2023

Tiny Language Models Enriched with Multimodal Knowledge from Multiplex Networks

Clayton Fields
Boise State University

Osama Natouf
Boise State University

Andrew McMains
Boise State University

Catherine Henry
Boise State University

Casey Kennington
Boise State University

Publication Information

Fields, Clayton; Natouf, Osama; McMains, Andrew; Henry, Catherine; and Kennington, Casey. (2023). "Tiny Language Models Enriched with Multimodal Knowledge from Multiplex Networks". In A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen, and R. Cotterell (Eds.), *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning* (pp. 47-57). https://doi.org/10.18653/v1/2023.conll-babylm.3

Tiny Language Models Enriched with Multimodal Knowledge from Multiplex Networks

Clayton Fields claytonfields@u.boisestate.edu
Osama Natouf osamanatouf@u.boisestate.edu
Andrew McMains andrewmcmains@u.boisestate.edu
Catherine Henry catherinehenry@u.boisestate.edu
Casey Kennington caseykennington@boisestate.edu
Department of Computer Science
Boise State University

Abstract

Large transformer language models trained exclusively on massive quantities of text are now the standard in NLP. In addition to the impractical amounts of data used to train them, they require enormous computational resources for training. Furthermore, they lack the rich array of sensory information available to humans, who can learn language with much less exposure to language. In this study, performed for submission in the BabyLM challenge, we show that we can improve a small transformer model's data efficiency by enriching its embeddings by swapping the learned word embeddings from a tiny transformer model with vectors extracted from a custom multiplex network that encodes visual and sensorimotor information. Further, we use a custom variation of the ELECTRA model that contains less than 7 million parameters and can be trained end-to-end using a single GPU. Our experiments show that models using these embeddings outperform equivalent models when pretrained with only the small BabyLM dataset, containing only 10 million words of text, on a variety of natural language understanding tasks from the GLUE and SuperGLUE benchmarks and a variation of the BLiMP task.

1 Introduction

The field of natural language processing is now dominated by large-scale transformer models such as GPT-3 (Brown et al., 2020). These models are characterized not only by their enormous size—billions of parameters—but also the huge datasets that are used in their pretraining. The 200 billion text tokens used to train GPT-3 are dwarfed by the 1.4 trillion used to train Chinchilla (Hoffmann et al., 2022). Huge model sizes and enormous pretraining datasets make research on pretraining language models impractical for all but the most lavishly funded industry research groups.

Beyond the practical problems posed by such massive data inputs, the current methods for language modeling require vastly more resources to learn and perform language tasks than human beings do. American children, for example, begin speaking around the age of 1 year on average (Gilkerson et al., 2017), and studies suggest that they have only heard around 5 million words before the onset of recognizable words (i.e., beyond babbling). Even the medium-sized BERT model was trained with a 3.3 billion word corpus (Devlin et al., 2019)—orders of magnitude more than human beings require to begin speaking and reach language proficiency. This disparity suggests that current methods in natural language procession (NLP) are missing crucial aspects of language learning.

One thing which models trained exclusively on text undoubtedly lack is a genuine connection between concrete words, such as *red*, and the physical world they describe. Human beings learn to speak with the aid of their sensory impressions, emotions and a rich environment of social cues (Smith and Gasser, 2005), which is to say that human language is grounded in the human sensory experience (Harnad, 1990). To use the same example, the word *red* is grounded in the visual perception of color. In contrast, transformer NLP models only have access to text and can only define words in terms of other words, following the distributional hypothesis of linguistic meaning. The lack of concrete sensory information is one possible reason why transformers require so much text and compute to learn perform basic human language tasks.

In this study, conducted as part of the BabyLM challenge (Warstadt et al., 2023), we seek to improve a tiny transformer model's data efficiency by providing it with a facsimile of that missing sensory information. Specifically, we follow the approach taken by Kennington (2021) and replace a pretrained model's word embeddings with vector representations that encode visual and sensorimotor information. Our approach differs in that we extract our embeddings from a custom multiplex network

Proceedings of the 27th Conference on Computational Natural Language Learning: Volume 2: The BabyLM Challenge, pages 47–57

that captures visual and sensorimotor relationships between words. Multiplex networks are multi-layer graphs, and researchers such as Ciaglia et al. (2023) have demonstrated their potential for representing various types of semantic relationships. Our multiplex network consists of two layers: a visual layer and a sensorimotor layer, which we explain below.

