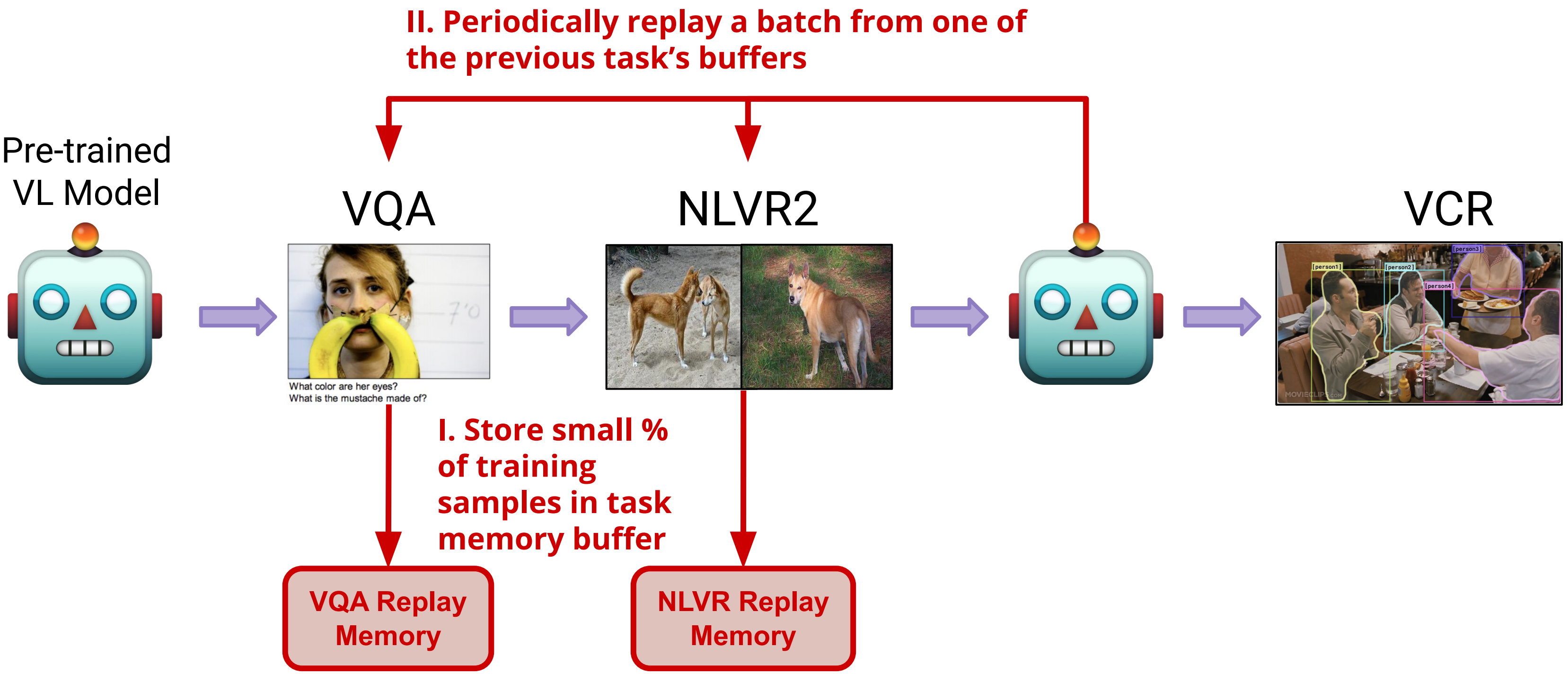


III. Continual Learning Algorithms

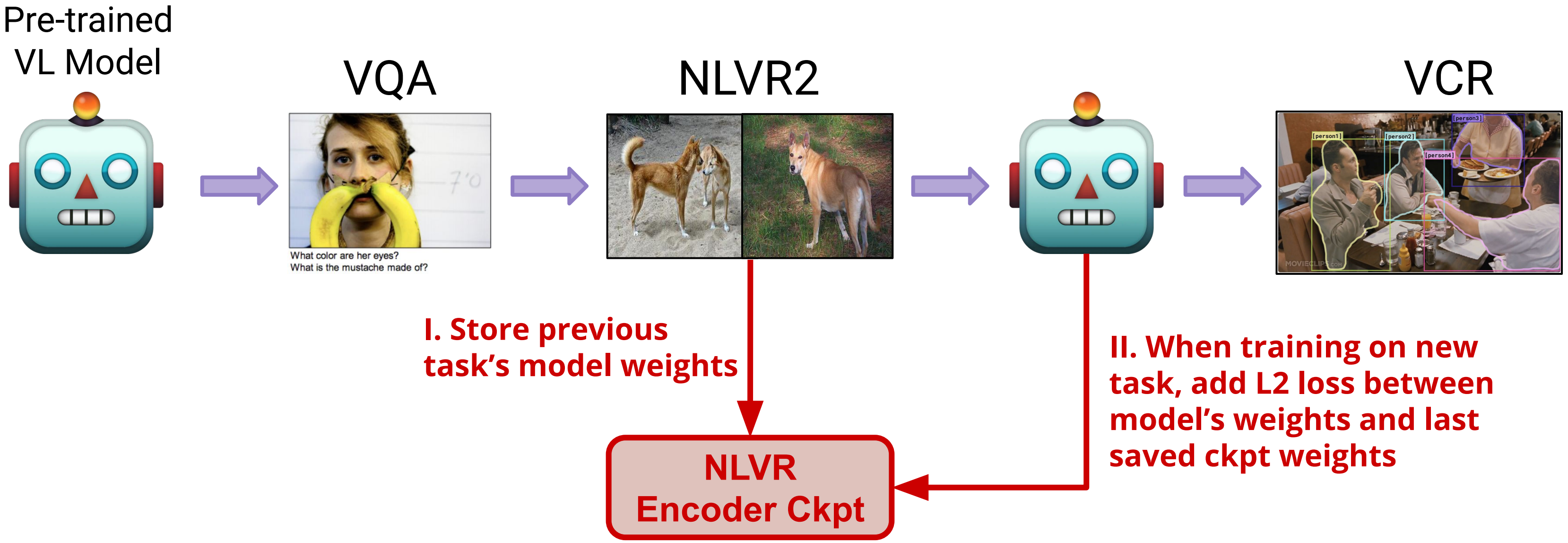
Currently, CLiMB supports 6 different Continual Learning algorithms:

- **Sequential Fine-tuning:** Fine-tune full encoder and task-specific layers
- **Frozen Encoder:** Train only task-specific layers
- **Frozen Bottom-K:** Fine-tune only top encoder layers and task layers
    - We set K=9
- **Experience Replay (ER)**
