Interpreting and Grounding Pre-trained Representations for NLP

Richard Johansson and Lovisa Hagström

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

CHALMERS | UNIVERSITY OF GOTHENBURG

Disclaimer!

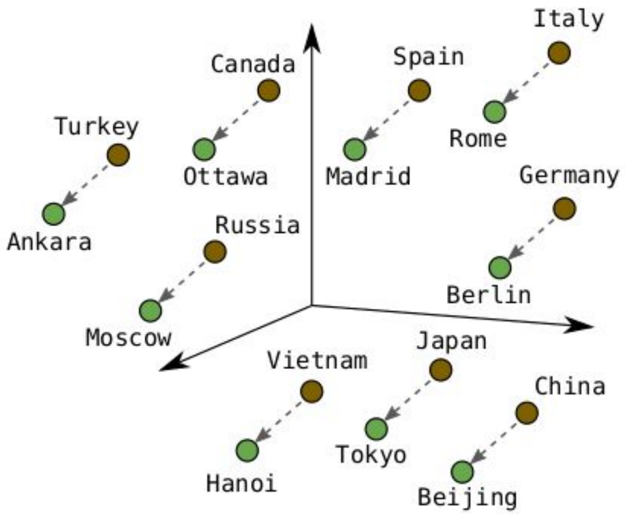
More ideas than results!

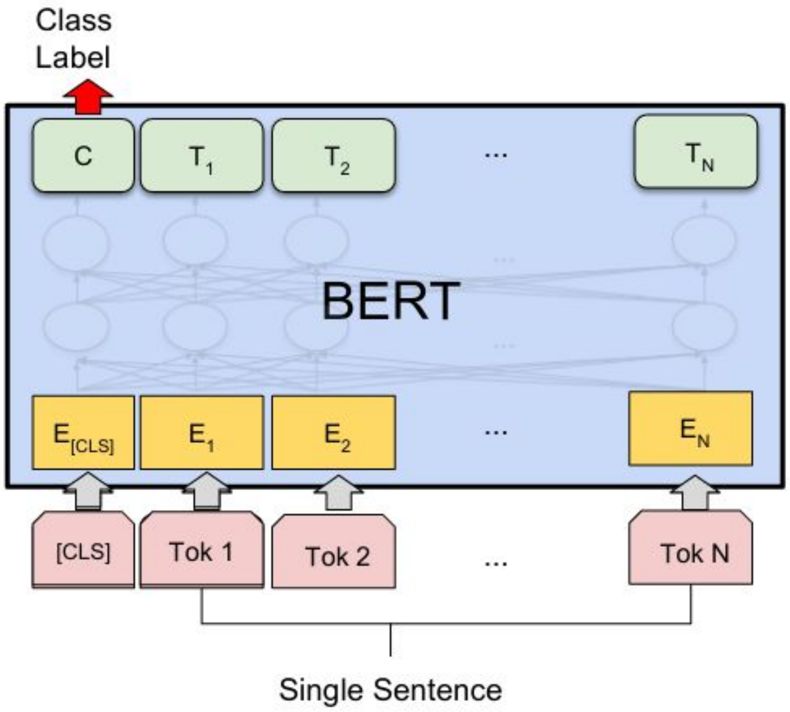
More questions than answers!