As one of the goals of our study, and the BabyLM challenge in general, is to increase a model's data effeciency, we pretrain our models with the cognitively plausible 10M word dataset provided by the BabyLM organizers. Additionally, with the aim of making research on pretraining transformer models more accessible, we use a tiny variation of ELECTRA (Clark et al., 2020) with fewer than 7 million parameters that can be trained on a single modestly priced GPU. This approach allows us to simultaneously address the topics of data efficiency and parameter efficiency. The contributions of our study can be summarized as follows:

- We show that tiny models can be as effective as models twice their size in a scarce pretraining data regime.
- We show that embeddings from a multiplex network that encodes visual and sensorimotor information related to English words can improve the data efficiency of a small transformer model.
- Models using these embeddings can perform as well as similar models that are trained with ten times the amount of pretraining data.

In the following section we present some work related to the topics associated with our modeling approach. In Section 3, we introduce both the pretraining datasets and the data we use to evaluate our models' downstream performance. Section 4 describe the ELECTRA model we use as the basis for our study and the multiplex network from which we extract our novel embeddings. Finally, we describe our experiments in Section 5 and conclude in Section 6.

2 Related Work

Data Efficient Pretraining for Language Models To date, model compression techniques for transformers have recieved more attention than data efficiency. There has, however, been some research directly addressing pretraining data types and sizes for transformers. Micheli et al. (2020) and Martin et al. (2019) experimented with reducing the absolute amount of training data in French language models. They showed that full sized French language transformer models can perform well on select tasks with significantly less pretraining data. Warstadt et al. (2020b) and Zhang et al. (2020) investigated the effect of different pretraining data volumes on the grammatical knowledge of the RoBERTa-base model using probing techniques. Another example is the BabyBERTa model introduced in Huebner et al. (2021). Here the authors used the CHILDES (MacWhinney, 2000), a small dataset of transcribed, child-directed speech to train a variation of RoBERTa (Liu et al., 2019). Notably, the CHILDES dataset is one the components of the dataset used in this study.

Small-Scale Language Models The process of creating transformers with fewer parameters and less computational demands has been an active area of research. A number of techniques for compressing transformers exist, but knowledge distillation is probably the most common. In knowledge distillation, a full-sized teacher model is used to train a smaller student network. DistilBERT (Sanh et al., 2019), TinyBERT (Jiao et al., 2019), MiniLM (Wang et al., 2020) and MobileBERT (Sun et al., 2020) are popular examples of compact transformers distilled using full sized BERT models as teachers. These methods produce effective smaller models, but they can't directly address the amount of input data required, and the training process still requires a using a full-sized teacher model trained with a large text corpus.

Multiplex Networks and Language Multiplex networks have been explored as a way of modeling the language acquisition process in human children (Stella et al., 2017, 2018). Citraro et al. (2023) used a complex network that incorporated phonetic, cooccurance, frequency, length, polysemy, (among others) to explore potential mental strategies for early word learning. Ciaglia et al. (2023) recently brought aspects of NLP into multiplex networks by including word-level embedding knowledge, which we build on here.

3 Data & Benchmarks

In this section we describe the datasets we use both for pretraining and for downstream evaluation of our models. As this paper was intended as part of

Dataset	Domain	Words-10M	Words-100M	Reference
CHILDES	Child-directed speech	0.44M	4.21M	MacWhinney (2000)
British National Corpus (BNC)	Dialogue	0.86M	8.16M	Consortium (2007)
Children's Book Test	Children's books	0.57M	5.55M	Hill et al. (2016)
Children's Stories Text Corpus	Children's books	0.34M	3.22M	Edenbd (2021)
Standardized Project Gutenberg	Written English	0.99M	9.46M	Gerlach and Font-Clos (2018)
OpenSubtitles	Movie subtitles	3.09M	31.28M	Lison and Tiedemann (2016)
QCRI Educational Domain Corpus	Video subtitles	1.04M	10.24M	Abdelali et al. (2014)
Wikipedia	Wikipedia	0.99M	10.08M	Wikimedia
Simple Wikipedia	Wikipedia (Simple)	1.52M	14.66M	Wikimedia
Switchboard Dialog Act Corpus	Dialogue	0.12M	1.18M	Stolcke et al. (2000)

Table 1: Composition of the BabyLM Datasets, from Warstadt et al. (2023).

the BabyLM competition, we use only the datasets provided by the organizers and their evaluation pipeline to assess our results. Although this information is described in Warstadt et al. (2023) and its associated references, we provide a brief summary in the interest of completeness and readability.

BabyLM Datasets The pretraining data provided for the BabyLM competition consists of two datasets with roughly 10 million and 100 million words. We will refer to these as the BabyLM-10M and the BabyLM-100M datasets. These datasets are meant to be developmentally plausible and are inspired by language input for children. The compositions of the datasets are described in Table 1 with references for each source dataset. The 10M word dataset is a uniform sample of the 100M word dataset.

