Learning to Generate Reviews and Discovering Sentiment

Anonymous authors
Paper under double-blind review

Abstract

We explore the properties of byte-level recurrent language models. When given sufficient amounts of capacity, training data, and compute time, the representations learned by these models include disentangled features corresponding to high-level concepts. Specifically, we find a single unit which performs sentiment analysis. These representations, learned in an unsupervised manner, achieve state of the art on the binary subset of the Stanford Sentiment Treebank. They are also very data efficient. When using only a handful of labeled examples, our approach matches the performance of strong baselines trained on full datasets. We also demonstrate the sentiment unit has a direct influence on the generative process of the model. Simply fixing its value to be positive or negative generates samples with the corresponding positive or negative sentiment.

1 Introduction and Motivating Work

Representation learning (Bengio et al. (2013)) plays a critical role in many modern machine learning systems. