

Boise State University

ScholarWorks

Computer Science Faculty Publications and
Presentations

Department of Computer Science

2021

Enriching Language Models with Visually-Grounded Word Vectors and the Lancaster Sensorimotor Norms

Casey Kennington
Boise State University

—

Enriching Language Models with Visually-grounded Word Vectors and the Lancaster Sensorimotor Norms

Casey Kennington

Department of Computer Science

Boise State University

caseykennington@boisestate.edu

Abstract

Language models are trained only on text despite the fact that humans learn their first language in a highly interactive and multimodal environment where the first set of learned words are largely concrete, denoting physical entities and embodied states. To enrich language models with some of this missing experience, we leverage two sources of information: (1) the Lancaster Sensorimotor norms, which provide ratings (means and standard deviations) for over 40,000 English words along several dimensions of embodiment, and which capture the extent to which something is experienced across 11 different sensory modalities, and (2) vectors from coefficients of binary classifiers trained on images for the BERT vocabulary. We pre-trained the ELECTRA model and fine-tuned the RoBERTa model with these two sources of information then evaluate using the established GLUE benchmark and the Visual Dialog benchmark. We find that enriching language models with the Lancaster norms and image vectors improves results in both tasks, with some implications for robust language models that capture holistic linguistic meaning in a language learning context.

1 Introduction

Children learn their first spoken language in a highly interactive setting where generally the first words children learn are concrete words that denote physical objects, which is an important developmental step in child first language acquisition (Kuperman et al., 2012a; McCune, 2008; Clark, 2013). This is partly because handling the Symbol Grounding Problem—the ability to connect symbolic knowledge of language with representations of the physical world (Harnad, 1990)—must take place before children learn more abstract concepts later in their cognitive development (Borghi et al., 2019; Ponari et al., 2018). Importantly, the physical world is not just the visual world; children learn that

words ground into proprioceptive states (e.g., a hand grasp around an object has specific muscle activations tied to the word *grab*), interoceptive states (i.e., affect and valence), as well as all other sensory modalities (e.g., the word *stinky* grounds into olfactory, the word *loud* grounds into auditory). These claims are evidenced in a large body of child development and cognitive science literature. Smith and Gasser (2005), for example, identified that babies’ experience of the world is profoundly multimodal: babies live in a physical world full of rich regularities that organize perception, action and thought; babies learn in a social world to learn a shared linguistic communicative system that is symbolic. Furthermore, a growing body of literature from linguistics and computational linguistics makes a strong case that the process of language learning (indeed, general human cognition) is embodied, interactive, and enacted; i.e., movement in the world is required (Pulvermüller, 1999; Lakoff and Johnson, 1999; Barsalou, 2008; Johnson, 2008; Smith and Samuelson, 2009; Di Paolo et al., 2018; Bisk et al., 2020); see also the prior work in developmental robotics research; e.g., Cangelosi and Schlesinger (2015), Chapter 7.¹

Taken together, it is clear that aspects of the physical world are necessary for holistic knowledge of semantic meaning, which has implications for how language is modeled computationally. In particular, what does this mean for language models that are trained purely on text (likely largely written by adults), such as BERT (Devlin et al., 2018) or GPT-3? These models have clearly led to important advances for natural language processing tasks and applications, but it is also clear that language models trained only on text are missing critical semantic information (Bender and Koller, 2020).

¹Here we follow Dourish (2001) that embodiment is *possessing and acting through a physical manifestation in the world*; i.e., having sensory input is only part of embodiment—the ability to *act* in the world is essential.

In this paper, we contribute to a growing body of recent work that attempts to address these limitations by (1) leveraging multimodal and sensorimotor knowledge of the Lancaster Sensorimotor Norms (Lynott et al., 2019) and (2) using vectorized representations of images by treating both (1) and (2) as embeddings of language models for GLUE and Visual Dialog benchmarks. In the following section, we explain related work—a growing body of literature that is adding multimodal information to language models, then we explain our two embeddings that we will use. We explore how these embeddings can be used to enrich the ELECTRA language model’s pre-training and fine-tuning, and evaluate on the GLUE benchmark (Experiment 1, Section 4), and how they can be used to replace input embeddings for a pre-trained RoBERTa model for the Visual Dialog task (Experiment 2, Section 5). Our experiments shed light on how useful multimodal information can be in a task that is text-only (Experiment 1) and a task that is multimodal (Experiment 2). Our results show that our parsimonious method to unifying vision (and sensorimotor knowledge) in existing language models shows improvements in multimodal benchmarks with accessible hardware (i.e., a single GPU) as a step towards models that can be trained in settings similar to that of child language learners.

2 Related Work

Language models are trained on text. Günther et al. (2018) took up the question *do words inherit sensorimotor activation from purely linguistic context?* and showed that experience is necessary for reactivating experiential traces, but this reactivation is not a necessary condition for understanding the corresponding aspects of word meaning. We take this to mean that humans are very adept at learning new concepts from language exposure alone (i.e., abstract concepts); e.g., someone who has never seen a zebra before, but hears them described as “horses with vertical black and white stripes” can compose a connotation of what zebra denotes without direct visual exposure. However, this only works if an agent that has learned the language has the knowledge of *horses*, *black*, *white*, *stripes*, and *vertical* concepts—i.e., via direct experience, not just through linguistic exposure or encyclopedic definitions. These claims are further backed up by neuroscience research that showed that neural assemblies encode concrete content words (i.e., words

that denote visual objects) and verbs (i.e., words that denote actions) are learned and represented in different brain regions (Pulvermüller, 1999; Borghesani et al., 2019).

Rogers et al. (2020) provides a recent primer and overview of research that has attempted to uncover strengths and weaknesses of BERT and related language models (so-called *BERTology*). While our work does fit into that growing body of literature, our criticisms on current language models specifically lies in the fact that they are only trained on easy-to-obtain text. This criticism is born out in Forbes et al. (2019) which showed that BERT can guess affordances and properties of objects because that information can be found in text (e.g., a typical chair has the affordance of being sittable, and a property of having legs), but has no notion of how objects are related semantically to each other, and Da and Kasai (2019) further showed that real-world perceptual properties are likely to be assumed instead of inferred. Furthermore, Bender and Koller (2020) make a strong case that BERT learns form instead of meaning, and while the fact that BERT performs so well on many tasks is difficult to dispute, models trained on text are missing semantic information crucial for holistic language understanding.

Since before BERT which has proven powerful in many language processing tasks, efforts have been made to encode multimodal (i.e., more than just text as a learning modality) information into embeddings and language models (Takano and Utsumi, 2016; Kiros et al., 2014; Zellers et al., 2021) and recent, continued efforts towards bridging grounded visual representations to distributional representations of word meanings give credence to the claim that text-only models like BERT are missing crucial semantic information because enriching BERT with visual information improves performance in several known tasks (Kim et al., 2019; Lu et al., 2019; Li et al., 2019). These models usually treat language and vision as separate pipelines; our method directly endows the language model with visual and sensorimotor knowledge.

3 Data

In this section, we motivate and introduce of multimodal information we will use in our experiments.

The Lancaster Sensorimotor Norms The Lancaster Sensorimotor norms (Lynott et al., 2019) provide ratings (means and standard deviations) for

40,000 English words along dimensions of embodiment which capture the extent to which a concept is experienced across 11 different sensory modalities, and measures derived from those categories, listed below (each has an example word that rates highly for that modality):

- Auditory - sound; *ping*
- Gustatory - having to do with eating; *cream*
- Haptic - muscle movement; *handshake*
- Interoceptive - having to do with affect or emotion; *headache*
- Olfactory - smell; *incense*
- Visual - visual; *barcode*
- Foot-leg - haptics for foot/leg; *run*
- Hand-arm - haptics for hand/arm; *pointing*
- Head - having to do with the head; *eye*
- Mouth - haptics for mouth; *kiss*
- Torso - haptics for torso; *breath*
- Max-strength.perceptual - the highest rating across the 11 sensorimotor dimensions
- Minkowski3.perceptual - treating the 11 modalities as a vector, this represents the distance of the vector from the origin with influence of weaker dimensions attenuated
- Exclusivity.perceptual - the extent to which a concept (out of the 11) which is experienced through a single perceptual modality

The last three can be seen as aggregates from the 11 modalities; they also have *.action* values representing the extent to which a concept is experienced as an action (as opposed to *.perceptual*), and *.sensorimotor* values representing the extent a concept is experience as sensorimotor. As these norms were derived from surveys given to adults, these norms represent the degree to which the survey participants assigned those words to those categories. Though this does not represent a neurophysiological grounding of words to those modalities learned through interaction and embodiment, this serves as a useful approximation. The final set is a vocabulary of 39,707 words (after removing rows which had null values), each represented as a vector of length 39 (i.e., 11 mean, 11 stdev columns; Max-strength, Minkowski, and Exclusivity columns for different ways of aggregating the modalities). We normalize each value in the vector independently to a value between 0-1 by dividing each value over its max value. We call this the Lancaster vectors.

We performed t-SNE on the Lancaster vectors (mapping to 2 dimensions) to determine if clus-

ters would reveal any intuitions about the kinds of semantic relatedness that the words might have with each other. Some clusters emerged such as foods (presumably because they have similar *gustatory* ratings), leg-movement verbs (e.g., *walk, jump, sit*), colors with eye-related words (e.g., *purple, green, blue, dark, see, eyes*), soft things (e.g., *hug, tummy, pillow, clothes*), audio-related words (e.g., *talk, story, sound, music, lie, say*), among others.

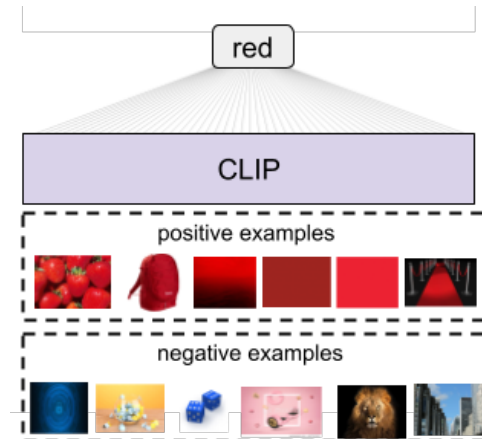


Figure 1: The *red* WAC classifiers are trained on positive and negative examples of images from Google Images for the word *red*; each image is then passed through the CLIP model. We train a binary logistic regression classifier, then extract the coefficients for the *red* vector.

Words-as-Classifiers Image Vectors The Words-as-Classifiers (WAC) approach to grounded semantics is quite simple: train a binary classifier for each word in a corpus where the features to that classifier are derived from images (Kennington and Schlangen, 2015). Each classifier is given positive and negative examples of visual denotations of each word by the images and learns a “fitness” score by the classifier. For example, the *red* classifier is given images of objects that are referred to as red in a corpus, and randomly assigned negative examples of things that are not referred to as red, as depicted in Figure 1. We follow Kiros et al. (2018) and use Google Image Search to find images using the BERT vocabulary, resulting in 27,152 words and corresponding images (some words did not result in images, and we did not download images for filler words). For each word, we perform an image search and download the top 100 images. We then follow Schlangen et al. (2016) and process each image by passing them through the recent CLIP (Jia

frozen or not during pre-training, then train for 100,000 steps.⁴ We then fine-tune the resulting ELECTRA model on the GLUE tasks using the multimodal vectors following standard fine-tuning protocols; that is, we add a linear layer with a softmax to the pre-trained model and use the ADAM solver with a learning rate of $2e-5$ for 3 epochs. As the WAC vectors were larger than ELECTRA’s expected embedding size of 128, we applied UMAP to reduce the dimensionality to 128; similarly for the WAC and Lancaster concatenated embeddings. For cases where there was no vector for a word (e.g., the [unmapped] words or words outside of the vocabulary of the Lancaster vectors), we simply used zero vectors. For Lancaster vectors, we set the ELECTRA embedding size to 39. We explored freezing the embeddings; our hypothesis is that not freezing the embeddings will lead to better results because the training regime can overpower the embeddings, but retain the multimodal knowledge.

For a broader comparison, we also compared to GloVe (Pennington et al., 2014) and several ablations where we concatenate multimodal vectors with the GloVe vectors (we used the evaluation script for GloVe provided by Wang et al. (2018)). We also use the same training and evaluation regime for the WAC and Lancaster vectors, and a concatenation of the two, on their own treating them as word-level embeddings similar to GloVe.