- **Elastic Weight Consolidation (EWC)**
- **Adapters**

Experience Replay



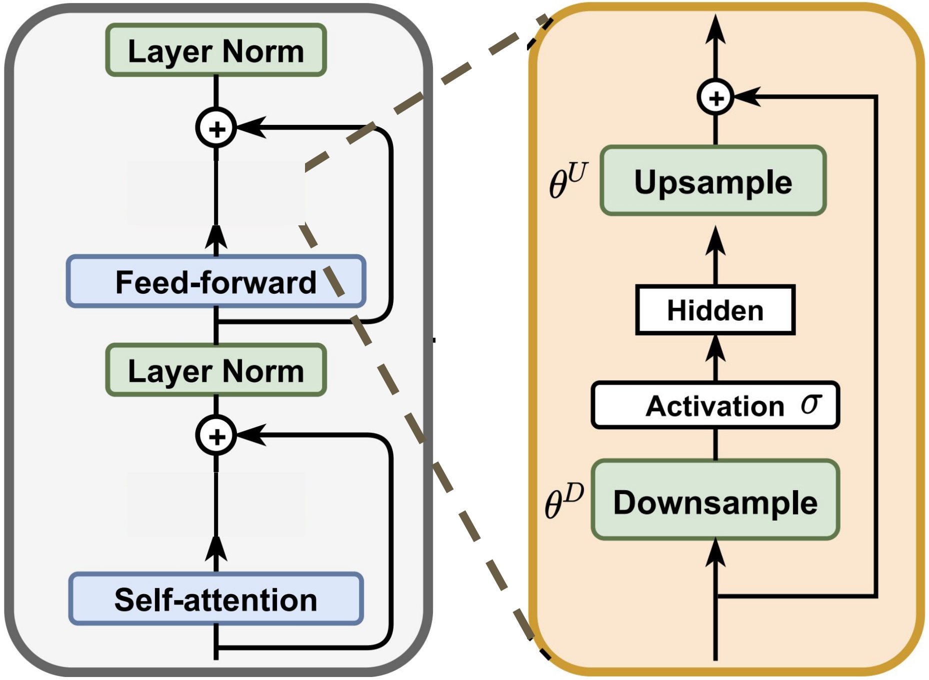
Elastic Weight Consolidation



Adapters

Insert new task-specific parameters into Transformer layers

- Transformer parameters kept frozen - **no forgetting!**
- **Fewer learnable parameters, faster to train**
- **Comparable performance as full model fine-tuning**
- **No cross-task knowledge transfer**