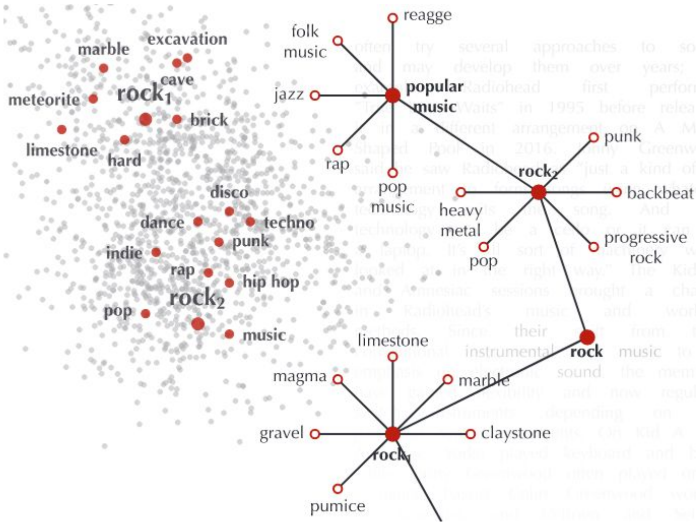


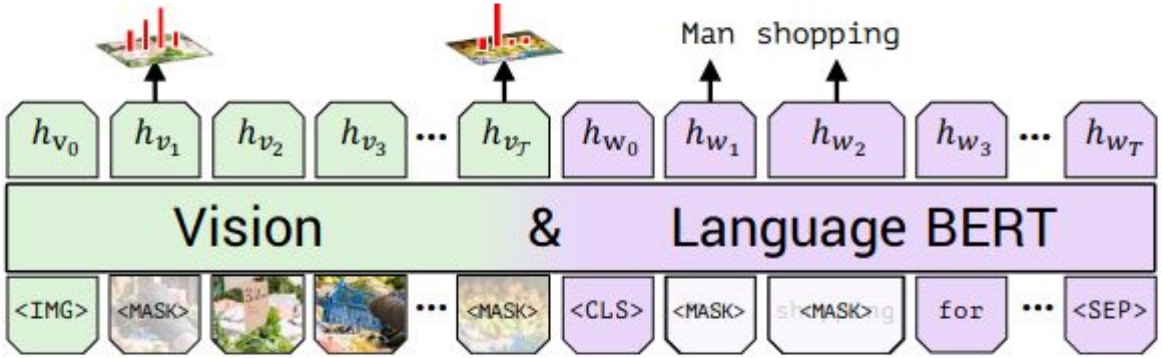

Learning language representation models from corpora





Extra-linguistic training signals



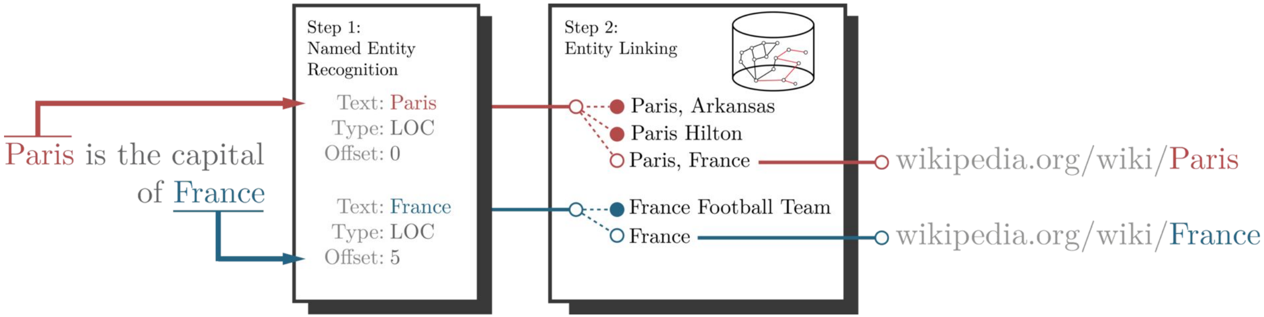


Interpreting representations; making representations interpretable

- *What information is stored in this vector?*
- *What parts of the model deal with coreference?*
- *Is it theoretically possible for model X to carry out task Y?*
- *Can we make new representations where it is easier to understand what is going on?*

Applications in industrial NLP (with Recorded Future)

University of Herefordshire's entire IT system offline after a cyber attack.



To appear at ACL 2020

Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data

Emily M. Bender
University of Washington
Department of Linguistics
ebender@uw.edu

Alexander Koller
Saarland University
Dept. of Language Science and Technology
koller@coli.uni-saarland.de

Abstract

The success of the large neural language models on many NLP tasks is exciting. However, we find that these successes sometimes lead to hype in which these models are being described as "understanding" language or capturing "meaning". In this position paper, we argue that a system trained only on form has *a priori* no way to learn meaning. In keeping with the ACL 2020 theme of "Taking Stock of Where We've Been and Where We're Going", we argue that a clear understanding of the distinction between form and meaning will help guide the field towards better science around natural language understanding.

1 Introduction

The current state of affairs in NLP is that the large neural language models (LMs), such as BERT (Devlin et al., 2019) or GPT-2 (Radford et al., 2019), are making great progress on a wide range of tasks, including those that are ostensibly meaning-sensitive. This has led to claims, in both academic and popular publications, that such models "understand" or "comprehend" natural language or learn its "meaning". From our perspective, these are overclaims caused by a misunderstanding of the relationship between linguistic form and meaning.

We argue that *the language modeling task, because it only uses form as training data, cannot in principle lead to learning of meaning*. We take the term *language model* to refer to any system trained only on the task of string prediction, whether it ... sentences, and

the structure and use of language and the ability to ground it in the world. While large neural LMs may well end up being important components of an eventual full-scale solution to human-analogous NLU, they are not nearly-there solutions to this grand challenge. We argue in this paper that genuine progress in our field — climbing the right hill, not just the hill on whose slope we currently sit — depends on maintaining clarity around big picture notions such as *meaning* and *understanding* in task design and reporting of experimental results.

After briefly reviewing the ways in which large LMs are spoken about and summarizing the recent flowering of 'BERTology' papers (§2), we offer a working definition for "meaning" (§3) and a series of thought experiments illustrating the impossibility of learning meaning when it is not in the training signal (§4,5). We then consider the human language acquisition literature for insight into what information humans use to bootstrap language learning (§6) and the distributional semantics literature to discuss what is required to ground distributional models (§7). §8 presents reflections on how we look at progress and direct research effort in our field, and in §9, we address possible counterarguments to our main thesis.

2 Large LMs: Hype and analysis

Publications talking about the application of large LMs to meaning-sensitive tasks tend to describe the models with terminology that, if interpreted at face value, is misleading. Here is a selection from academically-oriented pieces (emphasis added):

... that **understands** sentence

Experience Grounds Language

Yonatan Bisk* Ari Holtzman* Jesse Thomason*
Jacob Andreas Yoshua Bengio Joyce Chai Mirella Lapata
Angeliki Lazaridou Jonathan May Aleksandr Nisnevich Nicolas Pinto Joseph Turian

Abstract

Language understanding research is held back by a failure to relate language to the physical world it describes and to the social interactions it facilitates. Despite the incredible effectiveness of language processing models to tackle tasks after being trained on text alone, successful linguistic *communication* relies on a shared experience of the world. It is this shared experience that makes utterances meaningful.

Natural language processing is a diverse field, and progress throughout its development has come from new representational theories, modeling techniques, data collection paradigms, and tasks. We posit that the present success of representation learning approaches trained on large, text-only corpora requires the parallel tradition of research on the broader physical and social context of language to address the deeper questions of communication.

...provements in hardware and data collection galvanized progress in NLP across many ...ark tasks. Impressive performance has been ...d in language modeling (Radford et al., ...llers et al., 2019b; Keskar et al., 2019) and ...ction question answering (Devlin et al., ... et al., 2019b; Lan et al., 2020) through ...ta and massive models. With models ...uman performance on such tasks, now ...t time to reflect on a key question:

Where is NLP going?

...e consider how the data and world ...er is exposed to define and con... ...f that learner's semantics. Mean... ...from the statistical distribution ...heir use by people to comm... ...umptio...

> Meaning is not a unique property of language, but a general characteristic of human activity ... We cannot say that each morpheme or word has a single or central meaning, or even that it has a continuous or coherent range of meanings ... there are two separate uses and meanings of language – the concrete ... and the abstract.
>
> *Zellig S. Harris (Distributional Structure 1954)*

trained solely on text corpora, even when those corpora are meticulously annotated or Internet-scale.