Results Table 1 shows the results on the GLUE benchmark. The word-level embeddings of GloVe, WAC, and Lancaster are shown in the top 5 rows of the table. As expected, these word-level embeddings are not state-of-the-art, but we notice that both Lancaster and WAC vectors perform comparably against the GloVe vectors despite only being trained on images (WAC) or derived from the Lancaster norms. Of note is a significant advantage of using the Lancaster vectors alone compared to using any other embedding or combination for the WNLI task which is co-reference and natural language inference for fiction books. This suggests that inference on fiction is helped by knowing which modalities affect each word. Interestingly, the best performing model for COLA was GloVe and Lancaster word-level embeddings; COLA is

⁴This takes about 12 hours of training on our 12GB GPU, which we opted for because it represents more data and train time than ELECTRA-small, but still a small enough amount of time to establish using this model in a co-located, interactive learning setting similar to the setting where children learn their first language.

	MRPC acc	MRPC f1	COLA corr	WNLI acc
GloVe	0.745	0.807	0.691	0.563
GloVe+lan	0.735	0.799	0.449	0.563
lan	0.711	0.778	0.691	0.596
wac	0.748	0.812	0.313	0.563
lan+wac	0.619	0.670	0.382	0.535
ELECTRA	0.730	0.835	0.449	0.563
EL-wac	0.730	0.835	0.440	0.563
EL-wac ^f	0.760	0.833	0.0	0.563
EL-lan	0.708	0.819	0.39	0.535
EL-lan-wac ^f	0.792	0.859	0.459	0.563

Table 1: GLUE development set results with GloVe, ELECTRA, Lancaster (lan), WAC models and several combinations, with (*f* and without weight freezing during ELECTRA pre-training).

a grammaticality test, which is important in language understanding, but arguably less critical in early-stage child language acquisition.

All other rows show the ELECTRA baseline and ELECTRA that uses some variation of WAC, Lancaster, or both as embeddings (denoted with the EL- prefix). The bottom part of the table compares ELECTRA with a variant of ELECTRA that uses WAC embeddings (both with and without freezing the embedding weights), ELECTRA with Lancaster embeddings and ELECTRA with WAC embeddings concatenated with the Lancaster embeddings (where the length of the WAC embeddings plus the size of ELECTRA is 128). Contrary to our hypothesis, we observe that when ELECTRA uses WAC with frozen weights, the performance on the benchmark performs better than all others, including the ELECTRA baseline. This could suggest that ELECTRA can make effective use of the visual and Lancaster embeddings by adjusting weights in the other layers of the model. The EL-lan-wac variant performed well above the ELECTRA baseline, substantiating the hypothesis that enriching the model with multimodal knowledge can improve results. Taken together, we find the results encouraging because the relatively short training regime still yielded respectable results, suggesting that ELECTRA with a visual or other multimodal embedding can be useful with less training as is the case when children learn language.

5 Experiment 2: Replacing RoBERTa embeddings, fine-tuning on Visual Dialog

The evaluation in Experiment 1 was made up of text-based tasks. In this experiment, we use an evaluation that requires knowledge of the visual world by evaluating the Lancaster and WAC vectors on the Visual Dialog task (Das et al., 2019), termed *visdial*. Moreover, Experiment 1 used pre-training on a subset of the data for only 100,000 steps. In this experiment, we evaluate using a fully pre-trained RoBERTa model by replacing its embeddings with the WAC and Lancaster vectors.

Task Following Murahari et al. (2019), given an image, dialogue history consisting of question-answer pairs, and a follow-up question about the image, the task of visdial is to predict a free-form natural language answer to the question. The visdial dataset introduced in Das et al. (2019) also includes evaluation metrics and human-annotated answers to the natural language queries about the image. Five human annotators identified which responses out of 100 candidates could be considered correct. This allows multiple answers to be correct (e.g., *yes* and *yeah* are semantically identical).

Metrics We report the following metrics:

- **R@1** Rate of times the top-ranked answer is a correct one; i.e., accuracy.
- **R@5** Rate of times correct answers are in the top-five ranked answers.
- **MRR** *Mean Reciprocal Rank* is the multiplicative inverse of the rank of the first correct answer.
- **NDCG** *Normalized Discount Accumulative Gain* is a measure of ranking quality that takes the top K ranked options, where K is the number of answers marked as correct by a least one annotator; in this measure, the fraction of annotators that marked a particular answer as correct is taken into account.

