Contrastive Visual Semantic Pretraining Magnifies the Semantics of Natural Language Representations

Robert Wolfe
University of Washington
rwolfe3@uw.edu

Aylin Caliskan
University of Washington
aylin@uw.edu

Abstract

We examine the effects of contrastive visual semantic pretraining by comparing the geometry and semantic properties of contextualized English language representations formed by GPT-2 and CLIP, a zero-shot multimodal image classifier which adapts the GPT-2 architecture to encode image captions. We find that contrastive visual semantic pretraining significantly mitigates the anisotropy found in contextualized word embeddings from GPT-2, such that the intra-layer self-similarity (mean pairwise cosine similarity) of CLIP word embeddings is under .25 in all layers, compared to greater than .95 in the top layer of GPT-2. CLIP word embeddings outperform GPT-2 on word-level semantic intrinsic evaluation tasks, and achieve a new corpus-based state of the art for the RG65 evaluation, at .88. CLIP also forms fine-grained semantic representations of sentences, and obtains Spearman's $\rho = .73$ on the SemEval-2017 Semantic Textual Similarity Benchmark with no fine-tuning, compared to no greater than $\rho = .45$ in any layer of GPT-2. Finally, intra-layer self-similarity of CLIP sentence embeddings decreases as the layer index increases, finishing at .25 in the top layer, while the self-similarity of GPT-2 sentence embeddings formed using the EOS token increases layer-over-layer and never falls below .97. Our results indicate that high anisotropy is not an inevitable consequence of contextualization, and that visual semantic pretraining is beneficial not only for ordering visual representations, but also for encoding useful semantic representations of language, both on the word level and the sentence level.

1 Introduction

Large-scale "natural language supervision" using image captions collected from the internet has enabled the first "zero-shot" artificial intelligence (AI) image classifiers, which allow users to create their own image classes using natural language, yet outperform supervised models on common language-and-image tasks (Radford et al., 2021). The image encoders of such models have been shown to form "multimodal" representations in the upper layers, such that the same neurons fire for photographic, symbolic, and textual depictions of a concept (Goh et al., 2021). Research on these state of the art "visual semantic" (joint language-and-image) models has focused primarily on their benefits for encoding semantically legible representations of images. In this paper, we seek to answer a straightforward but as yet unexplored question: what benefits does contrastive visual semantic pretraining have for representations of natural language?

The CLIP ("Contrastive Language Image Pretraining") image classification model introduced by Radford et al. (2021) provides a unique opportunity to observe the effects of visual semantic pretraining on a contextualizing language model. While most other visual semantic architectures combine language and image features in the inner layers of the model (Lu et al., 2019), CLIP separates the language model from the vision model until the end of the encoding process, at which point it projects a representation formed by each model into a joint language-image embedding space (Radford et al., 2021). CLIP is trained to maximize the cosine similarity of a projected image with its projected natural language caption, while minimizing the cosine similarity of the projected caption with all of the other images in the batch (Radford et al., 2021), a training objective known as "contrastive learning" or "contrastive representation distillation" (Tian et al., 2019). The separation of the language model from the vision model prior to projection allows us to consider the two models independently of each other, such that we can study representations of natural language trained for a visual semantic objective, rather than representations which combine language and image features in the inner layers of the model. Moreover, because CLIP encodes natural language using GPT-2, a "causal" language

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 3050 - 3061
May 22-27, 2022 ©2022 Association for Computational Linguistics

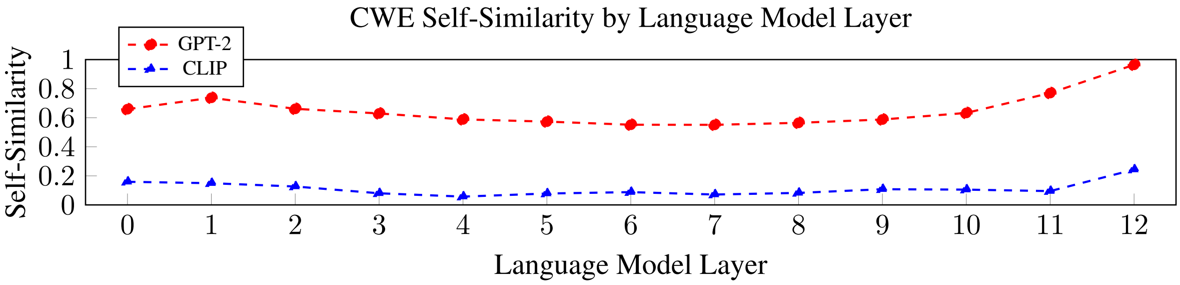

Figure 1: CLIP CWEs are much less self-similar than GPT-2 CWEs, despite being trained with the same architecture, suggesting that pretraining objective leads to significant differences in contextualized representations which are not the result of the contextualization process itself, nor of the model architecture.

model trained solely on next-word prediction, we can directly compare representations formed using the same architecture, but for two very different objectives: one solely linguistic, the other visual semantic.