You can't learn language from the radio. Nearly every NLP course will at some point make this claim. The futility of learning language from linguistic signal alone is intuitive, and mirrors the belief that humans lean deeply on non-linguistic knowledge (Chomsky, 1965, 1980). However, as a field we attempt this futility: trying to learn language from the *Internet*, which stands in as the modern radio to deliver limitless language. In this piece, we argue that the need for language to attach to "extralinguistic events" (Ervin-Tripp, 1973) and the requirement for social context (Baldwin et al., 1996) should guide our research.

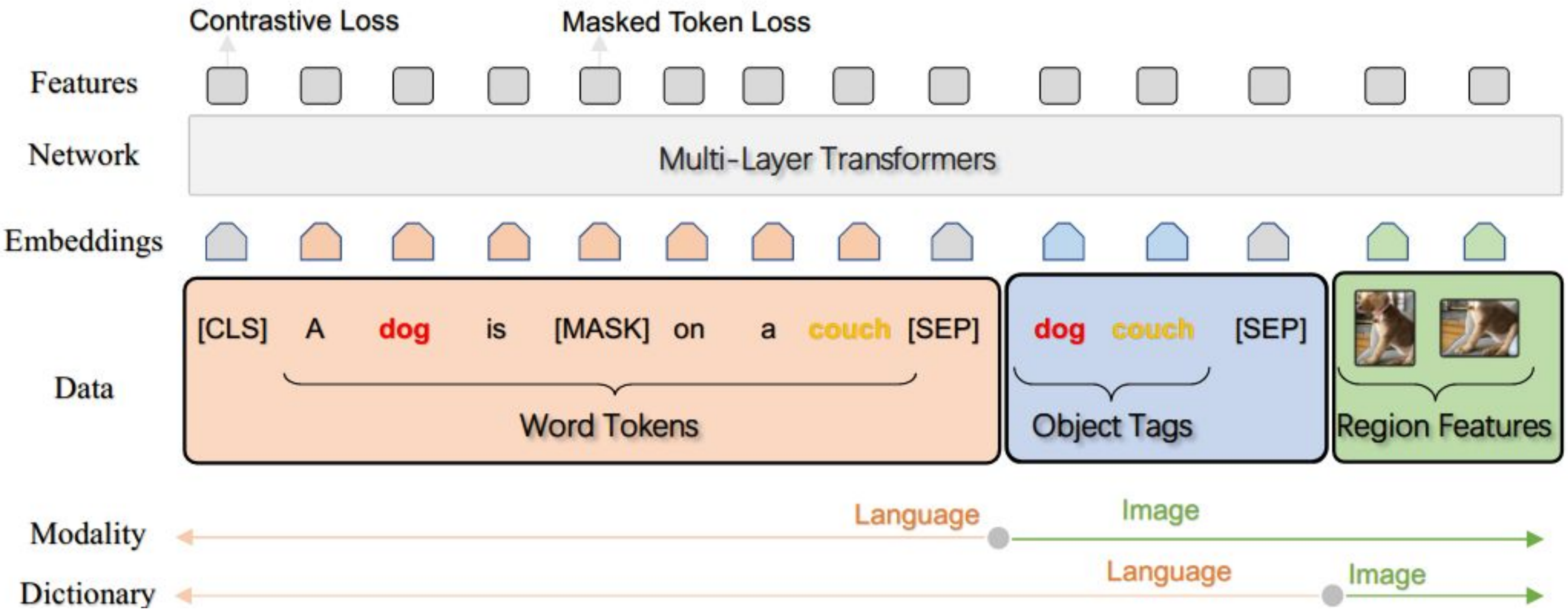
Drawing inspiration from previous work in NLP, Cognitive Science, and Linguistics, we propose the notion of a World Scope (WS) as a lens through which to audit progress in NLP. We describe five WSs, and note that most trending work in NLP operates in the second (Internet-scale data).

We define five levels of **World Scope**:

WS1. Corpus *(our past)*
WS2. Internet *(most of current NLP)*
WS3. Perception *(multimodal NLP)*
WS4. Embodiment
WS5. Social

These World Scope... the cont...

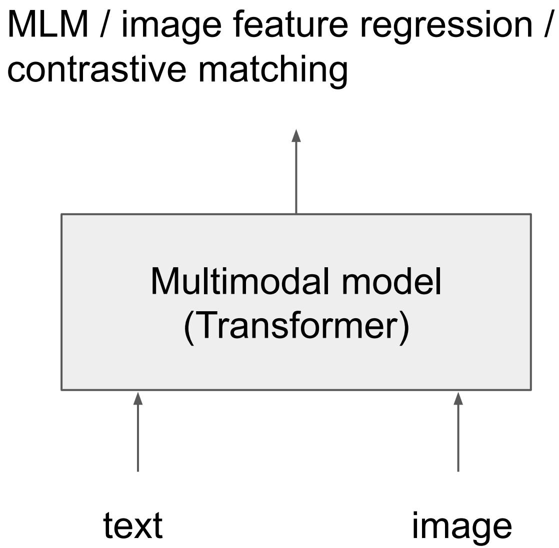
Multimodal language models



Li et al. (2020), *Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks*

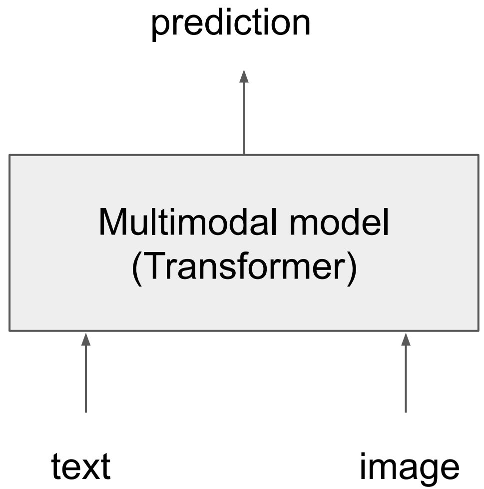
Visual-Linguistic Pretraining

- LXMERT
- VilBERT
- ImageBERT
- VisualBERT
- OSCAR
- 12-in-1
- VinVL
- Ernie-VIL
- ...



Visual-Linguistic tasks / benchmarks

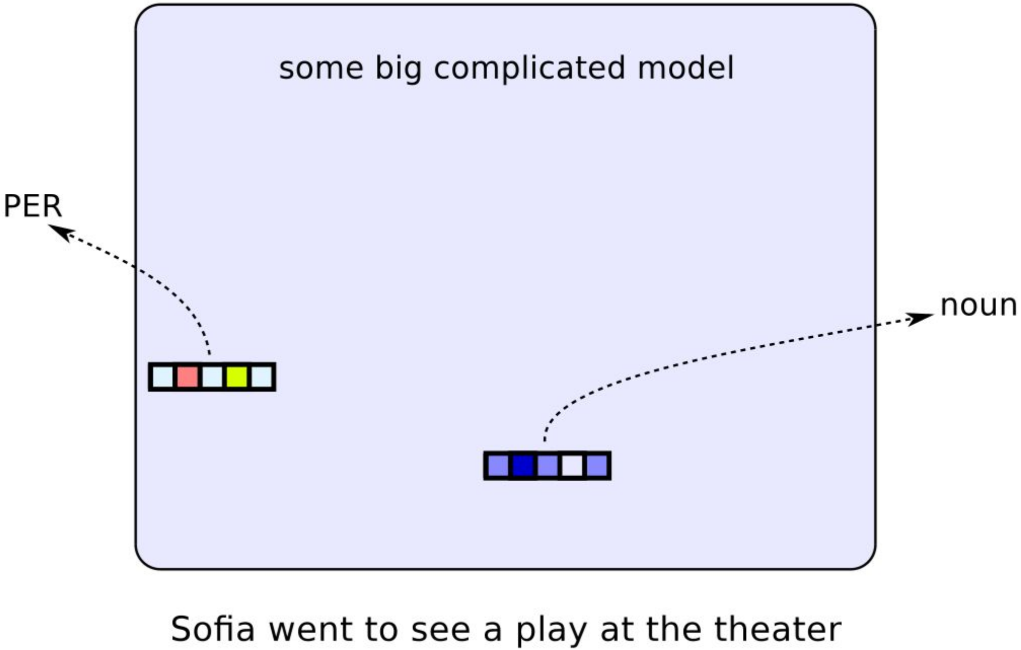
- Text-Image matching
  - Image to text retrieval/classification
  - Text to image retrieval/classification
- Text-Image generation
  - Image to text generation: Image captioning
  - Text to image generation (e.g. DALL-E)
- Text-Image classification
  - Visual Question Answering (VQA / GQA benchmarks)
  - Visual Commonsense Reasoning (VCR)
  - Natural Language for Visual Reasoning (NLVR)



Are text representations affected by multimodal training?

- Do text representations “store” some visual information?
- Do NLP applications work better when representations are trained multimodally?
  - ... at least in some narrow cases?
  - maybe primarily when the text discusses visual properties?

Investigating text representation models



Querying language models

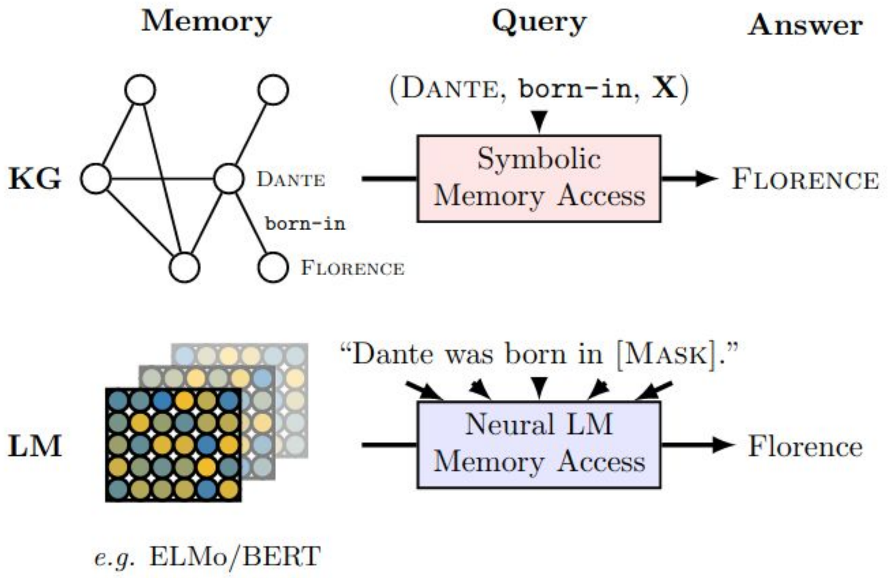


Figure 1: Querying knowledge bases (KB) and language models (LM) for factual knowledge.

Language Models as Knowledge Bases (Petroni et al. 2019)

Querying language models for prototypical colors

green

red

yellow

The color of grass is [MASK].

polar bear

strawberry

lemon

...

cf. also the idea of “memory colors” in vision and cogsci research

Initial findings

Lovisa Hagström, Tobias Norlund & Richard Johansson

Main idea

Do NLP applications work better with multimodal training?

Main idea

Do NLP applications work better with multimodal training?

- For example, can a multimodal text+image model develop a better understanding of colors than a unimodal text model?

Main idea

Do NLP applications work better with multimodal training?

- For example, can a multimodal text+image model develop a better understanding of colors than a unimodal text model?
- Could the multimodal model benefit from this understanding also on a pure text task?

A task and a dataset for evaluating color understanding

- The simplest evaluation task we could think of for evaluating how well grounded a model is in visual contexts without explicit use of images.

A task and a dataset for evaluating color understanding

- The simplest evaluation task we could think of for evaluating how well grounded a model is in visual contexts without explicit use of images.
- We query the models about typical colors of objects (memory colors) to investigate whether the models have knowledge of the meaning of different colors

A task and a dataset for evaluating color understanding

- The simplest evaluation task we could think of for evaluating how well grounded a model is in visual contexts without explicit use of images.
- We query the models about typical colors of objects (memory colors) to investigate whether the models have knowledge of the meaning of different colors
  - *Grass - Green*

A task and a dataset for evaluating color understanding

- The simplest evaluation task we could think of for evaluating how well grounded a model is in visual contexts without explicit use of images.
- We query the models about typical colors of objects (memory colors) to investigate whether the models have knowledge of the meaning of different colors
  - *Grass - Green*
  - *Lemon - Yellow*

A task and a dataset for evaluating color understanding

- The simplest evaluation task we could think of for evaluating how well grounded a model is in visual contexts without explicit use of images.
- We query the models about typical colors of objects (memory colors) to investigate whether the models have knowledge of the meaning of different colors
  - *Grass - Green*
  - *Lemon - Yellow*
  - *Coal - Black*

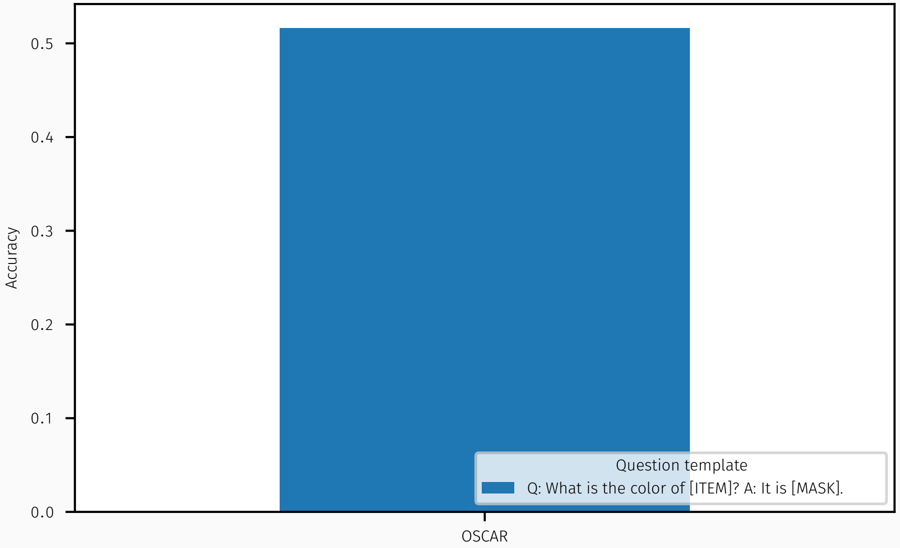
A task and a dataset for evaluating color understanding

- The simplest evaluation task we could think of for evaluating how well grounded a model is in visual contexts without explicit use of images.
- We query the models about typical colors of objects (memory colors) to investigate whether the models have knowledge of the meaning of different colors
  - *Grass - Green*
  - *Lemon - Yellow*
  - *Coal - Black*
- 124 item color pairs in total

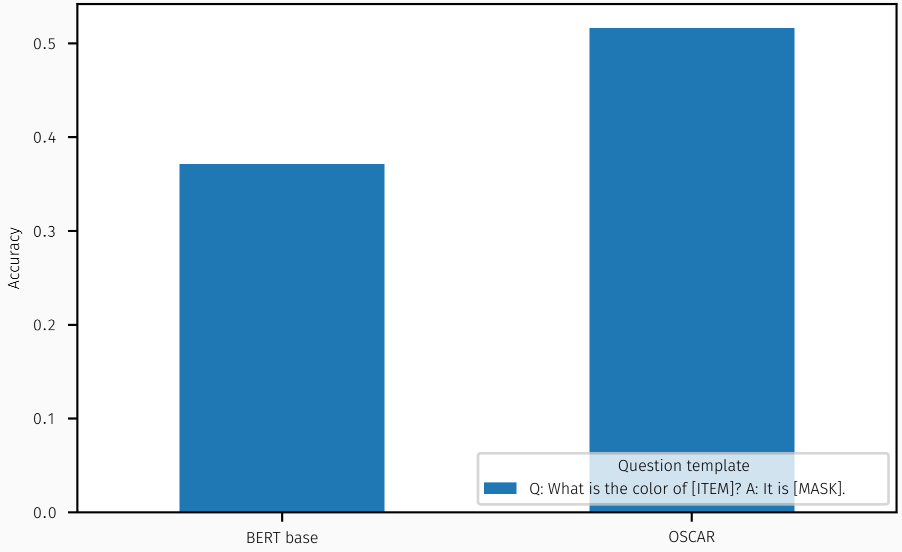
A task and a dataset for evaluating color understanding

- The simplest evaluation task we could think of for evaluating how well grounded a model is in visual contexts without explicit use of images.
- We query the models about typical colors of objects (memory colors) to investigate whether the models have knowledge of the meaning of different colors
  - *Grass - Green*
  - *Lemon - Yellow*
  - *Coal - Black*
- 124 item color pairs in total
- Includes 10 colors (*yellow*, *blue*, *green*, *white*, *red*, *orange*, *black*, *pink*, *brown*, *grey*)

Model performances on the item-color dataset



Model performances on the item-color dataset



Model performances on the item-color dataset

The multimodal model OSCAR has a better performance on our item-color evaluation set than the unimodal BERT base model.

Model performances on the item-color dataset

The multimodal model OSCAR has a better performance on our item-color evaluation set than the unimodal BERT base model.

But is this due to OSCAR being more grounded than BERT?

Model performances on the item-color dataset

The multimodal model OSCAR has a better performance on our item-color evaluation set than the unimodal BERT base model.

But is this due to OSCAR being more grounded than BERT?

Can we rule out that the difference in performance is due to something other than grounding?

Model performances on the item-color dataset

Can we rule out that the difference in performance is due to something other than grounding?

Model performances on the item-color dataset

Can we rule out that the difference in performance is due to something other than grounding?