Baseline and Procedure We report the values for the model described in Murahari et al. (2019) for our baseline—work which builds on ViBERT (Lu et al., 2019), a parallel model of vision and language used for the visual dialogue task—and leverage their model with our custom, multimodal embeddings. Their model uses two transformers, one for the language modality and one for the visual modality. As explained in Lu et al. (2019),

the interaction between the two transformers is mediated by two co-attention layers where attention in one modality is conditioned on inputs from the other modality. Murahari et al. (2019) adapted the ViBERT model for the visdial task by using a pre-trained language model trained on English Wikipedia and the BooksCorpus (Zhu et al., 2015) using masked language modeling and next sentence prediction losses. They then frame the task as a next-sentence prediction task (whereas the original ViBERT was modeled to generate descriptions of images). They then use the Conceptual Captions (Sharma et al., 2018) and Visual Question Answering (VQA) (Antol et al., 2015) datasets (using masked image region, masked language modeling, and next sentence prediction losses) to train the two transformers. They then fine-tune on the visdial task (also using masked image region, masked language modeling, and next sentence prediction losses). The underlying architecture uses a pre-trained BERT language model (i.e., bert-base-uncased) as a starting point before training on the Wikipedia, BooksCorpus, Conceptual Captions, and VQA datasets. This constitutes our baseline. We don’t consider the dense representations from Murahari et al. (2019) due to hardware limitations.

We altered their architecture by replacing the RoBERTa pre-trained embedding layer with the Lancaster and WAC vectors, as depicted in Figure 3. We then fine-tuned on the visdial task using their training regime.⁵ We explain how we made vectors compatible with their architecture.

Vocabulary: RoBERTa & AoA Abstract words do not have concrete, visual denotations, such as *utopia* or *justice*, so it does not make theoretical sense to include a WAC embedding for words that are clearly abstract because whatever set of images represents those concepts may not have useful semantic information. Moreover, children learning their first language learn concrete concepts before they learn more abstract concepts (Borghi et al., 2019; Ponari et al., 2018). To explore if RoBERTa could make use of a WAC embedding that uses words that are more aimed at a child vocabulary, we report results of filtering out words not in the the Age-of-Acquisition (AoA) list (Kuperman et al., 2012b). AoA a list of 30,000 English words rated for the average age when children first speak those words (avg 11 years; std 3.0, most common words