GLUE and SuperGLUE Many of the datasets we use for fine-tuning and evaluation, are drawn from the GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) benchmarks. Each consists of a suite of natural language understanding tasks and they are among the most commonly used benchmarks for evaluating natural language understanding. From GLUE, we use 7 of the 9 tasks in the suite. COLA, a grammatical acceptability task and SST-2, a sentiment classification task, are both single sentence tasks. QQP and MRPC are both two-sentence paraphrase tasks. Finally, MNLI, QNLI and RTE are natural language inference tasks. From SuperGLUE we use three tasks: BoolQ and MultiRC are both question answering tasks, and WSC is a co-reference task.

BLiMP The Benchmark of Linguistic Minimal Pairs (BLiMP) is a set of 67,000 pairs of sentences designed to test a language model's grasp of English grammar introduced in Warstadt et al. (2020a). The full BLiMP set consists of 67 sets of 1,000 pairs of English sentences covering 12 different grammatical phenomena. The sentences were generated from grammars created by linguists with each pair containing one grammatically correct and one incorrect sentence that differ by only a single edit. On aggregate, the creators found that humans agreed with the labels over 96 percent of the time. For each pair, a language model trained with causal language modeling, e.g. GPT-3, is considered to be successful if it assigns a higher likelihood to the grammatically correct sentence. BLiMP was concieved as a zero-shot task and many popular language models can be evaluated on BLiMP without fine-tuning using either the log-likelihood or the pseudo-log-likelihood scoring method (Wang and Cho, 2019; Salazar et al., 2020). Unfortunately, the ELECTRA model (Clark et al., 2020) that we use in our experiments is not one of them and we therefore adopt a minimal fine-tuning approach to the BLiMP task. To keep with the zero-shot spirit of the task as much as possible, we cast BLiMP as binary choice task with only enough training to remove large variances from run to run. The details of the fine-tuning regime that we used can be found in Section 5.2.

MSGS The MSGS dataset, pronounced *messages*, was introduced in Warstadt et al. (2020b) with the goal of studying the inductive biases of NLP models. The task challenges models to classify sentences based on either surface features, e.g. *Does the sentence contain the word "the"?*, or linguistic features, *Does the sentence contain an irregular past-tensed verb?*. In total the set contains 4 surface features and 5 linguistic features. By pairing a sentences containing a surface feature and a linguistic feature, the task tests a model's preference for surface features versus more meaningful linguistic features. The MSGS dataset contains 20 tasks, one for each possible combination of surface

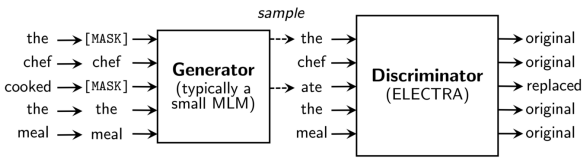

Figure 1: The ELECTRA model is a Generator-Discriminator ensemble. The Discriminator is tasked with determining if the Generator properly guessed a masked word; borrowed from (Clark et al., 2020).

and linguistic features, as well as 9 control tasks whose sentences contain only a surface or linguistic feature being tested. From this set we use 5 control tasks and 6 from the set of combinations. A more detailed description of these tasks can also be found in Section 5.2.

4 Method

4.1 ELECTRA-Tiny

In this subsection we describe ELECTRA (Clark et al., 2020), the language model that forms the basis of our experiments. In place of masked language modeling, ELECTRA pretrains a transformer encoder stack, structurally identical to BERT's, by corrupting some input tokens through replacing them with plausible alternative words sampled from a small generator network. A larger discriminator model then predicts whether each token is corrupted or not. After training, the generator is discarded and the discriminator is used for downstream tasks. See Figure 1 for an illustration of the ELECTRA model. Clark et al. (2020) show that this strategy leads to better results with less data and less compute than causal language modeling or standard masked language modeling, making it a natural choice for use in these experiments.

We make use of two architectural variations of ELECTRA. ELECTRA-Small is a scaled down version of the base model that was also introduced in Clark et al. (2020). This small version of ELECTRA has embedding vectors of dimension 128, 12 layers and a hidden size of 256. Following the original transformer architecture in (Vaswani et al., 2017), the intermediate size of each layer's feed-forward network is 4 times the model's hidden size, or 1024. In total, it contains only 14 million parameters and can be trained end to end using a single GPU. The other model we use is an even smaller variation that we call ELECTRA-Tiny and it was introduced and evaluated in Fields and Kennington (2023). ELECTRA-Tiny is very small, containing only 6.7 million parameters, approximately half as many as ELECTRA-Small. The tiny variation of ELECTRA however, is not simply scaled down with the same proportions. The model has an embedding size of 96, a hidden size of 196 and rather than a 4-fold expansion of the feed-forward network's intermediate layer, reduces the layer's size to 128. Finally, to compensate for the models decreased width, it contains 18 layers. The combination of an efficient pretraining method and small model sizes make these models ideal for our purposes. Figure 2 shows how ELECTRA-Small and ELECTRA-Tiny compare in their underlying sizes.