We observe differences between representations formed by GPT-2 and the CLIP language model ("LM") both on the word level and on the sentence level. We outline our contributions:

1. As shown in Figure 1, contrastive visual semantic pretraining mitigates the angular uniformity (known as anisotropy, measured using cosine similarity) observed by Ethayarajh (2019) in GPT-2 and other contextualizing LMs. The intra-layer self-similarity (mean pairwise cosine similarity, where 1.0 is maximally similar and 0.0 maximally dissimilar) of contextualized word embeddings (CWEs) is less than .25 in all layers of the CLIP LM, compared to greater than .50 in all layers and greater than .95 in the top layer of GPT-2. The five highest-magnitude neuron activations in a CWE from the CLIP LM make up 39% of its length in the top layer, compared to more than 97% of the length of a top layer GPT-2 CWE. This indicates that high anisotropy is not an inescapable consequence of contextualization, nor of using a specific language modeling architecture, but is dependent on pretraining objective, and is significantly reduced by using an objective which is both contrastive and visual semantic.

2. Contrastive visual semantic pretraining results in CWEs which outperform other static and contextualized word embeddings on word-level intrinsic evaluation tasks. CLIP word embeddings obtained in a "decontextualized" setting (wherein the model is given only the word with no other context) set new state of the art for a corpus-based method on the RG65 intrinsic evaluation task (Rubenstein and Goodenough, 1965), with Spearman's $\rho = .88$ in the eighth layer of the CLIP LM, and match state of the art for the ValNorm task, which evaluates the semantic quality of representations based on correspondence with pleasantness norms (Toney and Caliskan, 2021), with Pearson's $\rho = .88$ in layer 4. CLIP CWEs outperform GPT-2 CWEs on every intrinsic evaluation in a decontextualized setting, and for all but one evaluation also outperform the GPT-2 embeddings of Bommasani et al. (2020), who encode $100,000$ contexts and pool over the representations to form a static word embedding matrix.

3. Contrastive visual semantic pretraining encodes semantically useful sentence representations which obtain Spearman's $\rho = .73$ on the SemEval-2017 Semantic Textual Similarity (STS) Benchmark using the cosine similarity between sentence pairs. CLIP results on the STS benchmark outperform those of GPT-2, which never exceed $\rho = .45$ in any layer of the model. Moreover, we find that while GPT-2 sentence embeddings formed using the end-of-sequence (EOS) token exhibit intra-layer self-similarity $\geq .97$ in all layers, the self-similarity of CLIP sentence embeddings steadily decreases over the layers of the model, from .98 to .25 in the top layer, indicating that the contrastive visual semantic pretraining objective of the model forces the formation of fine-grained semantic representations of sentences, such that they can be associated with encoded images.

We make our code and data available at https://github.com/wolferobert3/clip_contrastive_acl_2022.

2 Related Work

We review prior work on visual semantic AI, on the geometry and semantic properties of representations formed by language models, and on semantic intrinsic evaluation tasks.

2.1 Foundation Models

We examine CLIP and GPT-2, both of which are "foundation models," a term coined by Bommasani et al. (2021) to describe the group of architecturally similar state of the art AI systems which have seen wide adoption across domains including language (Raffel et al., 2020), vision (Dosovitskiy et al., 2020), medicine (Rasmy et al., 2021), and programming (Chen et al., 2021), and which exhibit unexpected emergent properties such as strong performance on tasks on which they were not explicitly trained (Brown et al., 2020). GPT-2 and CLIP adapt the transformer neural network architecture, which uses an "attention" mechanism to draw information from the most relevant elements in the model's context window (Vaswani et al., 2017).

2.2 Contextualizing Language Models

GPT-2 is a contextualizing language model, meaning that it forms word representations which incorporate information from surrounding words ("context") (Radford et al., 2019). Such representations, referred to as "contextualized word embeddings" (Peters et al., 2018a), differ depending on the sense of the word used and the specific context in which the word occurs (Soler and Apidianaki, 2021), allowing such representations to overcome many of the limitations of static word embeddings, which use only one vector to represent each word (Collobert et al., 2011). GPT-2 is an autoregressive "causal" language model, meaning that it is trained to predict the next word, and employs "masked self-attention," such that the model can only draw information from words which precede the current word (Radford et al., 2019).

2.3 CLIP and Visual Semantic AI

CLIP is a "multimodal" model which combines language and image representations in a single joint visual semantic embedding space (Radford et al., 2021). CLIP can be used with either a ResNet (He et al., 2016) or a Vision Transformer (ViT) (Dosovitskiy et al., 2020) to encode images, and a language model (GPT-2) to encode captions (Radford et al., 2019). CLIP projects the encoded images and captions into a joint embedding space, where the model maximizes the cosine similarity of the correct image-caption pair while minimizing the cosine similarity of each caption with every other image in the batch (Radford et al., 2021). CLIP projects only a representation of the entire caption into the joint language-image space, and uses CWEs in order to produce this representation.