For example, the models have been trained on different datasets

Model performances on the item-color dataset

Can we rule out that the difference in performance is due to something other than grounding?

For example, the models have been trained on different datasets

- BERT: English Wikipedia + BookCorpus

Model performances on the item-color dataset

Can we rule out that the difference in performance is due to something other than grounding?

For example, the models have been trained on different datasets

- BERT: English Wikipedia + BookCorpus
- OSCAR: same data as for BERT + multimodal data (MS COCO, VQA, ...)

Model performances on the item-color dataset

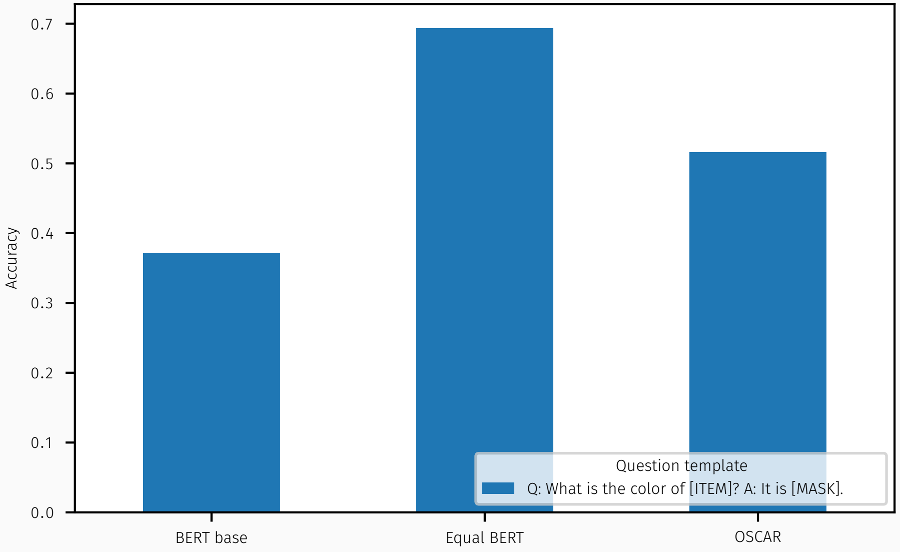
Can we rule out that the difference in performance is due to something other than grounding?

For example, the models have been trained on different datasets

- BERT: English Wikipedia + BookCorpus
- OSCAR: same data as for BERT + multimodal data (MS COCO, VQA, ...)

What if we make sure that the unimodal BERT model has been trained on the same textual data as OSCAR and then evaluate?

Model performances with equal footing



Model performances with equal footing

Can we rule out that the difference in performance is due to something other than grounding or training on different datasets?

Model performances with equal footing

Can we rule out that the difference in performance is due to something other than grounding or training on different datasets?

The models may also have varying sensitivity to the prompt they are evaluated with.

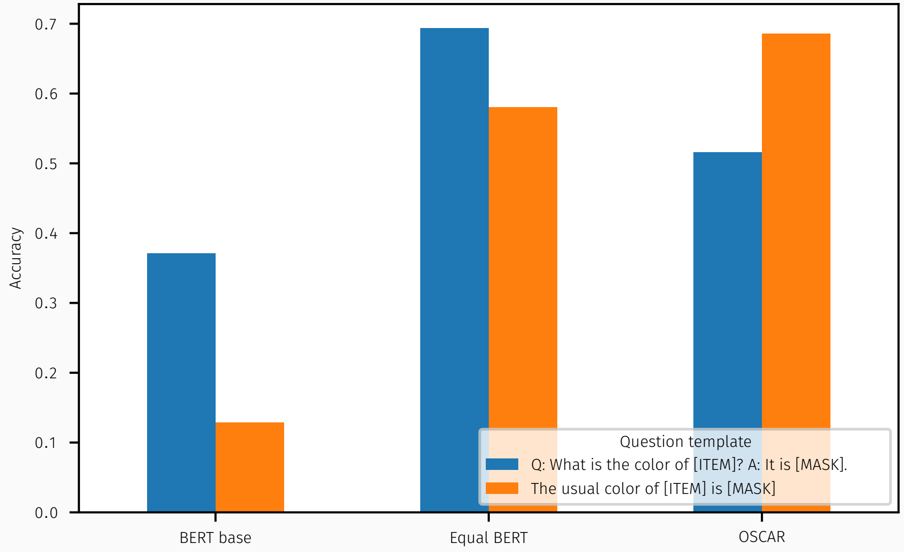
Model performances with equal footing

Can we rule out that the difference in performance is due to something other than grounding or training on different datasets?

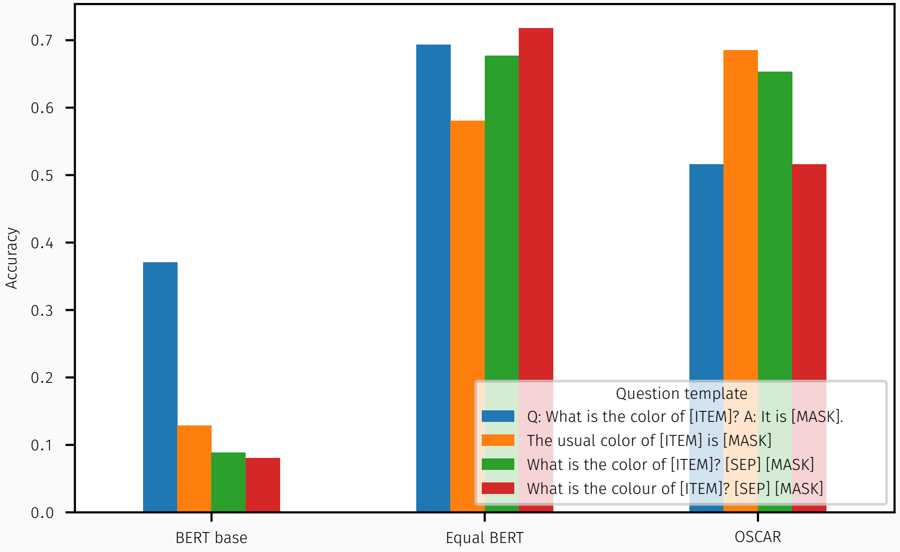
The models may also have varying sensitivity to the prompt they are evaluated with.

