Measuring Progress in Fine-grained Vision-and-Language Understanding
Coarse-grained vs. Fine-grained Tasks
Coarse-grained vs. Fine-grained Tasks

Coarse-grained Image Retrieval

A person is riding a horse.
Coarse-grained vs. Fine-grained Tasks

Coarse-grained Image Retrieval

A person is riding a horse.

Fine-grained Verb Understanding

A man \textit{jumping} into a river.

\begin{center}
\begin{tabular}{ll}
\text{Pos} & man, jump, river \\
\text{Neg} & man, kayak, river
\end{tabular}
\end{center}
What Matters for Fine-grained V&L Understanding?
What Matters for Fine-grained V&L Understanding?

Which models perform well on fine-grained tasks?

Localisation modelling > more Web data alone
What Matters for Fine-grained V&L Understanding?

Which models perform well on fine-grained tasks?
Localisation modelling > more Web data alone

How do data and losses impact fine-grained understanding?
Both data and losses needed; data diversity also matters
What Matters for Fine-grained V&L Understanding?

Which models perform well on fine-grained tasks?
Localisation modelling > more Web data alone

How do data and losses impact fine-grained understanding?
Both data and losses needed; data diversity also matters

How does fine-grained understanding evolve during training?
Performance can fluctuate during training, even becoming worse
Benchmarks

Fine-grained Tasks
Benchmarks

Fine-grained Tasks

- VALSE

6 phenomena: existence, plurality, counting, relations, actions, coreference
Benchmarks

Fine-grained Tasks

- VALSE
- VSR

65 relationships in 7 different categories (e.g., adjacency, proximity)
Benchmarks

Fine-grained Tasks

- VALSE
- VSR
- SVO-Probes

421 verbs with hard negatives for different parts of speech (subject, verb, object)
Benchmarks

Fine-grained Tasks

- VALSE
- VSR
- SVO-Probes
- Winoground

Tests a compositionality across 6 linguistic and visual phenomena
Benchmarks

Fine-grained Tasks
- VALSE
- VSR
- SVO-Probes
- Winoground

Coarse-grained Retrieval Tasks (Flickr30K, COCO)
Baselines
Baselines

Coarse-grained Models

- ALBEF (baseline)
Baselines

Coarse-grained Models

- ALBEF (baseline)
- BLIP (~ALBEF but w/ autoregressive LM)
Baselines

Coarse-grained Models

- ALBEF (baseline)
- BLIP (~ALBEF but w/ autoregressive LM)

Fine-grained Models

Newly proposed fine-grained models do not test on fine-grained tasks!
Baselines

Coarse-grained Models

- ALBEF (baseline)
- BLIP (~ALBEF but w/ autoregressive LM)

Fine-grained Models

- PEVL (ALBEF + bbox MLM)
Baselines

Coarse-grained Models
- ALBEF (baseline)
- BLIP (~ALBEF but w/ autoregressive LM)

Fine-grained Models
- PEVL (ALBEF + bbox MLM)
- X-VLM (ALBEF + bbox regression)
Baselines

Coarse-grained Models

- ALBEF (baseline)
- BLIP (~ALBEF but w/ autoregressive LM)

Fine-grained Models

- PEVL (ALBEF + bbox MLM)
- X-VLM (ALBEF + bbox regression)

Other coarse-grained Models (BLIP-2, ClipCap, Flamingo)
What Matters for Fine-grained V&L Understanding?

Which models perform well on fine-grained tasks?

How do data and losses impact fine-grained understanding?

How does fine-grained understanding evolve during training?
Which models perform well on fine-grained tasks?
Which models perform well on fine-grained tasks?

~ALBEF w/ autoregressive LM
Which models perform well on fine-grained tasks?

ALBEF + bbox prediction in MLM
Which models perform well on fine-grained tasks?

ALBEF + bbox regression head
Which models perform well on fine-grained tasks?

