Be Different to Be Better!
A Benchmark to Leverage the Complementarity of Language and Vision

Sandro Pezzelle[1], Claudio Greco[2], Greta Gandolfi[2], Eleonora Gualdoni[2], Raffaella Bernardi[2,3]
[1]Institute for Logic, Language, and Computation, University of Amsterdam
[2]CIMeC, [3]DISI, University of Trento
sites.google.com/view/bd2bb

SCAN ME

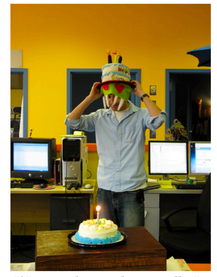
OVERVIEW

- Human communication, in real-life situations, is **multimodal** [1]
- Speakers do **not need to "repeat"** information provided by the environment
- In current computational approaches to language and vision the two modalities are ***aligned*** rather than ***complementary***
- **bd2bb**: Novel task where either modality is **necessary but not sufficient**

We introduce bd2bb, a novel benchmark that requires models combine *complementary* information from language and vision: +10K crowdsourced datapoints

While solving bd2bb is relatively easy for humans (~80% accuracy), SotA pre-trained multimodal encoders struggle to achieve similar results (~60% accuracy)

EXAMPLES



If I am in the mood to act silly, I will…

L: attend a dinner like this man holding a gift
L: buy him a cake and invite his friends to party
T: **act silly with this man and eat cake**
V: help my child cut their cake
V: have cake with soldiers

If I don't like this, I will…

sit next to the woman on the bench
get my face painted
avert my eyes from the man who looks silly
teach him how to tie a tie
wear a costume and march in a parade

DATASET

- Given an **image** depicting a real-life situation, 5 annotators provide 1) an **intention**, i.e., how they might feel/ behave if they were in that image (*If I…*); 2) a **target action**, i.e., what they would do based on that feeling/behavior (*I will…*)
- Intention: **ungrounded**; Action: **grounded**
- +10K valid *<image, intention, action>* datapoints + rich linguistic annotation

MULTIPLE-CHOICE TASK

IMAGE

If I have tons of energy

INTENTION

CANDIDATE ACTIONS

I will **play** baseball with the men
I will **play** a game of **tennis** with the **man**
I will compare images of me hitting the **tennis ball**
I will **play** baseball with the women
I will applaud my favourite **tennis player** of all time

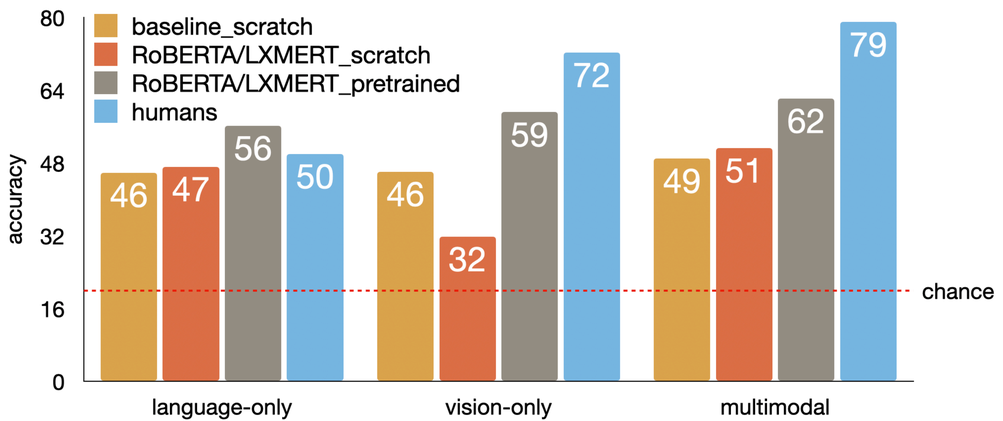
DISCUSSION

- **Multimodal integration is the key:** L&V models outperform unimodal systems
- **Best models are far from humans:** gap of ~17% accuracy (20% in hard test set)
- **Pre-trained is better:** largely pre-trained systems outperform models trained from scratch on the bd2bb task
- Importance of **richly-annotated resources** to test pre-trained systems
- Importance of multimodal tasks where **L&V are genuinely complementary**

MODELS & SETTINGS

	lang-only	vis-only	multimodal
baseline	✓	✓	✓
RoBERTa [2]	✓		
LXMERT [3]		✓	✓
humans	✓	✓	✓

RESULTS



REFERENCES

[1] Gunther R Kress. 2010. *Multimodality: A social semiotic approach to contemporary communication.* Taylor & Francis.
[2] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. *RoBERTa: A Robustly optimized BERT pretraining approach.* CoRR, abs/1907.11692.
[3] Hao Tan and Mohit Bansal. 2019. *LXMERT: Learning cross-modality encoder representations from transformers.* In Proceedings of EMNLP-IJCNLP, pages 5103–5114.

Funded by