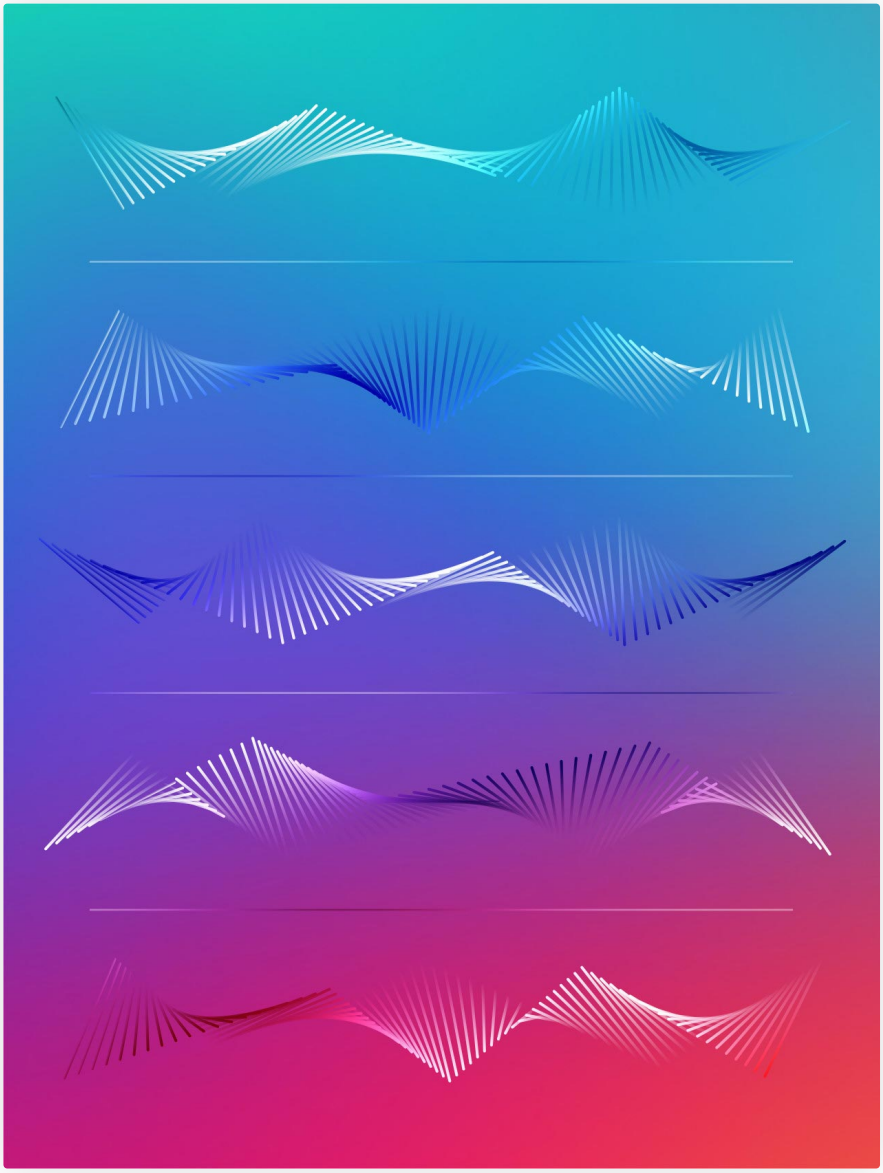


Unsupervised Sentiment Neuron

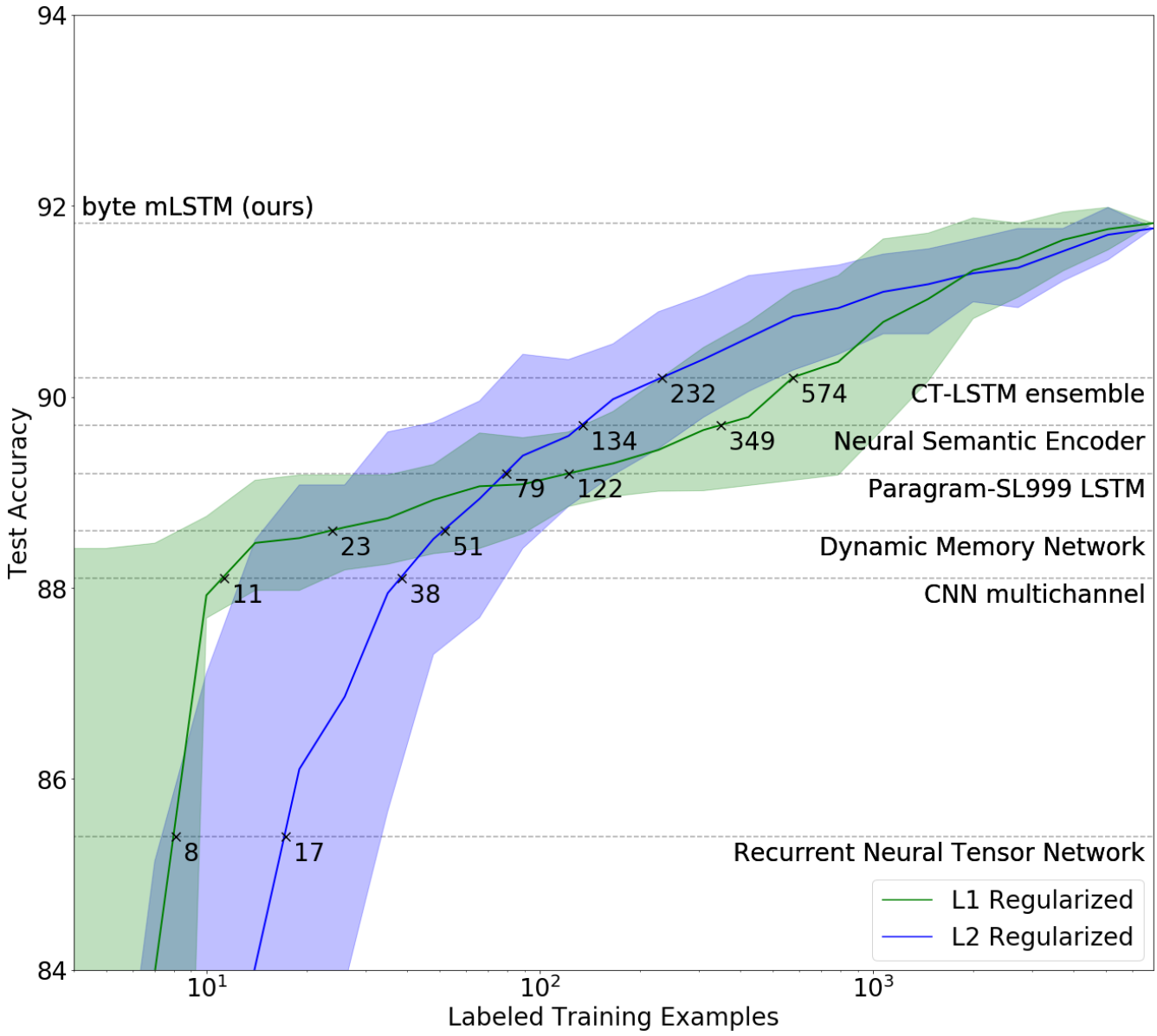
We've developed an unsupervised system which learns an excellent representation of sentiment, despite being trained only to predict the next character in the text of Amazon reviews.

April 6, 2017
6 minute read

A linear model using this representation achieves state-of-the-art sentiment analysis accuracy on a small but extensively-studied dataset, the Stanford Sentiment Treebank (we get 91.8% accuracy versus the previous best of 90.2%), and can match the performance of previous supervised systems using 30-100x fewer labeled examples. Our representation also contains a distinct "sentiment neuron" which contains almost all of the sentiment signal.

VIEW ON GITHUB VIEW ON ARXIV

Our system beats other approaches on Stanford Sentiment Treebank while using dramatically less data.



The number of labeled examples it takes two variants of our model (the green and blue lines) to match fully supervised approaches, each trained with 6,920 examples (the dashed gray lines). Our L1-regularized model (pretrained in an unsupervised fashion on Amazon reviews) matches [multichannel CNN] (https://arxiv.org/abs/1408.5882) performance with only 11 labeled examples, and state-of-the-art CT-LSTM Ensembles with 232 examples.

We were very surprised that our model learned an interpretable feature, and that simply predicting the next character in Amazon reviews resulted in discovering the concept of sentiment. We believe the phenomenon is not specific to our model, but is instead a general property of certain large neural networks that are trained to predict the next step or dimension in their inputs.

