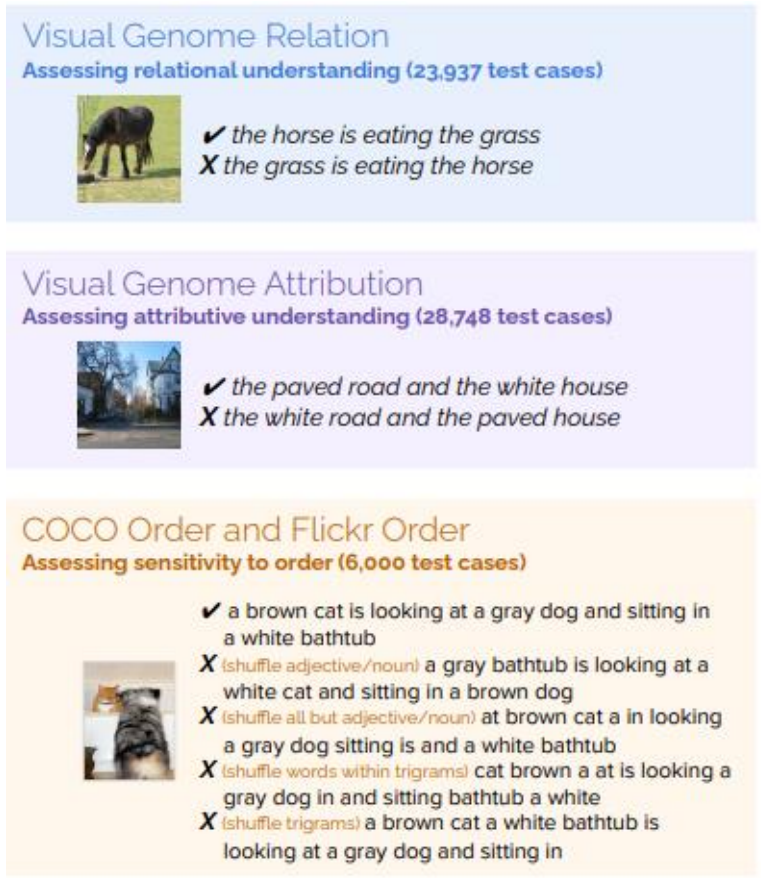
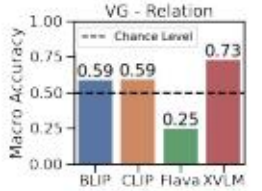
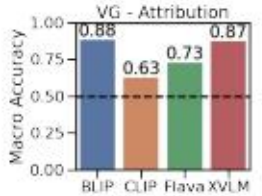
WHEN AND WHY VISION-LANGUAGE MODELS BEHAVE LIKE BAGS-OF-WORDS, AND WHAT TO DO ABOUT IT?

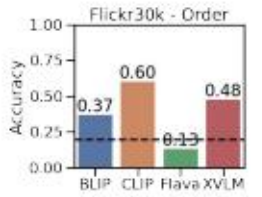
Краткое описание области исследования

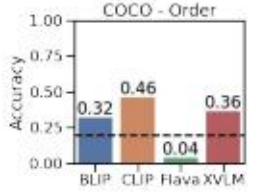
- Атрибутивное понимание
- Реляционное понимание
- Композициональная структура
- Анализ порядка
- Эвристики и обучение

