Providing Evidence that GPT2 Embeddings Encode Affect Using a Control Task.

Lin Khern Avery Chia
Department of Psychology
University of Illinois at Urbana-Champaign

Abstract

Natural Language Processing (NLP) applications and shared tasks such as SemEval are often interested in classifying the affect of text. A common solution is using Language Model (LM) embeddings as features in a classifier trained to predict Ekman's basic emotions (Ekman, 1992). This solution assumes that the LM has learned information useful to the classification task through its training objective. Unfortunately, this assumption has been challenged by the probing research community (Belinkov, 2022) and is untenable until tested. This paper applies Hewitt & Liangs' (2019) control task to test if GPT2-large embeddings contain usable affective information and finds evidence that it does.

1 Introduction

In shared tasks such as SemEval, neural network Language Models (LMs) are often used in conjunction with a model of emotion (e.g. Ekman, 1992). The logic of the approach is that the LMs have learned representations that encode emotional information, and that either fine-tuning or extracting features from the LMs will help in the classification task. Approaches of this sort involve training an additional classifier (these can be thought of as probe classifiers) on top of the LM. There are two assumptions in this approach:

A. Neural network LMs learn information pertaining to Ekman's basic emotions when trained on their language modeling task.

B. Ekman's basic emotions is a valid emotional model that reflects what is expressed by human communication in text.

2.1 Evidence from N400 Research

Assumption A has corroborating evidence from the Event-Related Potential literature in semantic memory and cognitive science. In particular, N400 research. Event-Related Potentials are brains waves derived from the scalp that are time-locked and linked to the occurrence of a particular event. These events are usually stimuli that are presented to the participants from which these ERPs are computed.

The N400 is a negative-going Event-Related Potential (ERP) that peaks at about 400ms upon the onset of a stimulus. While there appear to be several neural generators responsible for it, its latency is remarkably stable and it is identified instead by its functional sensitivities (Kutas & Federmeier, 2011). The waveform has been found to be sensitive to a suite of different variables: word frequency (Kutas & Federmeier, 2000) and repetition (Deacon et al., 2004), orthographic and lexical neighborhood (Laszlo & Federmeier, 2011), word- and sentence- contexts (Kutas, 1993; Lau et al., 2009), and discourse and pragmatic contexts (van Berkum, 2009). The N400 can also be elicited by many non-lexical stimuli, including nonwords (Deacon et al., 2004) and gestures (van Berkum, 2009). Due to these functional sensitivities, the current dominant theory of the N400 is that it reflects the access of semantic representations in the distributed network that is a human brain (Kutas &

Federmeier, 2011). The more negative-going the waveform is, the more semantic features needed to be brought online to support the given stimulus. This theory gels well with distributed feature theories of semantic memory (e.g. Osgood, 1952) and thus finds a natural fit with distributional models of language (a class that includes Transformer LMs) upon contact with the Distributional Hypothesis (Harris, 1954; Lenci, 2018).

There is work that suggests that GPT2-xl (extra large) embeddings encode human semantics rather well. Schrimpf et al. (2021) showed that GPT-2xl provided the best fits (out of several Transformer models) to several datasets of passive reading: 1) behavioral reading times, 2) electrocorticography, and 3) fMRI. Furthermore, the N400 computational modeling literature is moving in the direction of using GPT-2 to evaluate theories of the ERP. Lindborg & Rabovsky (2021) showed that metrics computed on the embeddings of GPT2-large are able to model N400 phenomena both qualitatively and quantitatively. Using model comparisons of generalized linear models, they find that metrics computed from layers 21-25 of GPT2-large provide the most predictive power of human N400s.

Arguably, the affect expressed by a sentence is part of the meaning/semantics of the sentence. If so, whether GPT2 encodes affective information is a natural question to ask.

2.2 Assumption A May Be Inappropriate

Unfortunately, due to the power of non-linear neural network models, assumption A in affective classification work is not tenable without testing. Probing classifiers are often used on LMs in service of application and research, with the promise of using/revealing the information learned by the LM in its training objective (Belinkov, 2022). Oftentimes, this usage involves training classifiers on LM embeddings in service of a different task. The logic of this is to see if the LM, in its language modeling objective, has learned information useful to this new task that is extractible. However, there is a concern that instead of learning to make use of linguistic information in the embeddings, probe classifiers learn arbitrary structure in the high-dimensional embeddings space that is useful (Belinkov, 2022). This concern is agnostic and can be either a concern that: 1) the LM did not learn the appropriate linguistic information, and 2) the LM has learned and encoded the appropriate information, but said information is hard to extract and unavailable.

Fortunately, clever controls have been designed to allay this concern. Hewitt & Liang (2019) proposed a control task, where a control classifier is trained and tested on a dataset with shuffled labels. Critically, the shuffling of these labels happen on the type- (not token-) level. This means that their control task has arbitrary structure that can be learned and generalized by a neural network classifier, and provides a good control to a classifier that has learned and generalized a (presumably) linguistic structure. This is different from simply shuffling the labels on the token-level, which will not necessarily provide any structure for the network to learn. By comparing the two classifiers on their performance, we can then talk about the selectivity of the representations: whether the representations encode linguistic information valuable to the task, or provide only arbitrary structure that has been taken advantage of (or structure that provides on-par performance with random, arbitrary structure). If the control classifier performs on par with the classifier, the latter is said to have occurred. If the classifier out-performs the control classifier, then we can say that there is linguistic structure that is readily extractible and useful for the task.

A positive evaluation of assumption A may suggest something interesting about assumption B. While widely used, Ekman's emotions have been criticized as a potentially invalid model of human emotions (Barrett et al., 2019). Finding that LM embeddings encode information useful for classifying Ekman's basic emotions may suggest that the emotional model accounts for some valuable emotional variance in produced language. On the other hand, a negative evaluation of assumption A may provide evidence to these criticisms.

3 Problem Definition

This project proposes to evaluate the question of whether a LM like GPT2 has truly learned structure that is appropriate for an affective classification task. This evaluation will be done using neural network classifiers trained on GPT2 embeddings of sentences to perform either: 1) emotion classification of Ekman's basic emotions, or 2) a control task as defined by Hewitt

& Liang (2019). Both classifier types will have GPT2 embeddings of size 1280 as input and a multi-label vector representing the possible Ekman emotions as output. A comparison of performance will be made between the classifier and the control classifier in order to determine if GPT2 embeddings have information specific to affective categorization.

4 Approach

4.1 Data

The labeled data on which the classifiers will be fit come from SemEval-2018, subtask 5 (Mohammad et al., 2018). In this subtask, teams had to build affective classifiers that classified the Ekman emotion present in tweets. This was a multi-label classification problem where each tweet could belong to one or more of the 11 emotion labels.

4.2 GPT2 Embeddings

GPT2-large and GPT2-xl are very large LMs. To narrow the scope of inquiry and make this project manageable and appropriate for an end-of-term paper, I will be using embeddings from layers 21 to 25 of GPT2-large. This is because Lindborg & Rabovsky (2021) have found that representations from these layers, out of all the GPT2-large layers, model the N400 the best. This presumably means that I stand the best chance of finding selective GPT2-large representations there. The GPT2-large model used in this project is the implementation by huggingface in PyTorch.

The procedure will be to fit two classifiers (one affective and one control) per layer of GPT2-large embeddings and compare their performance.

4.3 Building the Classifiers

Classifiers are fully-connected, multi-layer perceptron classifiers with at least one hidden layer and a sigmoidal output vector of size 11 (11 Ekman emotions). Non-linearity is introduced using the Rectified Linear Unit (ReLU) activation function in the hidden layers. Classifiers are implemented in PyTorch and hyperparameter tuned using Ray Tune, an industry-standard tool for distributed hyperparameter tuning.

Classifiers for each layer were tuned according to the following possible configurations in Table 1 and were selected based on fit to the validation set, evaluated using Binary Cross Entropy loss.

Number of Hidden Layers	1, 2
Number of Hidden Units	200, 500, 700, 1000
Learning Rate	loguniform(1e-4, 1e-1)
Batch Size	2, 4, 8, 16, 32

Table 1: Hyperparameter search space.

The decision to search 1-2 hidden layers was made to adhere to Hewitt & Liang (2019), who cited concerns of using probes that were too prone to learning arbitrary structure if they had ventured beyond two hidden layers. The decision to search the rest of the hyperparameter space was arbitrary and purely exploratory.

Control classifiers were trained using the same hyperparameter settings as their affective counterparts, but on shuffled labels.

Refer to Table 2 for the optimized hyperparamter settings of classifiers trained on GPT2 embeddings from layers 21 to 25.

	n_hidden	n_units	lr	batch_size
21	1	700	0.0005	8
22	2	700	0.002	16
23	2	1000	0.00016	2
24	1	1000	0.00068	8
25	2	500	0.00038	8

Table 2: Chosen hyperparameters per GPT2-large layer. Rows represent GPT2-large layers that classifiers were trained on. Columns represent hyperparameters.

4.4 Control Tasks

Control tasks were generated as per Hewitt & Liang (2019), where labels of subtask 5 were shuffled on the type level. This meant that every token of a particular type would be mapped to the label associated with that type. Doing so implements representations with arbitrary structure, which provides a good control to representations with presumably linguistic structure. The generation of my control tasks was done using a custom python object called ControlTaskGenerator. The object tokenizes a sequence of text using the GPT2-large tokenizer and assigns a randomized label for each type in the text. Then, using its transform() method, it assigns each tweet the label that has been mapped to its final token.

Each type was randomly assigned to have between 1 to 3 emotion labels to simulate the original dataset and cater to produce a multi-label classification task.

4.5 Computing Performance and Evaluating Selectivity

The metric of performance chosen was the micro-averaged multi-class F1 to allow for direct comparison with team task performance in Mohammad et al., (2018). Hewitt & Liangs' (2019) selectivity was then computed by subtracting the micro-F1 scores of the control classifier from that of the affective classifier. The more positive this resulting number, the more affective structure is present in GPT2-large embeddings over-and-above random structure.

5 Results

	Affective F1	Control F1	Selectivity
21	0.895	0.723	0.172
22	0.897	0.719	0.178
23	0.891	0.716	0.176
24	0.897	0.717	0.180
25	0.881	0.722	0.160

Table 3: Performance of affective and control classifiers per GPT2-large embedding layer, and their associated selectivity scores. All F1 metrics reported in this table are micro-averaged. Rows represent GPT2-large layers that classifiers were trained on.

5.1 Affective Classification Performance

Affective classifiers across all GPT2-large layers 21-25 performed better than the best system in subtask 5 of SemEval-2018 (Mohammad et al., 2018). They averaged a micro-F1 of 0.89 (compared to 0.70 of the best-performing system in the competition) (Table 3). The best-performing classifier was fit on embeddings from layer 24 and had a micro-F1 of 0.897. However, this is neither surprising nor groundbreaking given that Transformer LMs are far more powerful LMs than those that were publicly available in the 2018 competition, and gained popularly precisely because they broke ceilings in state-of-the-art performance in various linguistic and psycholinguistic tasks.

5.2 Selectivity

Control classifiers across all GPT2-large layers 21-25 yielded micro-F1 scores of about 0.72. Therefore, selectivity (affective F1 - control F1) is about 0.17 (Table 3). Selectivity seems to drop off in layer 25 (selectivity=0.159), with the highest value found in layer 24 (selectivity = 0.18).

6 Discussion and Conclusions

Language Models (LMs) feature heavily in applications, research, and computational linguistics competitions, especially in those domains that require a system to assign a string of text into certain linguistic or psychological categories. Classifiers are often trained to associate these LM embeddings with the appropriate labels. In the affective computing realm, these classifiers are often trained to predict Ekman's basic emotions using some form of LM representation. A key assumption to this approach is that the LMs have learned representations that pertain to Ekman's basic emotions when trained on their language modeling task. While there is corroborating evidence from cognitive science literature that Transformer LM embeddings encode semantic information, the assumption that these embeddings contain such information that is usable by a classifier remains untenable until tested. This concern stems from research that suggests that multi-layered neural network classifiers are powerful enough to learn and generalize arbitrary structure (Belinkov, 2022) and calls for appropriate analyses to ensure that past performance on the affective categorization task is due to the value of the embeddings and not due to the probe's power. Hewitt & Liang (2019) provides an appropriate experimental control where control classifiers are fit on control tasks with arbitrary structure. Selectivity is computed by subtracting control classifier performance from the linguistic classifier performance. The more positive the metric, the more affective structure is present in the LM embeddings over-and-above random structure.

My investigations have found evidence that GPT2-large embeddings are selective to an affective categorization task with Ekman's basic emotions. My affective classifiers that were trained on an affective categorization task consistently out-performed my control classifiers that were trained on a control task. There was variance in the selectivity of embeddings from the various GPT2-large layers I extracted from, with the lowest amount of selectivity coming from layer 25. This seems to be due to a combination of increased control classifier performance and decreased affective classifier performance.