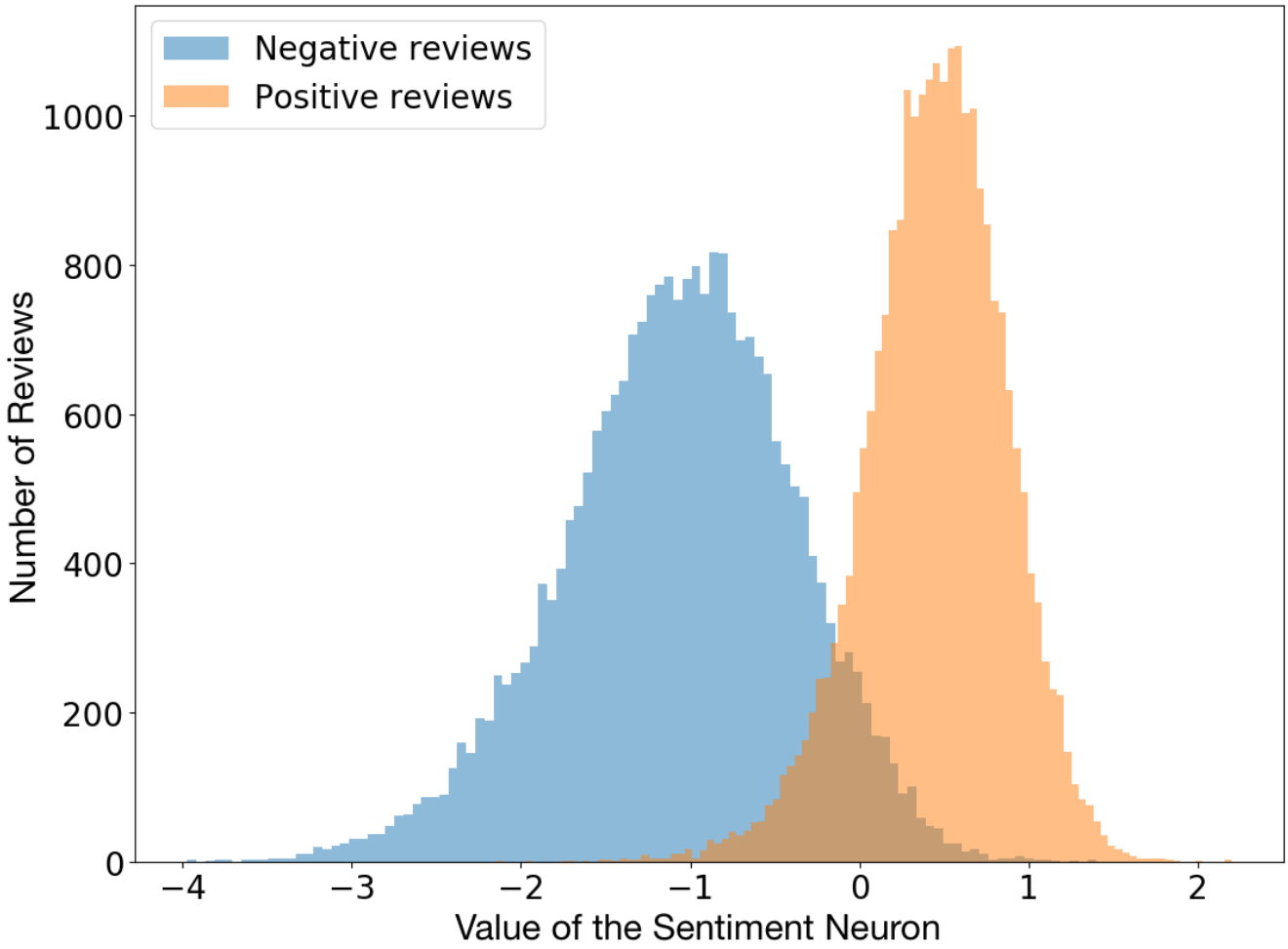
Methodology

We first trained a multiplicative LSTM with 4,096 units on a corpus of 82 million Amazon reviews to predict the next character in a chunk of text. Training took one month across four NVIDIA Pascal GPUs, with our model processing 12,500 characters per second.

These 4,096 units (which are just a vector of floats) can be regarded as a feature vector representing the string read by the model. After training the mLSTM, we turned the model into a sentiment classifier by taking a linear combination of these units, learning the weights of the combination via the available supervised data.

Sentiment neuron

While training the linear model with L1 regularization, we noticed it used surprisingly few of the learned units. Digging in, we realized there actually existed a single "sentiment neuron" that's highly predictive of the sentiment value.



The sentiment neuron within our model can classify reviews as negative or positive, even though the model is trained only to predict the next character in the text.

Just like with similar models, our model can be used to generate text. Unlike those models, we have a direct dial to control the sentiment of the resulting text: we simply overwrite the value of the sentiment neuron.

SENTIMENT FIXED TO POSITIVE	SENTIMENT FIXED TO NEGATIVE
Just what I was looking for. Nice fitted pants, exactly matched seam to color contrast with other pants I own. Highly recommended and also very happy!	The package received was blank and has no barcode. A waste of time and money.
This product does what it is supposed to. I always keep three of these in my kitchen just in case ever I need a replacement cord.	Great little item. Hard to put on the crib without some kind of embellishment. My guess is just like the screw kind of attachment I had.
Best hammock ever! Stays in place and holds it's shape. Comfy (I love the deep neon pictures on it), and looks so cute.	They didn't fit either. Straight high sticks at the end. On par with other buds I have. Lesson learned to avoid.