CLIP is not the first transformer-based model to form visual semantic representations: both Lu et al. (2019) and Li et al. (2019) adapt the BERT language model of Devlin et al. (2019) to produce visual semantic language-image representations, and Zhang et al. (2020) and Jia et al. (2021) use the same contrastive loss objective as CLIP. What makes CLIP unique is that it is the first image classifier to generalize to zero-shot image classification, such that users can define image classes "on-the-fly" using natural language, and obtain performance competitive with supervised computer vision models, without ever fine-tuning on the data for a task (Radford et al., 2021). CLIP improved the zero-shot state-of-the-art[1] on ImageNet (Deng et al., 2009) to 76.2% (Radford et al., 2021), from a previous best of 11.5% (Li et al., 2017).

2.4 Language Model Geometry

Ethayarajh (2019) find that CWEs in ELMo (Peters et al., 2018b), BERT (Devlin et al., 2019), and GPT-2 (Radford et al., 2019) are highly anisotropic (angularly uniform, based on measurements of cosine similarity). The effect is most pronounced in GPT-2, such that randomly selected embeddings in the top layer of the model have "nearly perfect" (*i.e.,* close to 1.0) cosine similarity (Ethayarajh, 2019). Cai et al. (2020) find that the inner layers of GPT and GPT-2 form contextualized word representations on a swiss-roll manifold, while BERT embeds words in clusters. Mitigating anisotropy has been shown to be beneficial for semantic representations, as Mu and Viswanath (2018) find that increasing the isotropy (angular dispersion) of static word embeddings improves performance on semantic intrinsic evaluation tasks. Voita et al. (2019) find that the pretraining objective of a contextualizing lan-

[1] Tiwary (2021) report that their Turing Bletchley model improves the zero-shot state of the art to 79.0%. This model is not available open source to the research community.

guage model affects what information is encoded in CWEs, and that embeddings in causal language models (like GPT-2) contain less mutual information with the input token and more mutual information with the next token in the sequence as the layer index increases. Tenney et al. (2019) shows that layers of BERT are devoted primarily to certain natural language processing (NLP) tasks, and that task complexity increases with the layer index.

2.5 Intrinsic Evaluation Tasks

Intrinsic evaluation tasks assess the quality of word or sentence embeddings by measuring the correlation of the geometric properties of the embeddings with human-rated judgments of similarity (Tsvetkov et al., 2016) or psycholinguistic norms (Toney and Caliskan, 2021). Bommasani et al. (2020) create static word embeddings by pooling over CWEs derived from tens of thousands of sentences from English Wikipedia, and study the performance of these embeddings on word-level intrinsic evaluation tasks. They find that embeddings from the upper layers of BERT and GPT-2 perform poorly relative to embeddings from earlier layers, and that embeddings formed by pooling over a word's CWEs significantly outperform embeddings formed from "decontextualized" words, input to the model with no surrounding context (Bommasani et al., 2020). We report results on the four intrinsic evaluation tasks analyzed by Bommasani et al. (2020), as well as the recently introduced ValNorm task (Toney and Caliskan, 2021), and a sentence-level intrinsic evaluation task, the Semantic Textual Similarity Benchmark (Cer et al., 2017).

3 Data

For comparison of our results on CWE anisotropy with the prior work of Ethayarajh (2019), we encode the text of the SemEval Semantic Textual Similarity tasks from 2012 through 2016 (Agirre et al., 2012, 2013, 2014, 2015), who used these datasets because they include instances of the same words used in different contexts and reflecting different word senses. We discard sentences too long to fit in the 77-token context window of the CLIP LM, which still leaves us with over 36,000 sentences.

3.1 Intrinsic Evaluation Tasks

We report results on five word-level tasks:

- **RG-65** (Rubenstein and Goodenough, 1965), a set of 65 noun pairs assigned scores between 0 and 4 based on their semantic similarity, as judged by 51 human participants in a controlled psychological study intended to evaluate the relationship between "similarity of context and similarity of meaning."

- **WordSim-353**, a word relatedness task consisting of 353 word pairs divided into two sets (Finkelstein et al., 2001). WS-353 was introduced in the context of information retrieval for search engines but is now widespread as an evaluation of word relatedness.

- **SimLex-999**, a word similarity task consisting of 666 noun-noun word pairs, 222 verb-verb word pairs, and 111 adjective-adjective word pairs (Hill et al., 2015).

- **SimVerb-3500**, a set of $3,500$ verb pairs rated on similarity by 843 study participants, and designed to remediate the lack of resources for evaluating verb semantics (Gerz et al., 2016).

- **ValNorm**, which measures the quality of an embedding based on how well it reflects the valence norms of the language on which was trained (Toney and Caliskan, 2021). ValNorm takes Pearson's correlation coefficient of human ratings in a valence lexicon with Single-Category Word Embedding Association Test (SC-WEAT) (Caliskan et al., 2017) pleasantness effect sizes for a word embedding.

Finally, we report results on a sentence-level task, the **Semantic Textual Similarity (STS) Benchmark**, a set of $8,628$ sentence pairs derived from SemEval STS tasks between 2012 and 2017 and rated on similarity (Cer et al., 2017). Sentences reflect three genres: news, forums, and captions. The test set, on which we report results without use of the training set, includes $1,379$ sentence pairs.

3.2 Language Model Architectures

While the CLIP LM is based on the GPT-2 architecture, there are minor differences between the models we examine.[2] The CLIP LM is a 63-million parameter version of the GPT-2 architecture, and uses 12 layers to form 512-dimensional CWEs within a 77-token context window (Radford et al., 2021). GPT-2 Small, the model studied by Ethayarajh (2019) and examined in this paper, forms

[2] We use the PyTorch models available via the Transformers library of Wolf et al. (2020).

768-dimensional CWEs over a 1,024-token context window, and has a total parameter count of 124-million (Radford et al., 2019). Though it consists only of image captions, the text component of the WebImageText corpus used to train CLIP has a "similar" word count to the WebText corpus used to train GPT-2, according to Radford et al. (2021).

4 Approach and Experiments

We outline our experiments, and discuss our approach for extracting both CWEs and sentence embeddings, and for computing self-similarity.

4.1 Geometry of CWEs

We use the self-similarity formula of Ethayarajh (2019) to study whether the contrastive visual semantic pretraining objective of CLIP has affected the anisotropy of GPT-2 CWEs:

$$s = \frac{1}{n^2 - n} \sum_{i} \sum_{j \neq i} cos(\vec{w_i}, \vec{w_j}) \qquad (1)$$

Note that cos in Equation 1 refers to cosine similarity, or the angular similarity of two vectors after normalization to unit length, a common method for measuring the semantic similarity of word embeddings. n refers to the number of word embeddings w used in the self-similarity measurement. Following Guo and Caliskan (2021), who report consistent results on semantic bias analyses by randomly sampling $10,000$ CWEs, we measure the self-similarity of $10,000$ randomly selected CWEs in contexts from the STS 2012-2016 tasks for every layer of CLIP and GPT-2. We collect CWEs for the same $10,000$ word indices from all layers, rather than randomly selecting new words at every layer.

Because Mu and Viswanath (2018) find that a few high-magnitude dimensions cause anisotropy and distort the semantics of static word embeddings, we also examine whether CLIP embeddings encode less of their magnitude in a few high-value dimensions. Mu and Viswanath (2018) find that there are usually $n/100$ such distorting dimensions in static word embeddings, where n refers to the embedding's dimensionality. Because GPT-2 small forms 768-dimensional embeddings, and CLIP forms 512-dimensional embeddings, we report the mean proportion of magnitude contained in the top 8 and the top 5 neuron activations for each model at each layer across $10,000$ embeddings.

4.2 Word-Level Intrinsic Evaluation Tasks

We examine the layerwise performance of CWEs extracted from the CLIP LM and from GPT-2 on the five word-level intrinsic evaluation tasks described in Section 3.1. For these tasks, we extract the vector corresponding to the last subtoken of every word, as prior work finds that the last subtoken in a causal language model fully encodes the semantics of words which a causal language model breaks into subwords (Guo and Caliskan, 2021). For each task, we input words in the "decontextualized" setting described by Bommasani et al. (2020) (*i.e.,* with no surrounding context). Unlike Bommasani et al. (2020), we also extract the BOS token and EOS token from the GPT-2 tokenizer, and add them to either side of the decontextualized word. We do this to keep the experiment consistent between the models, as adding the tokens is default behavior for the CLIP LM, but not for GPT-2. Because it is common to omit the BOS and EOS tokens when using GPT-2, we report results for GPT-2 both with the tokens and without them. To observe whether CLIP sentence embeddings have unique properties, since they are the only linguistic representations projected to the joint language-image space, we also report results on these tasks using the EOS token for the CLIP LM and GPT-2.

4.3 Sentence-Level Evaluations

We report layerwise performance using sentence representations obtained from CLIP and GPT-2 on the STS benchmark (Cer et al., 2017). For this task, we use the EOS token in both CLIP and in GPT-2. For GPT-2, we also use the last subtoken of the sentence, with no EOS token added.

Finally, we analyze the self-similarity of sentence embeddings from each model using Equation 1. In this case, w refers not to a word embedding, but to a sentence embedding. For this analysis, we use embeddings of all of the unique sentences in the test set of STS Benchmark (Cer et al., 2017).

5 Results

CLIP CWEs are less anisotropic than GPT-2 embeddings, and CLIP outperforms GPT-2 on word-level and sentence-level semantic evaluations.