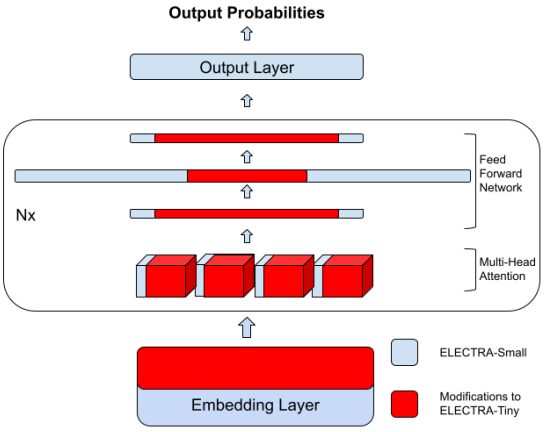

Figure 2: Relative size comparison of ELECTRA-Small (blue) with ELECTRA-Tiny (red). ELECTRA-Tiny has smaller embeddings, hidden size, and intermediate size, but has more hidden layers.

4.2 Enriching ELECTRA-Tiny with Multimodal Knowledge: Multiplex Networks

A multiplex network is a type of conceptual network that consists of multiple *layers*, where each layer represents a different type of relationship or interaction between nodes. In a multiplex network, the nodes can be the same across layers (which means nodes can be duplicated across different layers), but the relationships between them may vary. Figure 3 shows an example of a multiplex network that has five nodes composed of two layers.

A multiplex network can be represented as a multilayer graph, where the nodes are connected by edges in each layer and potentially across different layers. The layers can capture different aspects or modalities of the interactions between the entities. For example, in Ciaglia et al. (2023), the authors used a small vocabulary of words as the

nodes, where the layers of their multiplex network were represented by free association (i.e., when presented with a word, participants were asked to write the first word that comes to mind), visual relationships, sensory relationships, and distributional semantic relationships.

In any weighted multiplex network, the connections between nodes can have different types or strengths, depending on the layer and the weight on the edge. This allows for a more comprehensive representation of the relationships between nodes, as different layers capture different aspects of the relationships. Multiplex networks can provide richer information than a standard network that is made up of only one layer.

Ciaglia et al. (2023)'s network was composed of layers derived from word embeddings, free association, visual and sensorimotor vectors. Their work only included a vocabulary of 531; we extend their work by dramatically increasing the vocabulary covered in the multiplex network. Important for our work here is to only use layers that are cognitively plausible for a language learning child to have as they speak their initial words. As the embedding and free association layers were derived using adult written text and adult participants respectively, we leave them out of our model here and focus only on the visual and sensorimotor layers; modalities that children certainly have access to and from which they build their language learning.

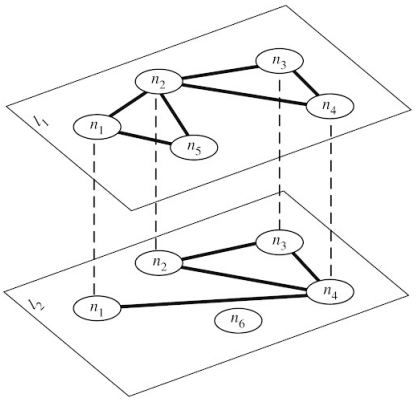

Figure 3: From Bródka et al. (2018). A visual representation of a multiplex network demonstrating interconnected layers. The dotted lines represent the interlayer connections (node relationships across layers) while the solid lines represent intralayer connections (node relationships within a layer).

Visual Layer: Words-as-Classifiers To represent the visual layer, we use the word-as-classifiers (WAC) model of grounded lexical semantics (Kennington and Schlangen, 2015). We train a binary logistic regression classifier that learns the *fitness* of identifying a visual aspect (e.g., *redness*) on images for each vocabulary word (e.g., images of dogs for the word *dog*) and randomly sampled negative images from other words. We use 100 positive images for each word with a ratio of 3 negative examples for each positive example. Each image is encoded as a vector for training using the CLIP model (Radford et al., 2021). Once each classifier for each word is trained, we take the learned coefficients and bias term as the vector (size 513) representing the word.

Sensorimotor Layer: Lancaster Sensorimotor Norms The Lancaster dataset (Lynott et al., 2020) uses the Lancaster Norms rating as a measure of perceptual and action strength on about 40K English words. Sensorimotor information plays a fundamental role in cognition and provides a useful connection between words and understanding. For example, the word *kick* has a strong sensorimotor grounding in leg and foot movement, the word *sour* is grounded in taste, and the word *ping* is grounded in auditory processing. For each word in the dataset, raters were asked to rate how strongly that word is associated with a specific perceptual modality including touch, hearing, smell, taste, vision, and interoception, and five action effectors including mouth/throat, hand/arm, foot/leg, head excluding mouth/throat, and torso. The dataset reports the mean and standard deviation of the ratings, as well as ways of aggregating the ratings, which we use as a vector (size of 39) for each word.

Constructing the Multiplex Network Kennington (2021) used the WAC and Lancaster vectors as the embedding layer for a language model in their experiments. We use the same modalities here, but we first combine the two modalities into a multiplex network and then extract the embeddings from the network to use for the embedding layer. This approach, we argue, is more cognitively plausible because words are associated by vision and sensorimotor modalities at a more categorical level, which is the basis of cognition (Harnad, 2017).

To create the network, we had to determine if two words had a relationship within a layer. To do so, we computed the cosine distance between all possible word pairs in each layer, forming relationships between words (i.e., the nodes) if the cosine

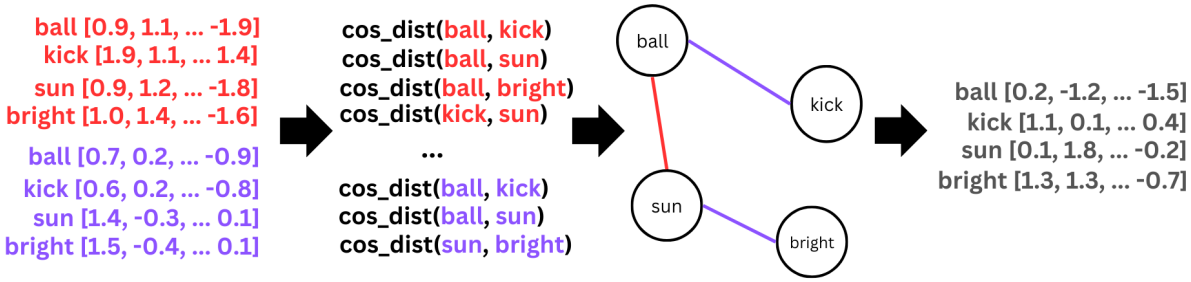

Figure 4: **Our Methodology**: vectors from vision (red, top) and sensorimotor (purple, bottom) are compared to each other using cosine distance. Word pairs that are above a specific threshold are added to the network, where connections from different modalities are retained in a multiple network (e.g., *ball* and *kick* are similar in sensorimotor vectors, whereas *ball* and *sun* are shaped similarly visually). Finally, we use the MultiVERSE algorithm to map from the multiplex network to the vector embeddings used in ELECTRA-Tiny.

distance was above a specific threshold. The selection of these thresholds was motivated by the goal of striking a balance between sparsity and density in the network, taking into account computational constraints associated with extracting the network embeddings. We then used those resulting embeddings in the ELECTRA-Tiny model. The process is depicted visually in Figure 4. In our experiments, we use three networks: visual only (cosine distance threshold of 0.94, vocabulary size of 21,235), sensorimotor only (cosine distance threshold of 0.27, vocabulary size of 22,054), and a multiplex combination of visual and sensorimotor (same thresholds as individual layers, vocabuary size of 35,607).

From Multiplex Network back to Embeddings: MultiVERSE We use the MultiVERSE algorithm, introduced in Pio-Lopez et al. (2021) to map from our multiplex network representation back to embeddings to be used in a language model. MultiVERSE is a network representation learning algorithm tailored for multiplex networks that aims to capture complex interactions by considering interdependencies between layers. By employing a unified framework integrating the multiplex network structure, node attributes, and meta-path-guided random walks, MultiVERSE learns low-dimensional node representations, clustering nodes with similar relationships. Importantly, the Random Walk with Restart algorithm explores different layers in parallel (i.e., the layers are represented as separate *sub*-networks instead of a complete network), retraining multiplex relationships. In contrast, other well-known algorithms that map from networks to embeddings, like node2vec (Grover and Leskovec, 2016), do not adequately retain multiplex information from each individual layer, making MultiVERSE a superior choice for mapping multiplex network representations to meaningful embeddings for downstream language model applications while maintaining the meaning from different relationships in different layers. This resulted in embeddings for many words in the ELECTRA vocabulary, but for the words that were not represented, we simply used zero vectors.

5 Experiments

5.1 Experiment 1: GLUE and SuperGLUE Tasks

In this experiment we determine to what extent embeddings extracted from our multiplex network can improve our small scale models on natural language understanding tasks. We begin by pretraining ELECTRA-Tiny on the BabyLM-10M word dataset described in Section 3 for 10 epochs using the Adam optimizer (Kingma and Ba, 2014), with a lerning rate of 1e-5 and a batch size of 64. Following Kennington (2021), our strategy is to swap the learned word embeddings from the pretrained model with our own embeddings prior to finetuning and evaluation. Using the MultiVERSE algorithm we extract three sets of embeddings from our network corresponding to the WAC visual layer, the Lancaster Norm layer and the multiplex combination of both layers. We then finetune the ELECTRA-Tiny model with each embedding set as well as a control model using ELECTRA's learned word embeddings. We fine-tune each model for ten epochs with a learning rate of 5e-5, and a batch size of 64. For the sake of comparison, we also trained ELECTRA-Tiny on the 100M word dataset and the original ELECTRA-Small (Clark et al., 2020) on the 10M dataset.