6.1 Limitations

Unfortunately, I cannot provide a more satisfying explanation for the pattern in layers 21-25 without more research. While it may be tempting to conclude from the analyses that later layers see a decrease in selectivity in general, we do not know for sure because I only tested 5 out of the 36 GPT2-large layers. Furthermore, any more interesting explanation will have to wait until a better hyperparameter search. My search space may have been insufficient, and a more thorough one may be able to find better affective categorization models. For example, I did not choose to tune for regularization, shape of the network, loss function, or activation function of the hidden layer, all of which may affect the probes' ability to learn generalizable structure.

Another limitation is the control task I used. I made the decision to simulate the 1x11 label vectors of the original task instead of plainly using the unique instances of those vectors. While unlikely, this may have produced labels that were computationally easier than the original labels, and made the control task easier than it should have been. Future work should address this potential issue by using unique instances of original label vectors instead of randomly generating them. If future work were to continue with randomly generating and simulating these multi-class labels, more care can be taken to ensure that the distribution of these labels mirror the actual data well.

Lastly, while Schrimpf et al., (2021) showed that GPT2-xl was a better model of human semantic behavior than GPT2-large, I was restricted to justifying and using GPT2-large due to computational concerns. It would be interesting to evaluate selectivity in GPT2-xl. Perhaps its embeddings are more selective, which would corroborate the findings of Schrimpf et al., (2021).

6.2 Conclusion

My work in this paper provides evidence that GPT2-large embeddings contain affective structure that is useful and extractible by simple multi-layer perceptron classifiers of Ekman's basic emotions. This corroborates with evidence from the ERP computational modeling literature that has found that GPT2 embeddings are useful for modeling semantic behaviors, covert and overt.

Not only does this work provide evidence that validates affective classifiers built on GPT2 embeddings, it suggests that Ekman's basic emotions do capture some kind of emotional variance in produced language. More importantly, this work suggests that we are on the right track using GPT2 (and perhaps other Transformer LMs) to model human semantics.

References

Belinkov, Y. (2022). Probing classifiers: Promises, shortcomings, and advances. Computational Linguistics, 48(1), 207-219.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. Psychological science in the public interest, 20(1), 1-68.

Deacon D, Dynowska A, Ritter W, Grose-Fifer J. Repetition and semantic priming of nonwords: Implications for theories of N400 and word recognition. Psychophysiology. 2004;41:60–74.

Ekman, P. (1992). Are there basic emotions?.

Harris ZS. 1954. Distributional structure. Word 10:146–62

Hewitt, J., & Liang, P. (2019). Designing and Interpreting Probes with Control Tasks. Proceedings of the 2019 Conference of Empirical Methods in Natural Language Processing.

Kutas M. In the company of other words: Electrophysiological evidence for single-word and sentence context effects. Lang Cogn Process. 1993;8:533–72. [Google Scholar]

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. Trends in cognitive sciences, 4(12), 463-470.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP). Annual review of psychology, 62, 621.

Laszlo, S., & Federmeier, K. D. (2011). The N400 as a snapshot of interactive processing: Evidence from regression analyses of orthographic neighbor and lexical associate effects. Psychophysiology, 48(2), 176-186.

Lau E, Almeida D, Hines PC, Poeppel D. A lexical basis for N400 context effects: evidence from MEG. Brain Lang. 2009;111:161–72.

Lenci, A. (2018). Distributional models of word meaning. Annual review of Linguistics, 4, 151-171.

Lindborg, A., & Rabovsky, M. (2021). Meaning in brains and machines: Internal activation update in large-scale language model partially reflects the N400 brain potential. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 43, No. 43).

Mohammad, S., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S. (2018, June). Semeval-2018 task 1: Affect in tweets. In Proceedings of the 12th international workshop on semantic evaluation (pp. 1-17).

Osgood, C. E. (1952). The nature and measurement of meaning. Psychological bulletin, 49(3), 197.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., ... & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. Proceedings of the National Academy of Sciences, 118(45), e2105646118.

Van Berkum JJA. The neuropragmatics of 'simple' utterance comprehension: An ERP review. In: Sauerland U, Yatsushiro K, editors. Semantics and Pragmatics: From Experiment to Theory. Basingstoke: Palgrave Macmillan; 2009. pp. 276–316.