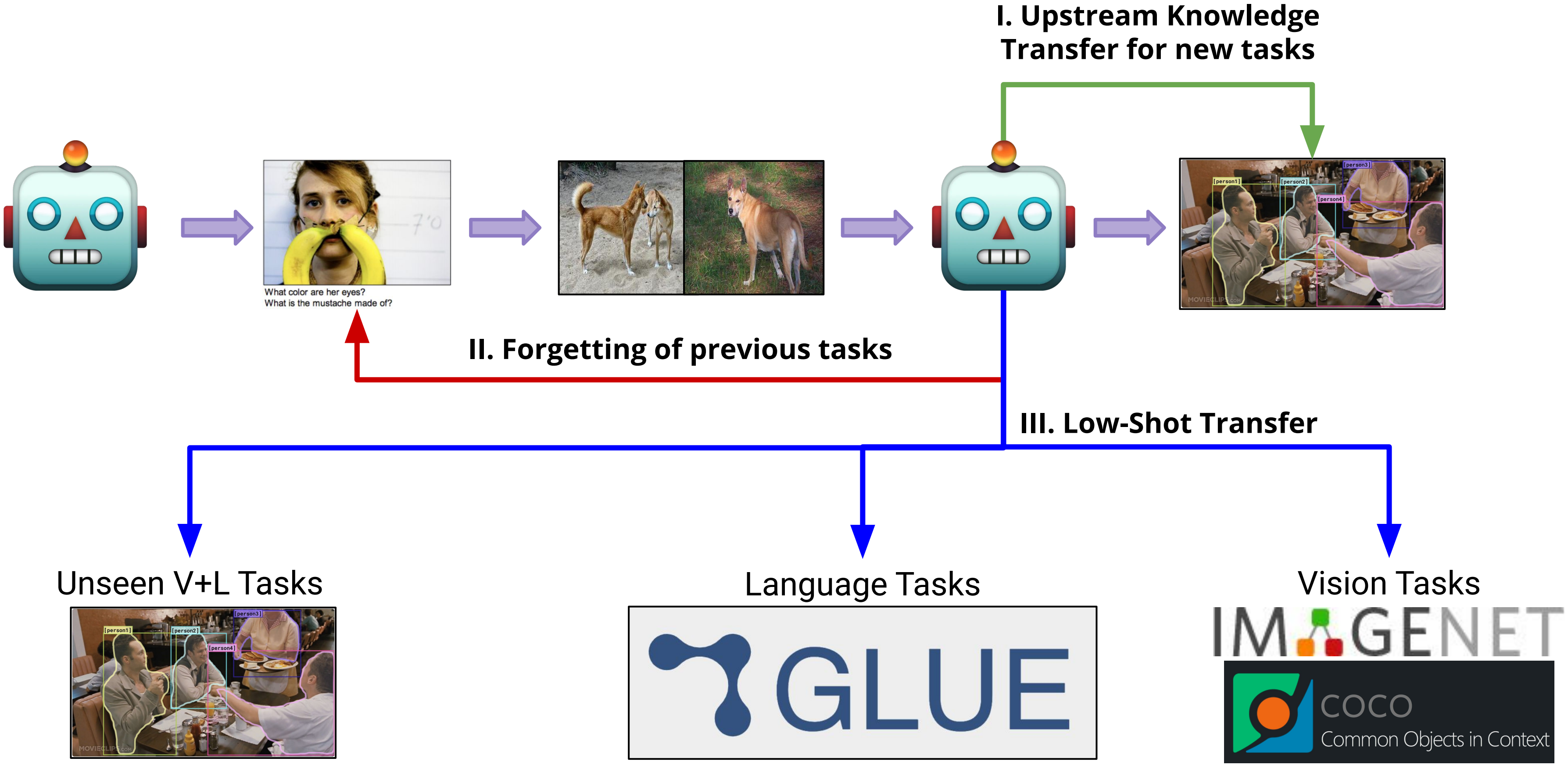


Transformer Layer

Adapter Module

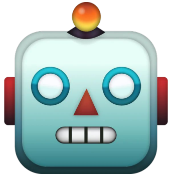
USC Viterbi
School of Engineering

IV. Evaluation



Upstream Evaluation I: Upstream Knowledge Transfer

With Continual Learning Algorithm A:

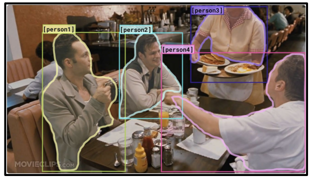


VQA

NLVR2

...