5.1 Embedding Geometry

As illustrated in Figure 1, the self-similarity of CWEs is lower in every layer of the CLIP LM than in GPT-2. Self-similarity in both models is at its

Performance by Intrinsic Evaluation Task										
Task	RG65		WS-353		SL-999		ValNorm		SV-3500	
Layer	Best	Top	Best	Top	Best	Top	Best	Top	Best	Top
GPT-2 - no BOS	.09 (1)	.01	.14 (1)	.12	.05 (5)	.02	.43 (7)	.25	.01 (8)	.00
GPT-2 - w/ BOS	.44 (7)	.23	.44 (9)	.25	.25 (8)	.11	.76 (7)	.33	.21 (8)	.07
CLIP	**.88 (8)**	.70	**.72 (6)**	**.51**	**.48 (9)**	**.39**	**.88 (4)**	.72	**.30 (4)**	**.17**
GPT-2 EOS	.32 (12)	.32	.31 (3)	.10	.16 (4)	.05	.61 (6)	.17	.10 (4)	-.01
CLIP EOS	.73 (12)	**.73**	.49 (5)	.45	.34 (11)	.34	.84 (5)	**.80**	.14 (11)	.13

Table 1: CLIP CWEs outperform GPT-2 CWEs on every intrinsic evaluation task examined. The "EOS" token corresponds to the model's sentence embedding. The best layer corresponds to the layer which a representation achieves the highest score for a task. All scores are Spearman's ρ, except for ValNorm, which uses Pearson's ρ.

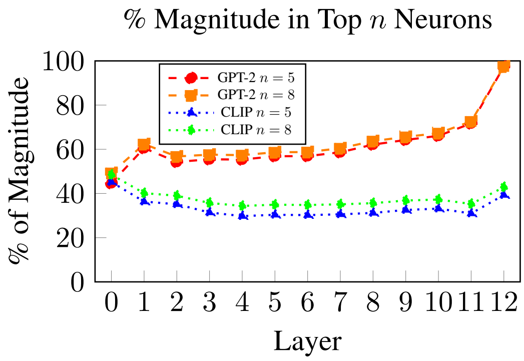

Figure 2: The five highest-magnitude neuron activations make up more than 97% of the length of GPT-2 CWEs, compared to 39% of the length of CLIP CWEs.

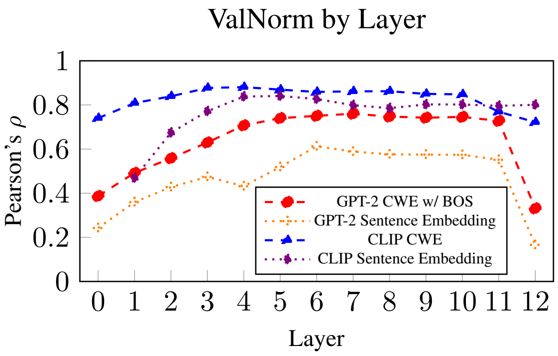

Figure 3: CLIP CWEs match the state of the art on the ValNorm intrinsic evaluation task in layer 4.

highest in the top layer, at .96 in GPT-2 and .24 in the CLIP LM. The self-similarity of CWEs in GPT-2 never falls below .55 in any layer, whereas the self-similarity of CWEs in CLIP falls to .06 in layer 4. As shown in Figure 2, we also find that the five highest-magnitude neuron activations in the top layer of GPT-2 make up more than 97% of the magnitude of GPT-2 CWEs, compared to only 39% of the magnitude of CLIP CWEs. For both models, there is a small increase (less than 3 percentage points in each layer) using the 8 highest neuron activations. Given that Mu and Viswanath (2018) found that high-magnitude dimensions cause high anisotropy and distort semantics in static word embeddings, and that Ethayarajh (2019) suggests increasing isotropy to improve CWE representational quality, we would expect that CLIP CWEs would have more semantic geometry than GPT-2 CWEs.

5.2 Word-Level Intrinsic Evaluation Tasks

As shown in Table 1, CLIP embeddings outperform GPT-2 embeddings on all five of the word-level intrinsic evaluation tasks we study, and non-trivially improve the corpus-based state of the art for the RG65 intrinsic evaluation to Spearman's $\rho = .88$.[3] As visualized in Figure 3, CLIP embeddings also match the state of the art for the ValNorm intrinsic evaluation task (Toney and Caliskan, 2021), previously achieved by the GloVe embeddings of Pennington et al. (2014). For every task except SV-3500, CLIP embeddings outperform the results obtained for GPT-2 by Bommasani et al. (2020), who create static word embeddings by pooling over CWEs obtained from $100,000$ encoded contexts, both in GPT-2 small and in GPT-2 medium, a 24-layer model which forms $1,024$-dimensional embeddings. For SV-3500, Bommasani et al. (2020) obtain Spearman's $\rho = .31$ in layer 6 of GPT-2 small from embeddings formed using CWEs $100,000$ from contexts.

Our results also indicate that adding the BOS token in GPT-2 significantly improves results on word-level semantic intrinsic evaluation tasks in the decontextualized setting. ValNorm scores im-

[3] According to the ACL leaderboard at https://aclweb.org/aclwiki/RG-65_Test_Collection_(State_of_the_art). Precisely, CLIP embeddings achieve Spearman's $\rho = .876$ on this task.

prove from .59 to .76 in layer 7, and RG65 scores improve from .01 to .44 in the same layer. On every test, simply adding the BOS token outperforms results reported by Bommasani et al. (2020) on embeddings obtained using the pooling methodology for $10,000$ contexts, both in GPT-2 small and GPT-2 medium Bommasani et al. (2020). While adding the BOS token does not match the results of applying the pooling method to 50,000 or 100,000 contexts, this marked improvement indicates that using the BOS token is a simple, computationally efficient, and easily replicated way of obtaining static reductions of CWEs, with better quality than representations requiring ten thousand contexts to form.

Finally, we find that CLIP EOS token embeddings outperform CWEs in the top layer on two of five word-level intrinsic evaluation tasks, and nearly equal the performance of CLIP CWEs on the other three tasks. ValNorm scores fall to .72 for CLIP CWEs in the top layer, but increase to .80 for CLIP EOS token embeddings in that layer; and RG65 scores fall to .70 in the top layer for CLIP CWEs, but increase to .73 for CLIP EOS token embeddings. CWEs lose some of their mutual information with the input word as the model forms predictions about the next word in the sequence (Voita et al., 2019), but our findings indicate that the EOS token must maintain the semantic information of a context in the top layers, such that it can be projected to the joint language-image space and accurately associated with an image.

Additional visualizations of CLIP and GPT-2 performance on word-level intrinsic evaluation tasks are included in Appendix A.

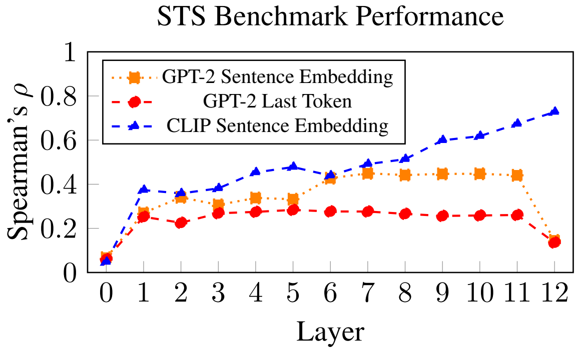

Figure 4: CLIP sentence embeddings outperform GPT-2 embeddings on the STS Benchmark.

5.3 Sentence Embeddings

As shown in Figure 4, sentence embeddings from the CLIP LM outperform GPT-2 sentence embeddings on the STS benchmark at every layer of the respective models, and the difference in performance grows in the upper layers. CLIP sentence embeddings obtain Spearman's $\rho = .73$ in the top layer, compared to no greater than .45 for GPT-2 embeddings. Even using the EOS token, GPT-2 sentence embeddings exhibit properties similar to CWEs in the model, and lose semantic information in the upper layers, while CLIP sentence embeddings improve in semantic quality through the top layer.

As shown in Figure 5, CLIP sentence embeddings become increasingly dissimilar as the layer index increases. This is in stark contrast to GPT-2, wherein sentence embeddings using the EOS token have self-similarity $\geq .97$ in every layer, and indicates that the contrastive visual semantic objective of CLIP forces fine-grained differentiation of sentence-level semantics.

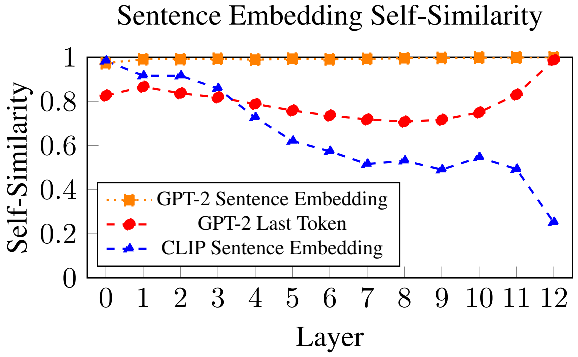

Figure 5: CLIP sentence embeddings become less self-similar as the layer index increases, while GPT-2 sentence embeddings remain highly anisotropic.

6 Discussion

Our findings are straightforward, but it is not obvious that they should occur. The training objective of CLIP is not to produce high-quality CWEs, or even sentence embeddings. Indeed, Radford et al. (2021) spend little time discussing the CLIP language model, noting that they did not see significant performance improvements by scaling up the size of the model. However, in creating the first broadly accurate zero-shot image classifier, Radford et al. (2021) have also created a zero-shot sentence encoder which substantially outperforms the version of its underlying architecture trained on language modeling. Moreover, without the need for computationally expensive pooling methodologies, and despite having less than half the parameter count of GPT-2 small, the CLIP LM produces CWEs which match or exceed the best performance ever realized with a corpus-based approach on two intrinsic evaluation tasks, and outperform embeddings formed from $100,000$ encoded contexts in GPT-2 medium (Bommasani et al., 2020).

CLIP embeddings show that the high anisotropy observed by Ethayarajh (2019) is not the inevitable result of contextualization, nor even of a specific language modeling architecture, but is connected to the pretraining objective of the model. When trained on a contrastive visual semantic objective, CWEs formed by CLIP have much lower self-similarity at every layer of the model in comparison with GPT-2. This is remarkable because CLIP does not actually project CWEs into the joint language-image space. While we might expect CLIP sentence embeddings, which are projected into the language-image space, to have different properties from the CWEs formed by GPT-2, it does not necessarily also follow that the CWEs formed by CLIP would also be so different from those in GPT-2. Indeed, we still observe the increased self-similarity in the top layer reported by Ethayarajh (2019), and the loss of semantic information related to the input token in the upper layers, as reported by Voita et al. (2019). However, these effects are much less pronounced in CLIP than they are in GPT-2, indicating that the contrastive visual semantic objective of the model has regularizing effects that shape more than just the projected sentence embedding.