Model	Data	COLA	RTE	MultiRC	QQP	QNLI	MNLI	MNLI-mm	SST2	Avg.
Tiny	10M	69.5	48.9	60.4	63.2	56.6	42.2	42.8	81.0	60.7
Tiny-L	10M	69.5	49.5	55.9	63.4	57.4	48.6	49.1	78.0	58.9
Tiny-V	10M	67.1	62.6	59.5	62.4	57.9	49.6	51.2	82.1	63.2
Tiny-LV	10M	67.1	62.6	59.5	62.4	57.9	49.6	51.2	82.1	63.2
Small	10M	69.3	50.5	56.3	62.2	57.0	39.5	39.0	81.3	59.9
Tiny	100M	69.5	54.5	56.0	64.3	58.8	51.5	51.5	81.7	62.8
Maj. Label	10M	69.5	53.1	59.9	53.1	35.4	35.7	35.7	50.2	52.6
OPT-125m	10M	64.6	60.0	55.2	60.4	61.5	60.0	57.6	81.9	63.4
RoBERTa	10M	70.8	61.6	61.4	73.7	77.0	73.2	74.0	87.0	71.4
T5-base	10M	61.2	49.4	47.1	66.2	62.0	48.0	50.3	78.1	60.9

Table 2: **GLUE and SuperGLUE results for the initial datasets on our various models.** Note that the last size models in the table are baselines included for the sake of comparison.

Model	Data	MRPC	RTE	MultiRC	QQP	QNLI	MNLI	MNLI-mm	SST2	Avg.
Tiny	10M	82.0	53.5	56.6	66.3	53.1	62.4	62.0	82.7	64.5
Tiny-L	10M	82.0	49.5	58.1	78.8	53.2	64.9	66.7	82.3	66.9
Tiny-V	10M	82.0	63.6	59.8	76.4	53.1	63.3	65.8	85.2	67.3
Tiny-LV	10M	81.6	54.5	57.0	76.5	59.4	66.3	67.2	84.4	67.1
Small	10M	82.0	50.5	58.1	73.2	53.1	60.3	60.7	81.1	64.5
Tiny	100M	82.0	53.5	50.9	79.3	61.3	67.4	67.6	86.6	67.2

Table 3: **GLUE and SuperGLUE results for our various model using the held-out portion of data.** Note that the last two models in the table are baselines, provided for the sake of comparison.

Results For ease of comparison we present our results for initial datasets and the portion held out by the BabyLM challenge organizers separately. The results for each model on the various tasks are displayed in Tables 2 and 3. Table 2 also includes the baseline values provided by the BabyLM organizers using the 10M dataset. Firstly, we note that the standard ELECTRA-Tiny model performs nearly identically with ELECTRA-Small when trained on the 10M word dataset and the T5-base model that was provided as a baseline by the BabyLM organizers. This indicates that larger models are not necessarily superior when using a very small training corpus. The model using embeddings derived from our Lancaster Norm layer showed little difference over the standard ELECTRA-Tiny model. The model using embeddings from the WAC visual layer, however, performed substantially better. In particular, it produced the highest score that we tested on the RTE textual entailment task. It also performed nearly as well on the MNLI tasks as the ELECTRA-Tiny model trained with ten times as much data.

The model using embeddings extracted from the combined layer performed nearly identically to the model containing embeddings from the visual layer on both data splits. This suggests that the model is better able to use the visual information provided via embeddings from the WAC layer of our multiplex than the embeddings extracted from the multiplex's Lancaster Norm layer on natural language understanding tasks. Whether the model benefits more from visual information than sensorimotor information or whether the disparity comes from the nature of our multiplex network can't be determined within the confines of this study. We can, however, definitively say that the visual information in our multiplex embeddings provide a significant boost to model performance in a low data training regime. These results verify prior work (Kennington, 2021), with some differences, that suggest that mapping the visual and sensorimotor sources of information to a network representation, then back to embeddings provides rich and useful information.

Several of the tasks produced identical scores for each model that we evaluated, even the model that was pretrained on the 100M dataset. Each variation yielded a score of 59.9 on BoolQ and 61.4 on WSC using both the initial and held-out datasets. Using the initial dataset each model scored 82.0 on MRPC. Each model also uniformly scored 69.5 on COLA using the held-out data. These results aren't displayed in the tables though their values contribute to the figures listed in the **Avg.** column of each table. Though this is somewhat surprising, we surmise that these scores would vary with additional data and extending training times.