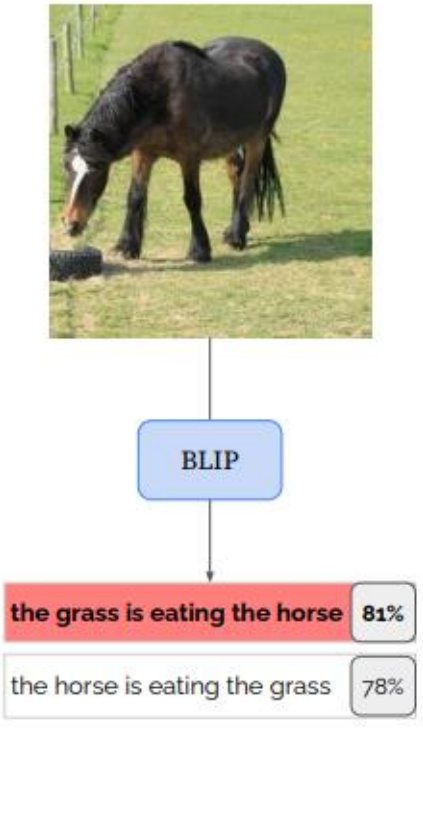




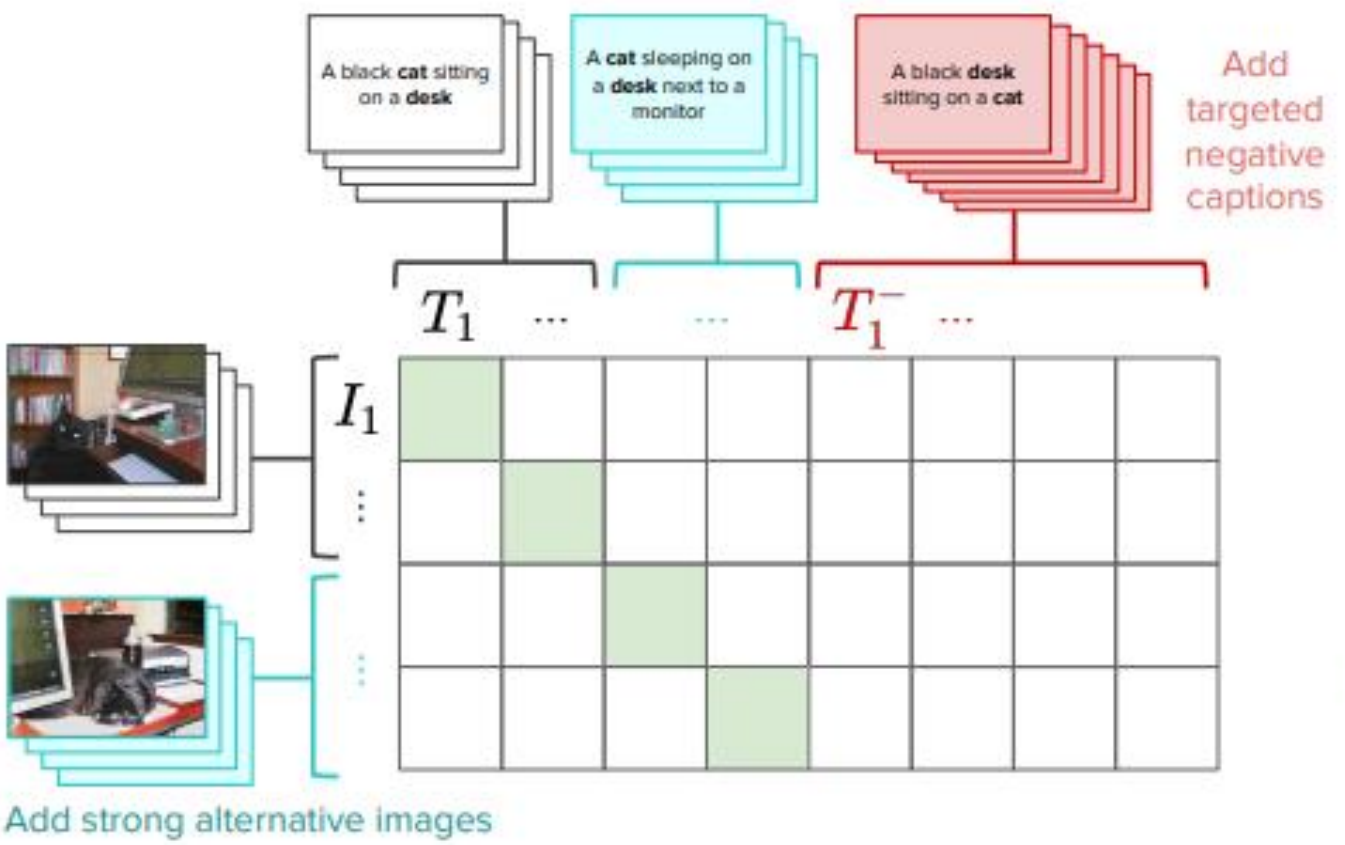


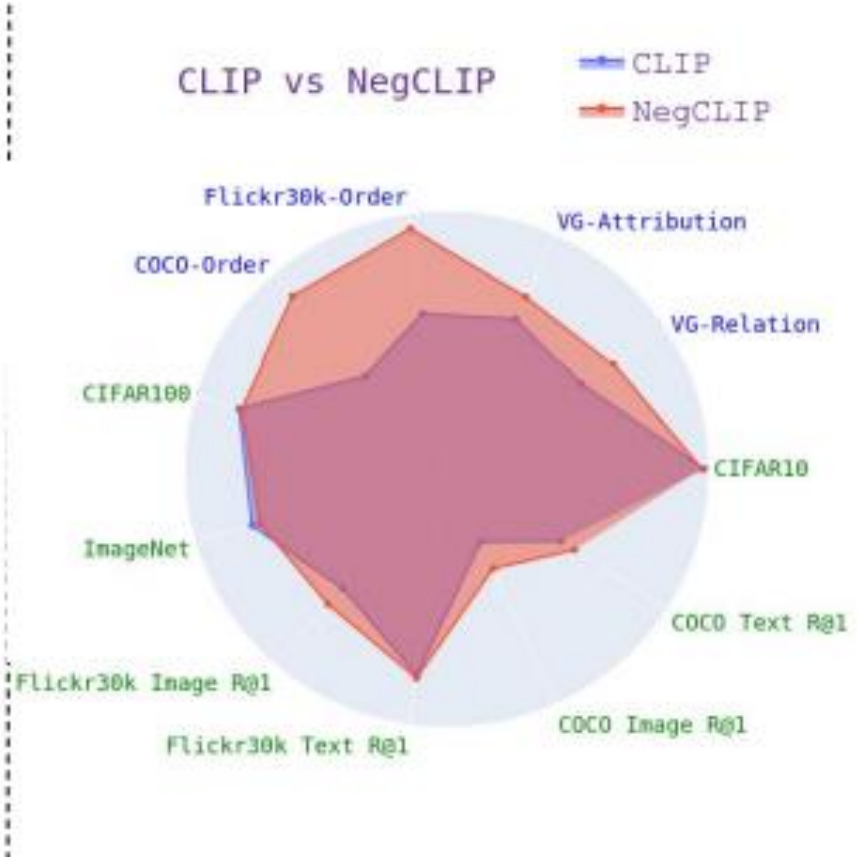






Краткий итог данной работы





Предыдущие исследования в данной области

- Исследование композициональных отношений в тексте и изображениях
- Смешанные модели текста и изображений
- Изучение проблем композициональности в моделях
- Исследование в области изображений и текста
- Данные и бенчмарки

Композициональность визуально-языковых моделей

- Stella Frank, Emanuele Bugliarello, and Desmond Elliott. Vision-and-language or vision-forlanguage? On cross-modal influence in multimodal transformers.
- Letitia Parcalabescu, Albert Gatt, Anette Frank, and Iacer Calixto. Seeing past words: Testing the cross-modal capabilities of pretrained V&L models on counting tasks.
- Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality.
- Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. Why is winoground hard? investigating failures in visuolinguistic compositionality.

Композициональность визуально-языковых моделей

- Ben Bogin, Shivanshu Gupta, Matt Gardner, and Jonathan Berant. Covr: A test-bed for visually grounded compositional generalization with real images.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database.
- Aimen Zerroug, Mohit Vaishnav, Julien Colin, Sebastian Musslick, and Thomas Serre. A benchmark for compositional visual reasoning.
- Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations.
- Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding.