Our findings suggest that language models trained on visual semantic objectives are likely to privilege the encoding of semantic information, which is essential to matching a caption to an image. The more isotropic representations we observe reflect the objective of the model, which requires differentiating fine-grained semantic information. That models trained on visual semantic objectives would form embeddings to reflect the semantics of a word or sentence more than would a causal language model makes intuitive sense. Through the lens of the training objective, it is more problematic for a causal language model to predict a syntactically invalid continuation of a sentence, such as an incorrect part of speech, than to predict a somewhat unexpected but still syntactically valid continuation of a sentence. When a language model is trained to encode and associate the correct text caption with a matching image, however, the semantic content of the text becomes at least as important as its syntactic properties.

6.1 Limitations and Future Work

Our work shows that a pretraining objective which is both visual semantic and contrastive in nature results in isotropic, highly semantic CWEs and sentence representations, in stark contrast to the representations formed by the same architecture when trained on a language modeling objective. However, further work is needed to address to what extent the results we observe are the result of contrastive training, and to what extent they are the result of visual semantic training. It is possible that a contrastive training objective, wherein the model must discriminate between correct and incorrect options, will result in isotropic and highly semantic embeddings even if both models produce linguistic representations. On the other hand, encoding language for the purpose of performing visual semantic tasks may be particularly important for achieving the effects seen in CLIP, as images lack a grammatical structure and are primarily semantic in composition. Future work might perform a direct assessment between representations obtained from the CLIP LM and representations learned by contrastive text-only models such as those recently introduced by Neelakantan et al. (2022).

This work examines semantics in contextualized representations without postprocessing, using cosine similarity as the similarity metric. While this is a common experimental design evaluated frequently in prior work, it is not the only way of assessing semantics in contextualized word embeddings. For example, recent work indicates that semantics can be better isolated in language models like GPT-2 by postprocessing and transforming the embedding space using methods such as removing high-magnitude directions with principal component analysis (Wolfe and Caliskan, 2022; Timkey and van Schijndel, 2021).[4] Future work might assess whether these postprocessing techniques, or methods which assess semantics using mutual information (Voita et al., 2019) or linear probes (Tenney et al., 2019), also indicate that contrastive multimodal pretraining magnifies semantics in the embedding space.

[4] CLIP still outperforms GPT-2 in nearly every case over intrinsic evaluation results reported after postprocessing, and CLIP embeddings may also exhibit improvements from comparable manipulations of the embedding space.

Finally, Radford et al. (2021) note that CLIP was first intended to be a zero-shot caption generator, a design which has since been realized using the SimVLM architecture of (Wang et al., 2021b). Analysis of such models, which are not yet available to the research community in a way which would allow analysis of the underlying architecture, may help to answer questions of whether the contrastive objective or the visual semantic setting is more important for regularizing anisotropy and representing semantics.

7 Conclusion

We find that contrastive visual semantic pretraining produces isotropic CWEs which outperform a language model based on the same architecture on semantic evaluations on both the word level and the sentence level. Our findings indicate that incorporating visual semantic objectives with language models may be useful both to regularize the anisotropy in CWEs and to improve the semantic quality of both word and sentence representations.

8 Ethical Considerations

While the contrastive visual semantic objective of CLIP produces semantically rich representations of natural language, we caution that the model is also known to encode harmful societal biases. Goh et al. (2021) find that the CLIP image encoder forms representations which reflect biases against communities marginalized based on religion and on immigration status, and Wang et al. (2021a) and Agarwal et al. (2021) report biases of underrepresentation and stereotypical associations which disproportionately affect women. Moreover, Radford et al. (2021) state that they use frequency-based heuristics to construct the WebImageText corpus on which CLIP trains. Other research on language models has shown that similar techniques can exacerbate biases against marginalized groups, who are often underrepresented in such datasets (Wolfe and Caliskan, 2021). Thus, while our findings are promising for the future of visual semantic AI systems, models like CLIP must be studied further to understand how they represent people, and what the ramifications of such representations are for society.

Acknowledgements

This material is based on research partially supported by the U.S. National Institute of Standards and Technology (NIST) Grant 60NANB20D212. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of NIST.

References

Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. 2021. Evaluating clip: Towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, et al. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 252–263.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 81–91.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. * sem 2013 shared task: Semantic textual similarity. In *Second joint conference on lexical and computational semantics (* SEM), volume 1: proceedings of the Main conference and the shared task: semantic textual similarity*, pages 32–43.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting pretrained contextualized representations via reductions to static embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2020. Isotropy in the contextual embedding space: Clusters and manifolds. In *International Conference on Learning Representations*.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde, Jared Kaplan, Harri Edwards, Yura Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv e-prints*, pages arXiv–2107.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414.