SENTIMENT FIXED TO POSITIVE	SENTIMENT FIXED TO NEGATIVE
Dixie is getting her Doolittle newsletter we'll see another new one coming out next year. Great stuff. And, here's the contents - information that we hardly know about or forget.	great product but no seller. couldn't ascertain a cause. Broken product. I am a prolific consumer of this company all the time.
I love this weapons look . Like I said beautiful !!! I recommend it to all. Would suggest this to many roleplayers, And I stronge to get them for every one I know. A must watch for any man who love Chess!	Like the cover, Fits good. . However, an annoying rear piece like garbage should be out of this one. I bought this hoping it would help with a huge pull down my back & the black just doesn't stay. Scrap off everytime I use it.... Very disappointed.

Examples of synthetic text generated by the trained model. Above, we select random samples from the model after fixing the sentiment unit's value to determine the sentiment of the review. Below, we also pass the prefix "*I couldn't figure out*" through the model and select high-likelihood samples only.

SENTIMENT FIXED TO POSITIVE	SENTIMENT FIXED TO NEGATIVE
I couldn't figure out the shape at first but it definitely does what it's meant to do. It's a great product and I recommend it highly	I couldn't figure out how to use the product. It did not work.At least there was no quality control; this tablet does not work. I would have given it zero stars, but that was not an option.
I couldn't figure out why this movie had been discontinued! Now I can enjoy it anytime I like. So glad to have found it again.	I couldn't figure out how to set it up being that there was no warning on the box. I wouldn't recommend this to anyone.
I couldn't figure out how to use the video or the book that goes along with it, but it is such a fantastic book on how to put it into practice!	I couldn't figure out how to use the gizmo. What a waste of time and money. Might as well through away this junk.
I couldn't figure out how to use just one and my favorite running app. I use it all the time. Good quality, You cant beat the price.	I couldn't figure out how to stop this drivel. At worst, it was going absolutely nowhere, no matter what I did.Needles to say, I skim-read the entire book. Don't waste your time.
I couldn't figure out how to attach these balls to my little portable drums, but these fit the bill and were well worth every penny.	I couldn't figure out how to play it.

Example

The diagram below represents the character-by-character value of the sentiment neuron, displaying negative values as red and positive values as green. Note that strongly indicative words like "*best*" or "*horrendous*" cause particularly big shifts in the color.

The sentiment neuron adjusting its value on a character-by-character basis.

It's interesting to note that the system also makes large updates after the completion of sentences and phrases. For example, in "*And about 99.8 percent of that got lost in the film*", there's a negative update after "*lost*" and a larger update at the sentence's end, even though "*in the film*" has no sentiment content on its own.

Unsupervised learning

Labeled data are the fuel for today's machine learning. Collecting data is easy, but scalably labeling that data is hard. It's only feasible to generate labels for important problems where the reward is worth the effort, like machine translation, speech recognition, or self-driving.

Machine learning researchers have long dreamed of developing unsupervised learning algorithms to learn a good representation of a dataset, which can then be used to solve tasks using only a few labeled examples. Our research implies that simply training large unsupervised next-step-prediction models on large amounts of data may be a good approach to use when creating systems with good representation learning capabilities.

Next steps

Our results are a promising step towards general unsupervised representation learning. We found the results by exploring whether we could learn good quality representations as a side effect of language modeling, and scaled up an existing model on a carefully-chosen dataset. Yet the underlying phenomena remain more mysterious than clear.

- These results were not as strong for datasets of long documents. We suspect our character-level model struggles to remember information over hundreds to thousands of timesteps. We think it's worth trying hierarchical models that can adapt the timescales at which they operate. Further scaling up these models may further improve representation fidelity and performance on sentiment analysis and similar tasks.
- The model struggles the more the input text diverges from review data. It's worth verifying that broadening the corpus of text samples results in an equally informative representation that also applies to broader domains.
- Our results suggest that there exist settings where very large next-step-prediction models learn excellent unsupervised representations. Training a large neural network to predict the next frame in a large collection of videos may result in unsupervised representations for object, scene, and action classifiers.

Overall, it's important to understand the properties of models, training regimes, and datasets that reliably lead to such excellent representations.

Authors
Alec Radford, Ilya Sutskever, Rafał Józefowicz, Jack Clark & Greg Brockman

Cover Artwork
Ludwig Pettersson

Filed Under
Research, Milestones