Localisation can help fine-grained understanding
Which models perform well on fine-grained tasks?

Localisation can help fine-grained understanding
But localisation loss matters!
Which models perform well on fine-grained tasks?

More data does not help as much as modelling
What Matters for Fine-grained V&L Understanding?

Which models perform well on fine-grained tasks?
Localisation modelling > more Web data alone

How do data and losses impact fine-grained understanding?

How does fine-grained understanding evolve during training?
What Matters for Fine-grained V&L Understanding?

Which models perform well on fine-grained tasks?
Localisation modelling > more Web data alone

How do data and losses impact fine-grained understanding?

How does fine-grained understanding evolve during training?
Data and Losses for Fine-grained Tasks

X-VLM adds 3 supervised datasets and 2 additional losses to ALBEF
Data and Losses for Fine-grained Tasks

X-VLM adds 3 supervised datasets and 2 additional losses to ALBEF
Data and Losses for Fine-grained Tasks

X-VLM adds 3 supervised datasets and 2 additional losses to ALBEF

- Object detection
  - COCO\textsubscript{OD}
  - VG\textsubscript{OD}

- Region description
  - VG\textsubscript{RD}
Data and Losses for Fine-grained Tasks

X-VLM adds 3 supervised datasets and 2 additional losses to ALBEF

- Object detection
  - COCO\textsubscript{OD}
  - VG\textsubscript{OD}

- Region description
  - VG\textsubscript{RD}

VG\textsubscript{RD} is the most useful dataset

Similar performance to training on all datasets
Data and Losses for Fine-grained Tasks

X-VLM adds 3 supervised datasets and 2 additional losses to ALBEF
Data and Losses for Fine-grained Tasks

X-VLM adds 3 supervised datasets and 2 additional losses to ALBEF
Data and Losses for Fine-grained Tasks

X-VLM adds 3 supervised datasets and 2 additional losses to ALBEF

object-centric visual view: an image region (not the whole image) is used
Data and Losses for Fine-grained Tasks

X-VLM adds 3 supervised datasets and 2 additional losses to ALBEF
X-VLM adds 3 supervised datasets and 2 additional losses to ALBEF

- Just adding supervised data does not help
Data and Losses for Fine-grained Tasks

X-VLM adds 3 supervised datasets and 2 additional losses to ALBEF

- Just adding supervised data does not help
- $L_{VMA}$ is slightly more helpful than $L_{BBOX}$
Data and Losses for Fine-grained Tasks

X-VLM adds 3 supervised datasets and **2 additional losses** to ALBEF

- Just adding supervised data does not help
- $L_{VMA}$ is slightly more helpful than $L_{BBOX}$
- $L_{VMA} + L_{BBOX}$ is best
What Matters for Fine-grained V&L Understanding?

Which models perform well on fine-grained tasks?
Localisation modelling > more Web data alone

How do data and losses impact fine-grained understanding?
Both data and losses needed; data diversity also matters

How does fine-grained understanding evolve during training?
What Matters for Fine-grained V&L Understanding?

Which models perform well on fine-grained tasks?
Localisation modelling > more Web data alone

How do data and losses impact fine-grained understanding?
Both data and losses needed; data diversity also matters

How does fine-grained understanding evolve during training?
Different Skills, Different Patterns
Different Skills, Different Patterns

A single checkpoint might not be adequate for all skills!
What Matters for Fine-grained V&L Understanding?

Which models perform well on fine-grained tasks?
Localisation modelling > more Web data alone

How do data and losses impact fine-grained understanding?
Both data and losses needed; data diversity also matters

How does fine-grained understanding evolve during training?
Performance can fluctuate during training, even becoming worse
Conclusion

Strong multimodal models trained at scale struggle with fine-grained understanding

- **Supervised losses** are promising
- As is **descriptive language** (region descriptions)

Fine-grained skills are learned at different times

- Pay attention to learning dynamics!
- How can we consistently improve over all fine-grained skills?