Model	Data	BLiMP	CR_LC	CR_RP	MV_LC	MV_RP	SC_LC	SC_RP
Tiny	10M	60.1	66.2	66.7	66.6	67.0	67.5	66.6
Tiny-L	10M	63.1	66.0	66.6	66.6	66.8	70.6	58.9
Tiny-V	10M	64.1	66.5	66.9	66.6	66.4	67.5	72.0
Tiny-LV	10M	65.1	66.6	66.7	66.6	67.2	68.0	67.6
Small	10M	61.8	66.0	66.7	66.6	66.4	67.4	63.6
Tiny	100M	64.6	66.0	68.2	66.6	67.6	71.3	68.5
OPT-125m	10M	N/A	66.5	67.0	66.5	67.6	80.2	67.5
RoBERTa	10M	N/A	67.7	68.6	66.7	68.6	84.2	65.7
T5-base	10M	N/A	66.7	69.7	66.6	66.9	73.6	67.8

Table 4: **BLiMP and MSGS results for various models.** Note that the last 5 models are baselines included for the sake of comparison. BLiMP scores are note included for baselines provided by the BabyLM organizers as they are not directly comparable to our scores produced through minimal fine-tuning.

5.2 Experiment 2: BLiMP and MSGS Syntactic Tasks

In this experiment we evaluate our models on a set of tasks devoted to testing their grammatical capacity and their inductive biases. Following the BabyLM guidelines, we use the BLiMP dataset to measure the grammatical capacity of our models. The evaluation pipeline for BabyLM treats BLiMP as a zero-shot task using the method of Wang and Cho (2019) or Salazar et al. (2020). Unfortunately, ELECTRA's novel pretraining task is not compatible with either method and produces scores at chance levels for every model variation. In order to make use of BLiMP, and to do so in the closest way possible to the zero-shot paradigm, we create a minimal fine-tuning regime for BLiMP. We treat BLiMP as a binary choice task and train for 1 epoch with only ten percent of each of BLiMP's 67 data subsets in the training split. We use the ADAM optimizer (Kingma and Ba, 2014), with a learning rate of 2e-5 and a batch size of 32. This allows us to obtain consistent results using minimal finetuning. We use the default methods and hyper-parameters provided and finetune for ten epochs with a learning rate of 5e-5 and a batch size of 64. Per BabyLM, we use 5 control tasks and 6 of the ambiguous evaluation tasks. Of the 5 controls, we have two surface features, Lexical Content (LC) and Relative Position (RP), and three linguistic features, Control Raising (CR), Main Verb (MV) and Syntactic Category (SC). The features are combined to form the MSGS tasks in which our models are measured for preference of linguistic features over surface features via Matthews correlation (Matthews, 1975).

Results As our results for BLiMP are not directly comparable to the zero-shot baselines of the BabyLM submissions, we list only the overall average BLiMP score over all 67 data subsets it contains. In the third column of Table 4 we see the BLiMP results for our various models. In each case, the embeddings derived from our multiplex network improved the results over our baseline ELECTRA-Tiny model trained on the 10M dataset. This result is somewhat surprising in that we had not expected concrete sensory information to benefit an abstract task such a grammatical acceptability. Further, we noticed no similar effect relative to the COLA task, the only grammatical acceptability task conducted in the first experiment. That said, we feel confident in claiming that our models derive definite benefit from multi-modal embeddings in a fine-tuning variation of BLiMP.

The results that we obtain for the various MSGS tasks are less definitive. The results for the main task are displayed in Table 4. None of our embeddings seem to have a significant effect, either positive or negative, on model performance for the main MSGS tasks. The only model that we trained that showed a broad increase in performance was the ELECTRA-Tiny model trained with the 100M word dataset. When considered with our other results, this suggests that a model's tendency to adopt favorable inductive biases may primarily be a function of dataset size.

6 Conclusion

In this study, performed in response to the BabyLM challenge, we have shown that small language models can be made more data efficient by enriching their embeddings with sensory information. In particular, the embeddings derived from the Words as Classifiers layer of our multiplex network improve model performance on a variety of tasks from GLUE, SuperGLUE and a version of BLiMP recast as a fine-tuning task. Embeddings derived from Lancaster Sensorimotor Norms likewise pro-

vided useful information for the language models that we evaluated on the BLiMP task, but were less effective on the GLUE and SuperGLUE tasks. Our results from the MSGS evaluations suggest that our models don't gain strong inductive biases toward deep linguistic features as defined by the MSGS task.

Limitations

Our choice to conduct our study on very small models means that our results cannot be assumed to generalize to much larger models. This of course limits the applicability of the findings we have presented. It also stands to reason that multimodal information, like the kind we used to enrich our models, could improve the performance of language models trained on traditional large-scale datasets. Due to the dataset restrictions of the BabyLM challenge, this was also outside the scope of our study and is left to future research.

Acknowledgements

We are grateful to the BabyLM Challenge organizers for making this important challenge happen. Thanks to the anonymous reviewers for their very useful feedback. This material is based upon work supported by the National Science Foundation under Grant No. 2140642.

References

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The amara corpus: Building parallel language resources for the educational domain. In *LREC*, volume 14, pages 1044–1054.

Piotr Bródka, Anna Chmiel, Matteo Magnani, and Giancarlo Ragozini. 2018. Quantifying layer similarity in multiplex networks: a systematic study. *R Soc Open Sci*, 5(8):171747.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Floriana Ciaglia, Massimo Stella, and Casey Kennington. 2023. Investigating preferential acquisition and attachment in early word learning through cognitive, visual and latent multiplex lexical networks. *Physica A: Statistical Mechanics and its Applications*, 612:128468.

Salvatore Citraro, Michael S Vitevitch, Massimo Stella, and Giulio Rossetti. 2023. Feature-rich multiplex lexical networks reveal mental strategies of early language learning. *Sci. Rep.*, 13(1):1474.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

B N C Consortium. 2007. British national corpus. *Oxford Text Archive Core Collection*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Edenbd. 2021. Children stories text corpus.

Clayton Fields and Casey Kennington. 2023. Exploring transformers as compact, data-efficient language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning*. Association for Computational Linguistics.

Martin Gerlach and Francesc Font-Clos. 2018. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics.

Jill Gilkerson, Jeffrey A Richards, Steven F Warren, Judith K Montgomery, Charles R Greenwood, D Kimbrough Oller, John HL Hansen, and Terrance D Paul. 2017. Mapping the early language environment using all-day recordings and automated analysis. *American journal of speech-language pathology*, 26(2):248–265.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. *KDD*, 2016:855–864.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D*, 42(1-3):335–346.

Stevan Harnad. 2017. To cognize is to categorize: Cognition is categorization. In *Handbook of Categorization in Cognitive Science*, pages 21–54.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The goldilocks principle: Reading children's books with explicit memory representations.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models.

Philip A Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. Babyberta: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th conference on computational natural language learning*, pages 624–646.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*.

Casey Kennington. 2021. Enriching language models with visually-grounded word vectors and the Lancaster sensorimotor norms. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 148–157, Online. Association for Computational Linguistics.

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.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

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. *arXiv preprint arXiv:1907.11692*.

Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2020. The lancaster sensorimotor norms: multidimensional measures of perceptual and action strength for 40,000 english words. *Behavior Research Methods*, 52:1271–1291.

Brian MacWhinney. 2000. *The CHILDES project: The database*, volume 2. Psychology Press.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. 2019. Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*.

Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

Vincent Micheli, Martin d'Hoffschmidt, and François Fleuret. 2020. On the importance of pre-training data volume for compact language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7853–7858, Online. Association for Computational Linguistics.

Léo Pio-Lopez, Alberto Valdeolivas, Laurent Tichit, Élisabeth Remy, and Anaïs Baudot. 2021. Multiverse: a multiplex and multiplex-heterogeneous network embedding approach. *Scientific reports*, 11(1):8794.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Linda Smith and Michael Gasser. 2005. The development of embodied cognition: Six lessons from babies. *Artif. Life*, (11):13–29.

Massimo Stella, Nicole M Beckage, and Markus Brede. 2017. Multiplex lexical networks reveal patterns in early word acquisition in children. *Sci. Rep.*, 7:46730.

Massimo Stella, Nicole M Beckage, Markus Brede, and Manlio De Domenico. 2018. Multiplex model of mental lexicon reveals explosive learning in humans. *Scientific Reports, Nature*, 8(1):2259.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. Mobilebert: a compact task-agnostic bert for resource-limited devices. *arXiv preprint arXiv:2004.02984*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Adv. Neural Inf. Process. Syst.*, 30.

Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36, Minneapolis, Minnesota. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

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. *arXiv preprint arXiv:1804.07461*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Gotlieb Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Adina Williams, Bhargavi Paranjabe, Tal Linzen, and Ryan Cotterell. 2023. Findings of the 2023 BabyLM Challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the 2023 BabyLM Challenge*. Association for Computational Linguistics (ACL).

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020a. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Alex Warstadt, Yian Zhang, Haau-Sing Li, Haokun Liu, and Samuel R Bowman. 2020b. Learning which features matter: Roberta acquires a preference for linguistic generalizations (eventually). *arXiv preprint arXiv:2010.05358*.

Wikimedia. simplewiki. https://dumps.wikimedia.org/simplewiki/20230301/. Accessed: 2023-7-31.

Yian Zhang, Alex Warstadt, Haau-Sing Li, and Samuel R Bowman. 2020. When do you need billions of words of pretraining data? *arXiv preprint arXiv:2011.04946*.