- Prompt engineering

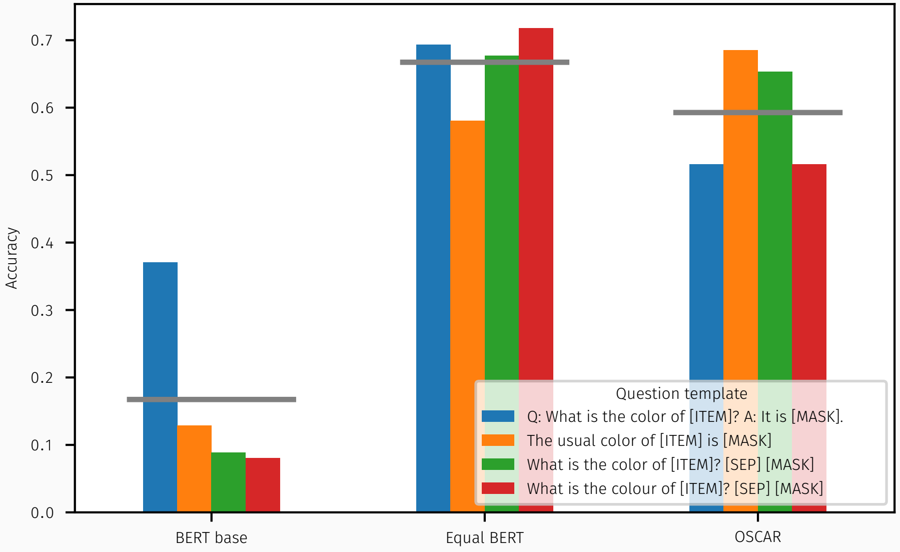
Model performances with equal footing and different prompts



Model performances with equal footing and different prompts



Model performances with equal footing and different prompts



Model performances with equal footing and different prompts

Can we rule out that the difference in performance is due to something other than grounding, training on different datasets or prompt sensitivity?

Model performances with equal footing and different prompts

Can we rule out that the difference in performance is due to something other than grounding, training on different datasets or prompt sensitivity?

Could it be due to the specific model used?

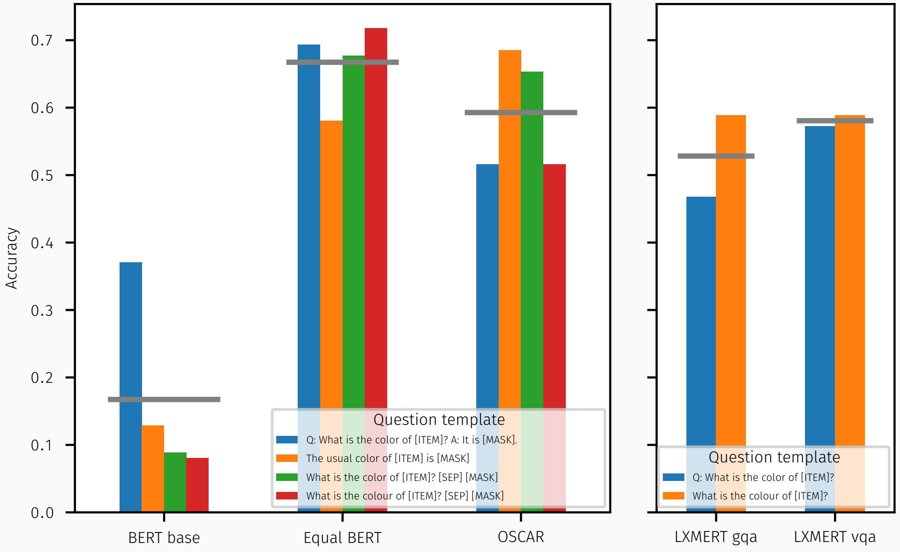
Model performances with equal footing and different prompts

Can we rule out that the difference in performance is due to something other than grounding, training on different datasets or prompt sensitivity?

Could it be due to the specific model used?

There are other multimodal models than OSCAR, for example LXMERT.

Model performances with another multimodal model

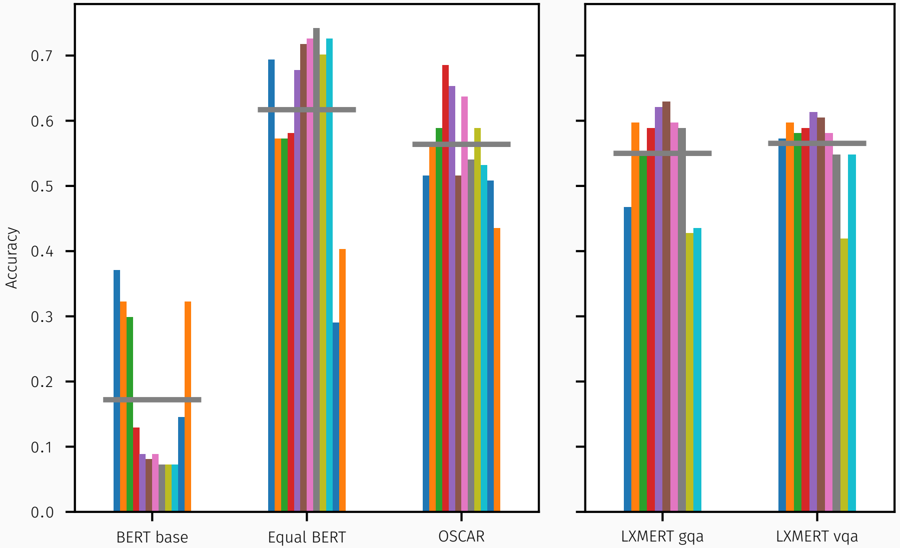


Many model performances on the item-color dataset

- Q: What is the color of [ITEM]? A: It is [MASK].
- Q: What is the color of [ITEM]? [SEP] A: It is [MASK].
- The color of [ITEM] is [MASK].
- The usual color of [ITEM] is [MASK]
- What is the color of [ITEM]? [SEP] [MASK]
- What is the colour of [ITEM]? [SEP] [MASK]
- What is the typical color of [ITEM]? [SEP] [MASK]
- What is the typical colour of [ITEM]? [SEP] [MASK]
- What is the usual color of [ITEM]? [SEP] [MASK]
- What is the usual colour of [ITEM]? [SEP] [MASK]
- [ITEM] usually has the color [SEP] [MASK]
- [ITEM] usually has the color of [MASK].

