Acquiring Linguistic Knowledge from Multimodal Input

Theodor Amariucai (21-943-337) - *MSc Computer Science*

Supervision

- Alexander Warstadt (alexanderscott.warstadt@inf.ethz.ch)
- Ryan Cotterell (ryan.cotterell@inf.ethz.ch)

Introduction

Natural language processing (NLP) is a field of machine learning that has seen tremendous progress in recent years. Pretraining has become a popular approach in NLP, where models are trained on large amounts of unsupervised data to learn general linguistic knowledge and improve their performance on downstream tasks. However, the current trend of pretraining language models (LMs) largely relies on text [1, 2, 3] (and to a lesser extent, vision [4, 5, 6]) input, overlooking other influential sensory data (e.g., speech) that is part of the learning signal for humans acquiring language.

Multimodal language models have been gaining attention due to their ability to reason about information transcending text. By incorporating multiple modalities, they can better capture the complex relationships between words and the visual or auditory world, obtaining a more comprehensive and accurate understanding of concepts.

Motivation

Existing Transformer-based [7], pre-trained vision and language models have achieved impressive performance at vision tasks and cross- and multimodal vision and language tasks [4]. However, even the most recent such models appear unlikely to improve natural language understanding (NLU) over their unimodal (text-only) counterparts [8, 9]. Moreover, large models are, in general, significantly less data-efficient than humans at tasks such as grammar learning, requiring hundreds of times more exposure to words than a child [9].

Human language acquisition goes beyond simple image-text pairs and is driven by additional stimuli (e.g., hearing). However, speech-text multimodal architectures [10] are an under-explored area in language acquisition research, with the incentives typically being automated speech recognition or translation

[11, 12, 13]. This could be a missed opportunity, as information embedded in speech wavelengths might more effectively reinforce grammaticality in models by more closely mapping to text (as opposed to visual environments).

We hypothesize that learning from multimodal inputs could improve language acquisition and narrow the gap between how humans and machines learn. This hypothesis has only been tested in limited pretraining paradigms [8] or incidental ablation studies with restricted evaluation and analysis [4].

Lastly, we seek to understand the impact on linguistic competence of varying quantities of text and multimodal training data and whether efficient trade-offs could be made in the input space.

Target of the Thesis

The goal is to investigate a state-of-the-art multimodal language model and enrich it with audio capabilities. The final deliverable is a controlled study backed by a novel, open source, speech-text model whose NLU [14, 15] and pseudo-log-likelihoods [16] are scrutinized under different sensory groundings.

By accessing new insights into the fundamental challenges of language modeling, we hope to foster innovation among NLP researchers, AI practitioners, computational linguists, and cognitive scientists.

Objectives

1. **Literature review** on the state-of-the-art in multimodal language model pretraining, including common objectives, benchmarks, and corpora.

2. Code **infrastructure setup** and controlled **experimentation** (minimizing differences in model sizes and training data):

   (a) Can trade-offs be made in the vision-text input space while preserving performance? Ablation study [1] on how FLAVA's [2] linguistic knowledge is affected by varying volumes of text and visual input.

3. **Development**

   (a) Consider SLAM [10], a closed source, early-fusion, speech-text model where NLU performance was shown to decrease after adding speech input. We respond to the authors' call for "better cross-modal alignments to alleviate the presumed capacity limitations" by enriching FLAVA with audio processing capabilities. Thus, we develop the first [3] open source, early-fusion, speech-text model.

[1] An option would be to 1) set text-only baselines, 2) independently vary text and non-text training data, 3) evaluate on text data.

[2] FLAVA is a recent breakthrough [8, 9] vision-text model that was shown to improve NLU.

[3] As far as my current literature review suggests.

(b) Can trade-offs be made in the *speech*-text input space while preserving performance? Ablation study on how the new model's linguistic knowledge is affected by varying volumes of text and speech input.

4. **Analysis and thesis writing**

   (a) Points of interest will be grammaticality (CoLA [15], BLiMP [14]) and data efficiency (# words the model has been exposed to).