Информация о порядке в language и vision

- Jack Hessel and Alexandra Schofield. How effective is bert without word ordering? implications for language understanding and data privacy.
- Joe O'Connor and Jacob Andreas. What context features can transformer language models use?
- Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. ¨ Masked language modeling and the distributional hypothesis: Order word matters pre-training for little.
- Allyson Ettinger. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models.
- Wieland Brendel and Matthias Bethge. Approximating CNNs with bag-of-local-features models works surprisingly well on imagenet.
- Ajinkya Tejankar, Bichen Wu, Saining Xie, Madian Khabsa, Hamed Pirsiavash, and Hamed Firooz. A fistful of words: Learning transferable visual models from bag-of-words supervision.

Использование негативов и контрастного обучения

- Ben Harwood, Vijay Kumar BG, Gustavo Carneiro, Ian Reid, and Tom Drummond. Smart mining for deep metric learning.
- Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning.
- Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning.
- Yao Qin, Chiyuan Zhang, Ting Chen, Balaji Lakshminarayanan, Alex Beutel, and Xuezhi Wang. Understanding and improving robustness of vision transformers through patch-based negative augmentation.
- Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation.

Перспективы будущих исследований

- Улучшение алгоритмов обучения
- Создание более сложных бенчмарков
- Применение в практических задачах