⁵<https://github.com/vmurahari3/visdial-bert>

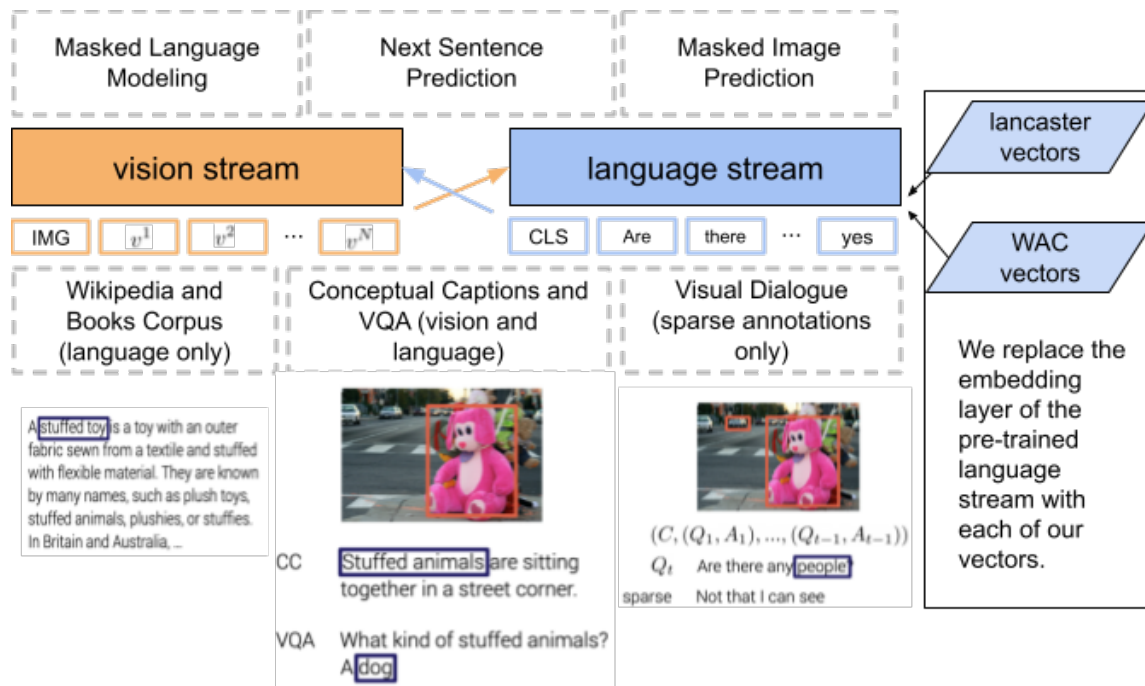


Figure 3: Adapted from Murahari et al. (2019). Our approach uses the same pre-training datasets, architecture, and losses. During the final fine-tuning on the Visual Dialogue data, however, we replace the RoBERTa embeddings with the Lancaster and WAC embeddings.

are for ages 2-14). This resulted in 9,627 remaining words in the vocabulary; all other words were set to an embedding of zeros.

Lancaster vectors Similar to AoA, the Lancaster Norms has a predefined vocabulary, which, when compared to the RoBERTa vocabulary results in 11,402 words in both. For each word in the RoBERTa vocabulary that was also in the Lancaster norms, we replaced the RoBERTa embedding with the Lancaster vector for that word; otherwise words retained the original RoBERTa embedding. As their model expects vectors of size 768 (the embedding size for RoBERTa), but the Lancaster vectors are only size 39, we padded the rest of the vector with zeros.

WAC vectors We use the vocabulary from the RoBERTa tokenizer as with the Lancaster Vectors, which results in a 27,152-word overlap with the WAC vectors. As the WAC vectors have a dimensionality of 513, smaller than the required size of RoBERTa’s 768, we padded zeros after each vector. All vectors that did not exist in the WAC set were zero vectors of size 768. We followed a training regime for two settings:

- **no-freeze** The embedding layer was not frozen so as to allow weight changes during

training.

- **2-freeze** The embedding layer was frozen for two epochs, then the weights were unfrozen for the rest of training; prior work has shown that freezing layers after a certain number of epochs can improve results (Liu et al., 2021); we opt for two because it still allows the fine-tuning to overpower the exiting embeddings if needed and preliminary results showed that freezing the weights for all epochs resulted in poor model performance.

We trained each model for 20 epochs in total, which is the default training setting for this task. We used the settings that were used to train the baseline model (i.e., learning rate of $2e-5$). We report the results of the baseline model and the variants of our above changes.⁶ Compared to Experiment 1 with the GLUE benchmark, the approach taken in this section fundamentally changes how the Lancaster and WAC embeddings are applied to RoBERTa; here the Lancaster and WAC embeddings are used on a pre-trained model. We hy-

⁶Note that the baseline we are comparing to here is lower than what is reported on the leaderboard <https://eval.ai/web/challenges/challenge-page/518/leaderboard>. This is partially due to the fact that our training regime was altered due to hardware limitations (i.e., we could only use a batch size of 8 on a single 12GB GPU).

no freeze	MRR	R@1	R@5	NDCG
baseline	64.92	50.52	82.98	56.82
lan	64.61	49.98	82.6	58.10
2-freeze	MRR	R@1	R@5	NDCG
lan	63.77	49.03	81.63	57.53
wac	66.38	52.25	83.8	61.47
lanwac	63.93	49.025	82.65	57.73
wac-aoa	66.79	52.75	84.05	60.44

Table 2: Experiment 2 results for the visdial task: baseline RoBERTa embedding, Lancaster norms (lan), WAC vectors, and concatenated (lanwac), not frozen, and only frozen for 2 epochs (bottom section).

hypothesize that RoBERTa will improve with WAC embeddings, as well as the Lancaster concatenated to WAC (denoted *lanwac*), though the Lancaster embedding on its own may be too small to make a difference. As words that are learned earlier in a child’s life are generally more concrete, we hypothesize that RoBERTa will improve when WAC only uses words from the AoA data as more abstract terms are represented by zero vectors.

Results Table 2 shows the results for the visdial task. Though it is clear that RoBERTa is doing the heavy lifting, when added to RoBERTa, the Lancaster and WAC vectors show improvements over the RoBERTa baseline for some metrics. As noted in Murahari et al. (2019), the NDCG metric is actually counter to MRR, but is important because it takes multiple dialogue response annotations into account. For cases where the Lancaster and WAC models yield better performance, these results suggest that a pre-trained language model can make use of adding multimodal knowledge in the form of vectors derived from multimodal knowledge (Lancaster) and visual (WAC) for the visdial task.