Deliverables

The project will result in the following concrete deliverables:

- Project description
  - Introduction / motivation of the problem
  - Thorough analysis of related work
  - Description of the resulting implementation including its design
  - Experimental evaluation
- Complete source code for the resulting prototype, benchmarking, data analysis, and scripts [4].
- Presentation of the results and demonstration of functionality.

Grading

The Master's Thesis (MT) is a graded semester performance. In order to successfully complete the MT, a grade of 4.0 or higher must be obtained. The supervisor establishes the assessment criteria in a written report, which can include a presentation. In principle, the following evaluation scale is applied:

Grade	Requirements
6.00	Work and results are publishable for international workshops
5.50	Thesis quality significantly exceeds expectations
5.00	Thesis meets expectations
4.50	Thesis partially meets expectations and has minor deficits
4.00	Thesis meets minimum quality requirements; but has major deficits and is clearly below expectations

[4] This is accompanied with enough documentation to allow complete reproduction of the experimental results.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” 2018.

[2] P. He, X. Liu, J. Gao, and W. Chen, “Deberta: Decoding-enhanced bert with disentangled attention,” 2020. [Online]. Available: https://arxiv.org/abs/2006.03654

[3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

[4] A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, and D. Kiela, “Flava: A foundational language and vision alignment model,” 2021.

[5] E. Bugliarello, R. Cotterell, N. Okazaki, and D. Elliott, “Multimodal pre-training unmasked: A meta-analysis and a unified framework of vision-and-language berts,” 2020.

[6] J. Lu, D. Batra, D. Parikh, and S. Lee, “Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks,” 2019. [Online]. Available: https://arxiv.org/abs/1908.02265

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2017.

[8] T. Yun, C. Sun, and E. Pavlick, “Does vision-and-language pretraining improve lexical grounding?” 2021. [Online]. Available: https://arxiv.org/abs/2109.10246

[9] A. Warstadt and S. R. Bowman, “What artificial neural networks can tell us about human language acquisition,” 2022. [Online]. Available: https://arxiv.org/abs/2208.07998

[10] A. Bapna, Y.-a. Chung, N. Wu, A. Gulati, Y. Jia, J. H. Clark, M. Johnson, J. Riesa, A. Conneau, and Y. Zhang, “Slam: A unified encoder for speech and language modeling via speech-text joint pre-training,” 2021. [Online]. Available: https://arxiv.org/abs/2110.10329

[11] J. Ao, R. Wang, L. Zhou, C. Wang, S. Ren, Y. Wu, S. Liu, T. Ko, Q. Li, Y. Zhang, Z. Wei, Y. Qian, J. Li, and F. Wei, “Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing,” 2021. [Online]. Available: https://arxiv.org/abs/2110.07205

[12] Z. Zhang, L. Zhou, J. Ao, S. Liu, L. Dai, J. Li, and F. Wei, “Speechut: Bridging speech and text with hidden-unit for encoder-decoder based speech-text pre-training,” 2022. [Online]. Available: https://arxiv.org/abs/2210.03730

[13] Y. Tang, H. Gong, N. Dong, C. Wang, W.-N. Hsu, J. Gu, A. Baevski, X. Li, A. Mohamed, M. Auli, and J. Pino, "Unified speech-text pre-training for speech translation and recognition," 2022. [Online]. Available: https://arxiv.org/abs/2204.05409

[14] A. Warstadt, A. Parrish, H. Liu, A. Mohananey, W. Peng, S.-F. Wang, and S. R. Bowman, "Blimp: The benchmark of linguistic minimal pairs for english," 2019. [Online]. Available: https://arxiv.org/abs/1912.00582

[15] A. Warstadt, A. Singh, and S. R. Bowman, "Neural network acceptability judgments," *arXiv preprint arXiv:1805.12471*, 2018.

[16] J. Salazar, D. Liang, T. Q. Nguyen, and K. Kirchhoff, "Masked language model scoring," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020. [Online]. Available: https://doi.org/10.18653%2Fv1%2F2020.acl-main.240