VCR

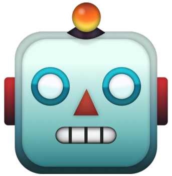


Score S^i_A

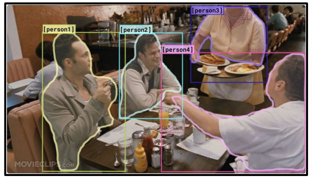
Tasks {1,...,i-1}

Task i

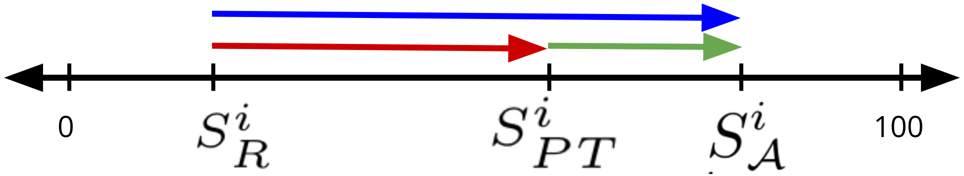
Without Continual Learning:



VCR



Score S^i_{PT}



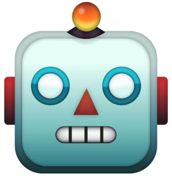
$$\mathbb{T}_{UK}(i) = \frac{S^i_A - S^i_{PT}}{S^i_{PT} - S^i_R}$$

USC Viterbi School of Engineering

Upstream Evaluation II: Forgetting Transfer

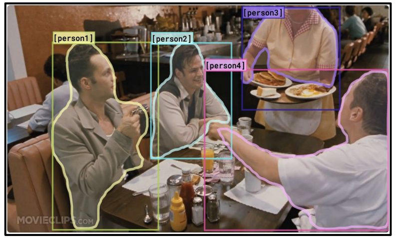
Directly evaluate on previously learned task j

Score $S_{\mathcal{A}}^{j \leftarrow i}$



...

Task j: VCR

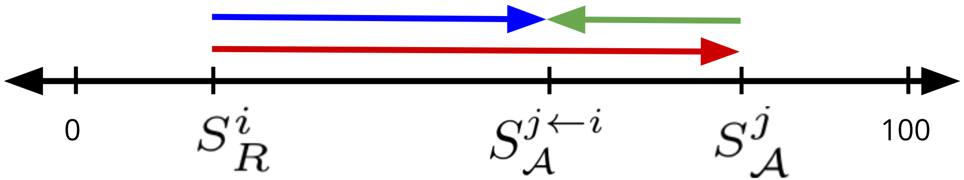


Score $S_{\mathcal{A}}^{j}$

...

Task i: SNLI-VE

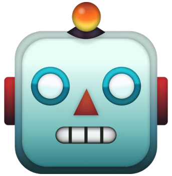
A man shoots a basketball
Contradiction



$$\mathbb{T}_F(j \leftarrow i) = \frac{S_{\mathcal{A}}^{j} - S_{\mathcal{A}}^{j \leftarrow i}}{S_{\mathcal{A}}^{j} - S_R^{j}}$$

Downstream Evaluation: Low-Shot Transfer

With Continual Learning Algorithm $\mathcal{A}$:

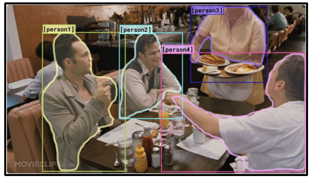


VQA

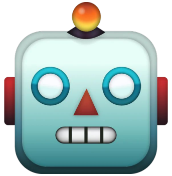
What color are her eyes?
What is the mustache made of?

...

VCR



Task i

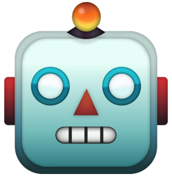


Low-Shot V+L/V/L Task

IMAGENET

Score $S_{\mathcal{A}}^{LS(i)}$

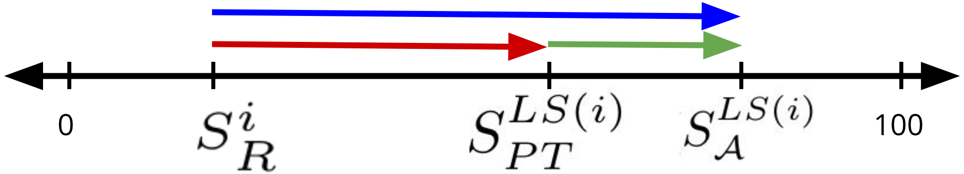
Without Continual Learning:



Low-Shot V+L/V/L Task

IMAGENET

Score $S_{PT}^{LS(i)}$



$$\mathrm{T}_{LS}^{M}(i) = \frac{S_{\mathcal{A}}^{LS(i)} - S_{PT}^{LS(i)}}{S_{PT}^{LS(i)} - S_R^i}$$

USC Viterbi School of Engineering

Experiments I: Upstream Continual Learning

- 4 V+L Tasks, ordered **VQA → NLVR2 → SNLI-VE → VCR**
- **ViLT**-based continual learning model
- **6** different Continual Learning algorithms

USC Viterbi School of Engineering

Results I: Upstream Continual Learning

Upstream Knowledge Transfer: How does Continual Learning affect model's ability to learn newly arriving tasks?

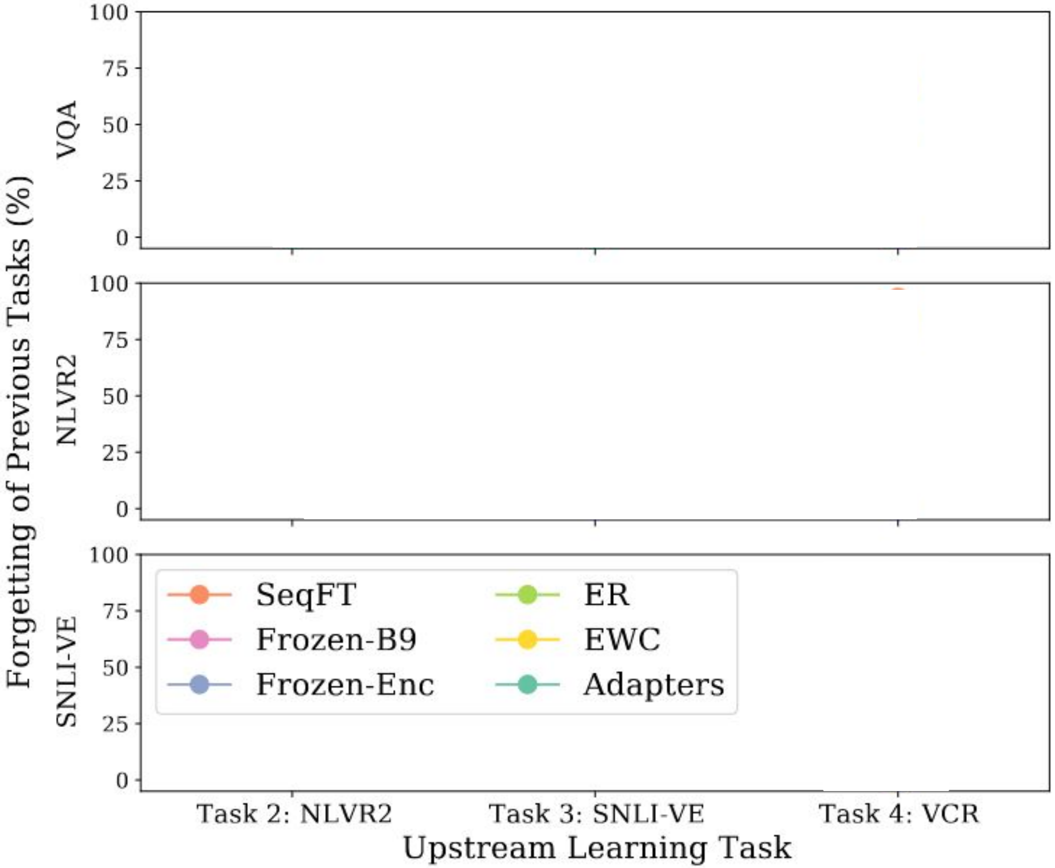
Alg $\mathcal{A}$	Params Trained	Task 1 VQAv2	Task 2 NLVR2	Task 3 SNLI-VE	Task 4 VCR
Direct FT	100%	[67.70]	[73.07]	[76.31]	[61.31]

- **More continual learning hurts ability to learn new tasks**
- **Adapters do not show negative transfer**, comparable to full model fine-tuning

Results I: Upstream Continual Learning

Forgetting: How does learning new tasks affect model's performance on already-learned tasks?