RoBERTa that uses the WAC embedding especially shows respectable results in the visdial task, particularly when the embedding uses the AoA vocabulary (we only considered AoA for WAC because WAC performed better than lanwac in this experiment). The WAC vectors were trained on very noisy data, yet despite the noise and the parsimonious model, there is some information about the visual world that enriches the baseline model. The variant trained with frozen weights for 2 epochs, then unfrozen for the remaining 18 epochs had respectable performance across all metrics.⁷

⁷As a sanity check, we also evaluated using randomly generated embeddings which performed only slightly worse

6 Conclusion

The main contribution of this paper is to explore using the Lancaster Sensorimotor Norms and the Words-as-Classifiers model as vectorized knowledge from the physical world on the GLUE and Visual Dialog tasks. Lancaster norms performed well on their own in one GLUE task compared to other word embeddings like GloVe, and coupled with the WAC vectors as the embedding in an ELECTRA model, they performed respectably on the GLUE task. The WAC vectors, when used as embeddings in the RoBERTa model performed well on the Visual Dialog task, particularly when the vocabulary was more restricted to the Age of Acquisition vocabulary. Crucially, this work differs from other visually grounded models because the grounded knowledge is part of the language model itself (i.e., the embeddings) rather than computed in parallel and added for a task-specific purpose. Moreover, standard language models cannot actually identify denotations when they are present; i.e., ELECTRA and RoBERTa are not actually capable of determining if an object is *red* or *soft* from observing that object—a basic ability for a language learning child—simply because those models cannot observe the world outside of text, though the purpose of the WAC (and models like ViBERT) model is to do just that: identify denotations; by coupling WAC with ELECTRA and RoBERTa, both models can make use of that capability.

This work is critical in our ongoing efforts towards a model that learns language in a co-located setting in an embodied platform. In particular, our knowledge from this paper informs us that the ELECTRA model with embeddings tied to WAC classifier weights is a good candidate for live interaction of a robot that is learning words from a human collaborator because the ELECTRA-WAC model can function with small amounts of data and the embedding layer can successfully be tied to weights of the WAC classifiers. We leave implementation and evaluation of this model on a robotic platform for future work.

Acknowledgements Thanks to the anonymous reviewers whose comments really helped strengthen the paper. Also thanks to NVIDIA for donating that GPU that was used for the experiments.

than baseline when frozen for 2 epochs, but the results of wac-aoa are significantly better.