- Q: What is the color of [ITEM]?
- The color of [ITEM] is what?
- The usual color of [ITEM] is what?
- What is the colour of [ITEM]?
- What is the typical color of [ITEM]?
- What is the typical colour of [ITEM]?
- What is the usual color of [ITEM]?
- What is the usual colour of [ITEM]?
- [ITEM] usually has the color of what?
- [ITEM] usually has what color?

Many model performances on the item-color dataset



Many model performances on the item-color dataset

Conclusions

Many model performances on the item-color dataset

Conclusions

- Performances of all models are highly dependent on the chosen question template.

Many model performances on the item-color dataset

Conclusions

- Performances of all models are highly dependent on the chosen question template.
- The unimodal BERT model performs better on our evaluation set than the multimodal models.

Many model performances on the item-color dataset

Conclusions

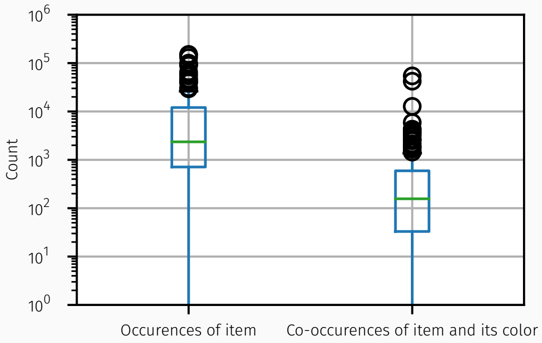
- Performances of all models are highly dependent on the chosen question template.
- The unimodal BERT model performs better on our evaluation set than the multimodal models.
- Could something be wrong with our evaluation task?

Our evaluation task does not work as intended

The information we are looking for can be found in the text data.

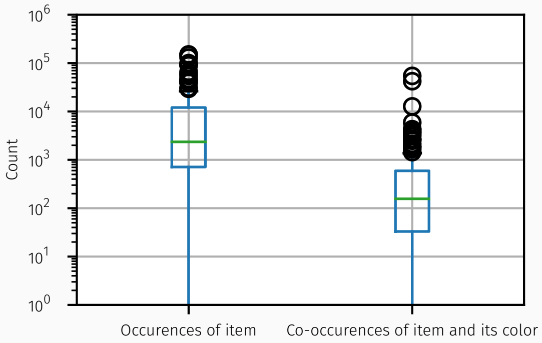
Our evaluation task does not work as intended

The information we are looking for can be found in the text data.



Our evaluation task does not work as intended

The information we are looking for can be found in the text data.



While we would want it to be revealed only by the visual input.

To conclude

Work-in-progress and future work

Work-in-progress and future work

- Remove the the parts of the pre-training dataset that reveal the evaluation task, then re-train and re-evaluate.

Work-in-progress and future work

- Remove the the parts of the pre-training dataset that reveal the evaluation task, then re-train and re-evaluate.
- Develop a model that can self-visualize.

Work-in-progress and future work

- Remove the the parts of the pre-training dataset that reveal the evaluation task, then re-train and re-evaluate.
- Develop a model that can self-visualize.
- Further evaluate the multimodal models on pure text tasks.

Questions to discuss

- How can we check if a model is grounded or not without explicit use of images or other multimodal data sources?

Questions to discuss

- How can we check if a model is grounded or not without explicit use of images or other multimodal data sources?
  - Is this question relevant?

Questions to discuss

- How can we check if a model is grounded or not without explicit use of images or other multimodal data sources?
  - Is this question relevant?
  - How do we build tools or sets for evaluating grounded models, without the risk of the model “cheating”?

Questions to discuss

- How can we check if a model is grounded or not without explicit use of images or other multimodal data sources?
  - Is this question relevant?
  - How do we build tools or sets for evaluating grounded models, without the risk of the model "cheating"?
  - Would we need to know exactly what is in the training data of the model that is being evaluated to rule out cheating?

Questions to discuss

- How can we check if a model is grounded or not without explicit use of images or other multimodal data sources?
  - Is this question relevant?
  - How do we build tools or sets for evaluating grounded models, without the risk of the model “cheating”?
  - Would we need to know exactly what is in the training data of the model that is being evaluated to rule out cheating?
  - Would the removal of “revealing” content in the training data be a way to avoid the risk of the model cheating?

Questions to discuss

- How can we check if a model is grounded or not without explicit use of images or other multimodal data sources?
    - Is this question relevant?
    - How do we build tools or sets for evaluating grounded models, without the risk of the model “cheating”?
    - Would we need to know exactly what is in the training data of the model that is being evaluated to rule out cheating?
    - Would the removal of “revealing” content in the training data be a way to avoid the risk of the model cheating?
    - How can we make sure that subsequent evaluation results are robust and significant?

Questions to discuss

- How can we check if a model is grounded or not without explicit use of images or other multimodal data sources?
  - Is this question relevant?
  - How do we build tools or sets for evaluating grounded models, without the risk of the model "cheating"?
  - Would we need to know exactly what is in the training data of the model that is being evaluated to rule out cheating?
  - Would the removal of "revealing" content in the training data be a way to avoid the risk of the model cheating?
  - How can we make sure that subsequent evaluation results are robust and significant?
- What tasks do we want to solve better with a grounded model?

Thank you for listening!