- **More fine-tuned params == more forgetting**
- **ER > EWC**
- **Adapters >>>>**
- **Forgetting more severe after VCR**

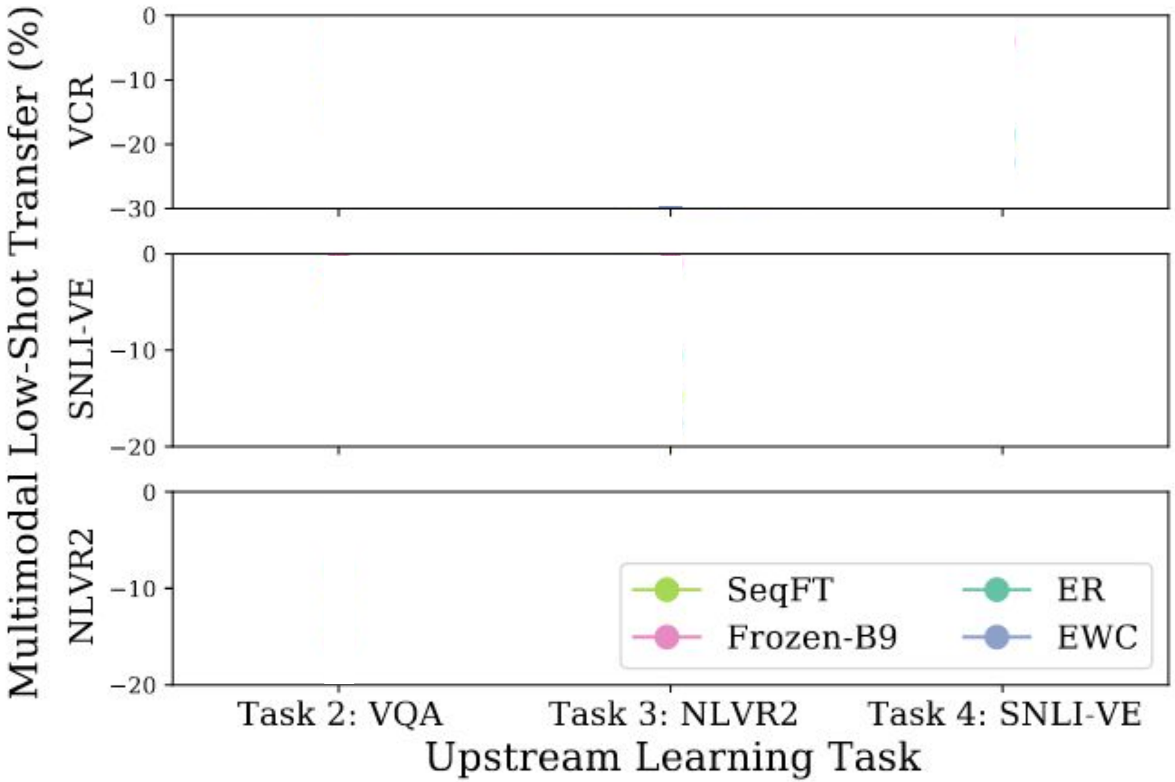


Effect of Continual Learning Task Order

Experiments and Results II: Downstream Low-Shot Transfer

Low-Shot Transfer to Unseen V+L Tasks

- **Low-Shot transfer is always negative**
- **Unsurprising — CL also hurts model transfer with full training data**