References

- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. [VQA: Visual question answering](#). In *Proceedings of the IEEE International Conference on Computer Vision*.
- Lawrence W Barsalou. 2008. [Grounded Cognition](#). *Annual Review of Psychology*, (59):617–645.
- Emily M Bender and Alexander Koller. 2020. [Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data](#). In *Association for Computational Linguistics*, pages 5185–5198.
- Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. [Experience Grounds Language](#). *arXiv*.
- Valentina Borghesani, Marco Buiatti, Evelyn Eger, and Manuela Piazza. 2019. [Conceptual and Perceptual Dimensions of Word Meaning Are Recovered Rapidly and in Parallel during Reading](#). *Journal of Cognitive Neuroscience*, 31(1):95–108.
- Anna M Borghi, Laura Barca, Ferdinand Binkofski, Cristiano Castelfranchi, Giovanni Pezzulo, and Luca Tummolini. 2019. Words as social tools: Language, sociality and inner grounding in abstract concepts. *Phys. Life Rev.*, 29:120–153.
- Angelo Cangelosi and Matthew Schlesinger. 2015. *Developmental robotics: From babies to robots*. MIT press.
- Eve V Clark. 2013. *First language acquisition*. Cambridge University Press.
- Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. [ELECTRA: Pre-training text encoders as discriminators rather than generators](#).
- Jeff Da and Jungo Kasai. 2019. [Cracking the Contextual Commonsense Code: Understanding Commonsense Reasoning Aptitude of Deep Contextual Representations](#). In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*. Association for Computational Linguistics.
- Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Stefan Lee, Jose M.F. Moura, Devi Parikh, and Dhruv Batra. 2019. [Visual Dialog](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(5).
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. [BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](#).
- Ezequiel A Di Paolo, Elena Clare Cuffari, and Hanne De Jaegher. 2018. *Linguistic bodies: The continuity between life and language*. Mit Press.
- Paul Dourish. 2001. [Where the Action Is: The Foundations of Embodied Interaction](#). *Where the action is the foundations of embodied interaction*.
- Maxwell Forbes, Ari Holtzman, Yejin Choi, and G Allen. 2019. [Do Neural Language Representations Learn Physical Commonsense?](#) *arXiv*.
- Fritz Günther, Carolin Dudschig, and Barbara Kaup. 2018. [Symbol Grounding Without Direct Experience: Do Words Inherit Sensorimotor Activation From Purely Linguistic Context?](#) *Cognitive Science*, 42:336–374.
- Stevan Harnad. 1990. [The symbol grounding problem](#). *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.
- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. [Scaling up visual and Vision-Language representation learning with noisy text supervision](#).
- Mark Johnson. 2008. *The meaning of the body: Aesthetics of human understanding*. University of Chicago Press.
- Casey Kennington and David Schlangen. 2015. [Simple Learning and Compositional Application of Perceptually Grounded Word Meanings for Incremental Reference Resolution](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 292–301, Beijing, China. Association for Computational Linguistics.
- Donghyun Kim, Kuniaki Saito, Kate Saenko, Stan Sclaroff, and Bryan A. Plummer. 2019. [MULE: Multimodal Universal Language Embedding](#).
- Jamie Ryan Kiros, William Chan, and Geoffrey E Hinton. 2018. [Illustrative Language Understanding: Large-Scale Visual Grounding with Image Search](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pages 922–933, Melbourne, Australia. Association for Computational Linguistics.
- Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. [Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models](#). In *arXiv preprint arXiv:1411.2539*, pages 1–13.
- Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012a. [Age-of-acquisition ratings for 30,000 English words](#). *Behavior Research Methods*, 44(4):978–990.

- Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012b. Age-of-acquisition ratings for 30,000 english words. *Behav. Res. Methods*, 44(4):978–990.
- George Lakoff and Mark Johnson. 1999. *Philosophy in the flesh: The embodied mind and its challenge to western thought*, volume 640. Basic books New York.
- Liunian Harold Li, Mark Yatskar, Da Yin, Cho Jui Hsieh, and Kai Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv*.
- Yuhan Liu, Saurabh Agarwal, and Shivaram Venkataraman. 2021. AutoFreeze: Automatically freezing model blocks to accelerate fine-tuning.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks.
- Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2019. The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, pages 1–21.
- Lorraine McCune. 2008. *How Children Learn to Learn Language*. Oxford University Press.
- Daniele Moro, Stacy Black, and Casey Kennington. 2019. Composing and embedding the words-as-classifiers model of grounded semantics.
- Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. 2019. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. *arXiv preprint arXiv:1912.02379*.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.
- Marta Ponari, Courtenay Frazier Norbury, and Gabriella Vigliocco. 2018. Acquisition of abstract concepts is influenced by emotional valence. *Dev. Sci.*, 21(2).
- Friedemann Pulvermüller. 1999. Words in the brain’s language. *Behavioral and Brain Sciences*.
- Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A Primer in BERTology: What we know about how BERT works. *arXiv*.
- David Schlangen, Sina Zariess, and Casey Kennington. 2016. Resolving References to Objects in Photographs using the Words-As-Classifiers Model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1213–1223.
- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernamed, image alt-text dataset for automatic image captioning. In *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*.
- L B Smith and L Samuelson. 2009. Objects in Space and Mind: From Reaching to Words. In *The Spatial Foundations of Language and Cognition*.
- Linda Smith and Michael Gasser. 2005. The Development of Embodied Cognition: Six Lessons from Babies. *Artificial Life*, (11):13–29.
- Katsumi Takano and Akira Utsumi. 2016. Grounded Distributional Semantics for Abstract Words. *CogSci*, pages 2171–2176.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.
- Zhaofeng Wu, Hao Peng, Noah A Smith, and Paul G Allen. 2021. Infusing Finetuning with Semantic Dependencies. *Transactions of ACL*.
- Rowan Zellers, Ari Holtzman, Matthew Peters, Roozbeh Mottaghi, Aniruddha Kembhavi, Ali Farhadi, and Yejin Choi. 2021. PIGLeT: Language grounding through Neuro-Symbolic interaction in a 3D world.
- Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision*.