Daniela Gerz, Ivan Vulic, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. Simverb-3500: A large-scale evaluation set of verb similarity. In *EMNLP*.

Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. 2021. Multimodal neurons in artificial neural networks. *Distill*, 6(3):e30.

Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 122–133.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

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. *arXiv e-prints*, pages arXiv–2102.

Ang Li, Allan Jabri, Armand Joulin, and Laurens van der Maaten. 2017. Learning visual n-grams from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4183–4192.

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 preprint arXiv:1908.03557*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 13–23.

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. 2022. Text and code embeddings by contrastive pretraining. *arXiv preprint arXiv:2201.10005*.

Jeffrey Pennington, Richard Socher, and Christopher D 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.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. Deep contextualized word representations. *CoRR*, abs/1802.05365.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018b. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):1–13.

Herbert Rubenstein and John B Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Aina Gari Soler and Marianna Apidianaki. 2021. Let's play mono-poly: Bert can reveal words' polysemy level and partitionability into senses. *Transactions of the Association for Computational Linguistics (TACL)*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2019. Contrastive representation distillation. In *International Conference on Learning Representations*.

William Timkey and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546.

Saurabh Tiwary. 2021. Turing bletchley: A universal image language representation model by microsoft.

Autumn Toney and Aylin Caliskan. 2021. Valnorm quantifies semantics to reveal consistent valence biases across languages and over centuries. *Empirical Methods in Natural Language Processing (EMNLP)*.

Yulia Tsvetkov, Manaal Faruqui, and Chris Dyer. 2016. Correlation-based intrinsic evaluation of word vector representations. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 111–115.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4396–4406, Hong Kong, China. Association for Computational Linguistics.

Jialu Wang, Yang Liu, and Xin Eric Wang. 2021a. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. *arXiv preprint arXiv:2109.05433*.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021b. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Robert Wolfe and Aylin Caliskan. 2021. Low frequency names exhibit bias and overfitting in contextualizing language models. *arXiv preprint arXiv:2110.00672*.

Robert Wolfe and Aylin Caliskan. 2022. Vast: The valence-assessing semantics test for contextualizing language models. In *Thirty-Sixth AAAI Conference on Artificial Intelligence*.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. 2020. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*.

A Intrinsic Evaluation Performance

We include visualizations showing the performance of CLIP and GPT-2 embeddings on the intrinsic evaluation tasks discussed in the paper.

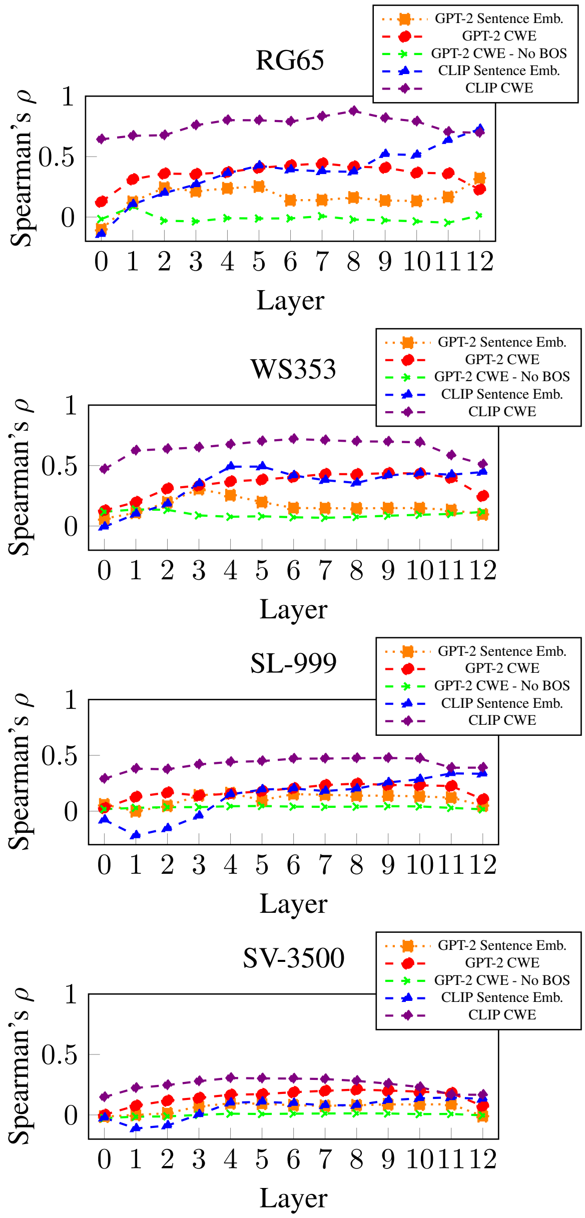

Figure 6: CLIP CWEs outperform other representations in almost every layer across four intrinsic evaluations, including achieving corpus-based state of the art on RG65 in layer 8, with Spearman's $\rho = .876$..