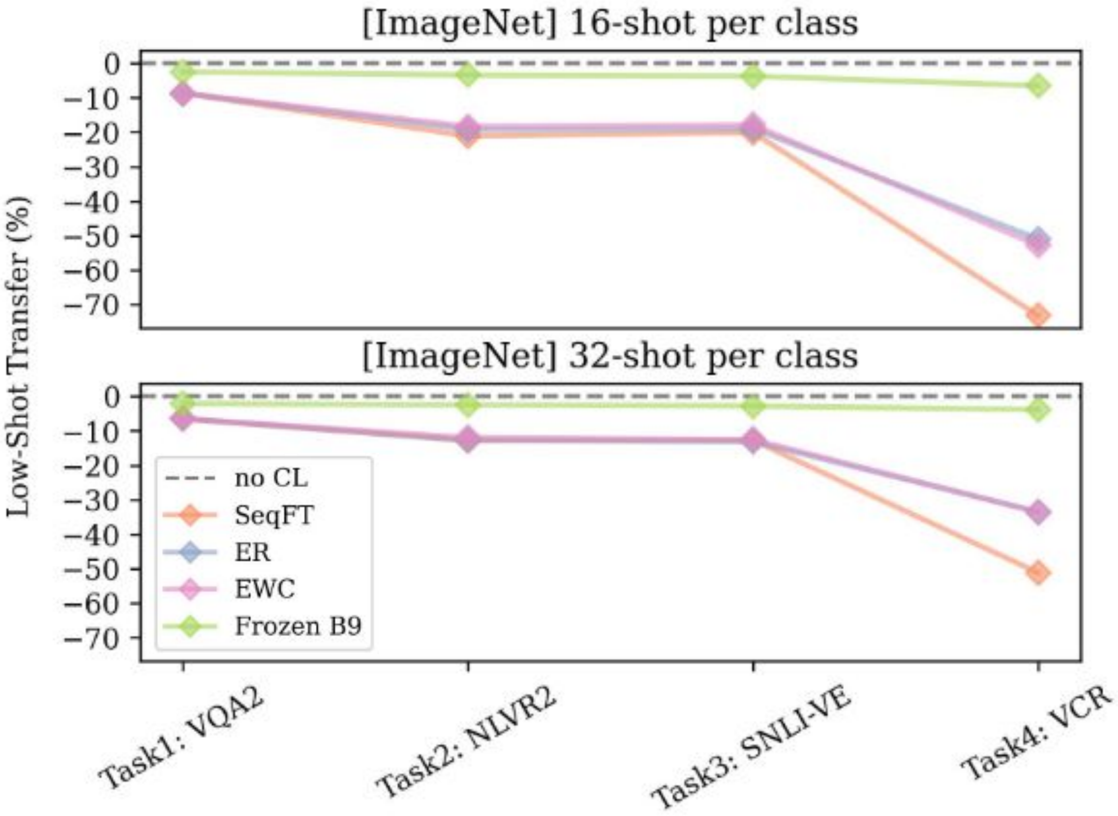


Experiments and Results II: Downstream Low-Shot Transfer

Low-Shot Transfer to Vision-Only Tasks

Language prompt: "This is an image."

- **ViLT achieves good low-shot performance on vision tasks**
- **CL hurts low-shot transfer**
- **NLVR2 and VCR have more negative effect**

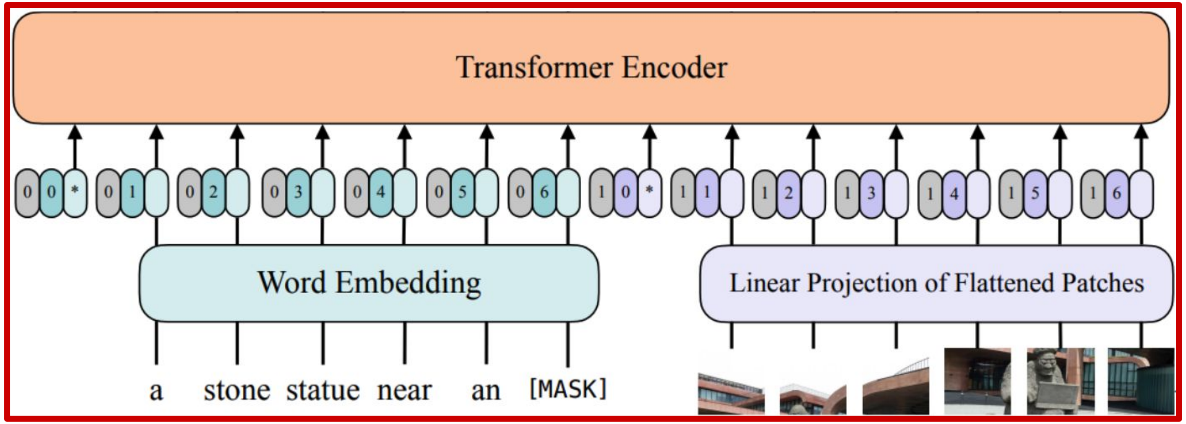


Experiments and Results II: Downstream Low-Shot Transfer

Low-Shot Transfer to Language-Only Tasks

Adapting ViLT for NLP tasks:

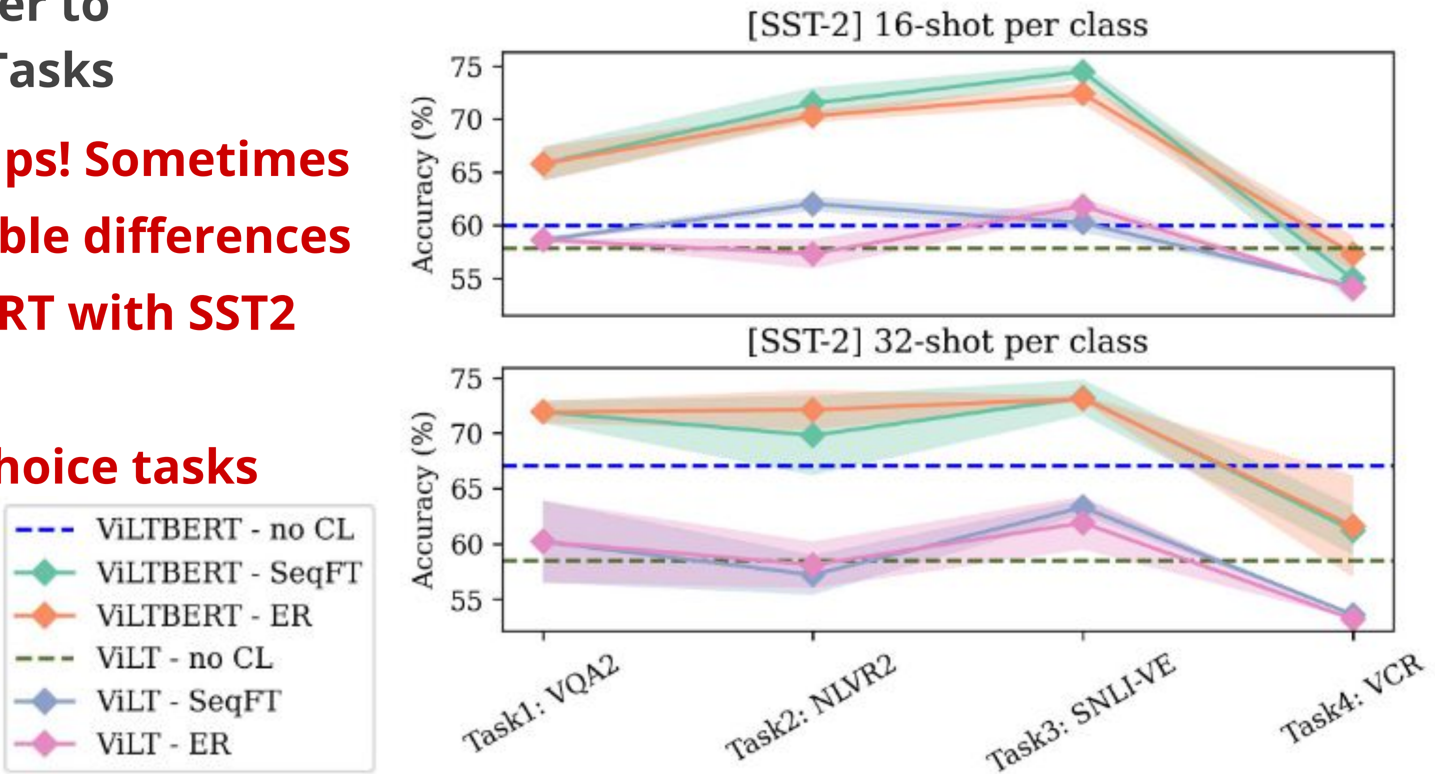
- Use **“average” MS-COCO image** for in-distribution visual input
- **Extend** language position embeddings
- **ViLT-BERT:** Replace language input embeddings with BERT representations



Experiments and Results II: Downstream Low-Shot Transfer

Low-Shot Transfer to Language-Only Tasks

- Upstream CL helps! Sometimes
- ViLT sees negligible differences
- CL helps ViLT-BERT with SST2
- VCR hurts SST2
- CL hurts multi-choice tasks



Conclusions

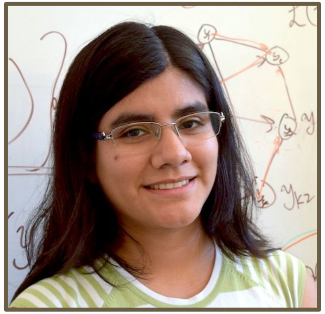
- We propose **CLiMB**, a benchmark to study CL in multimodal settings
- CLiMB is an **extensible community tool** for studying tasks, model architectures, and CL algorithms.
- **Existing Continual Learning methods fail** at:
  - generalizing well to sequences of multimodal tasks
  - Enabling low-shot adaptation to multi/unimodal tasks
- **Adapters are most effective** at preserving pre-trained model knowledge and forgetting mitigating
- There is **a need for new research** into Continual Learning strategies for this challenging multimodal setting.

Future Directions

- **Adapters that share knowledge across tasks**
- **Multimodal Adapters**
- Studying **multimodal distribution shifts**
- Building a **task-agnostic** modeling framework:
  - Sequence-to-sequence task formulations
  - Integrating **generalist models** into CLiMB
  - **Embodied** navigation, task completion

Acknowledgements

Ting-Yun Chang✨



Leticia Pinto Alva✨

Georgios Chochlakis

Mohammad Rostami

Jesse Thomason ✨

Thank You!!

https://github.com/GLAMOR-USC/CLiMB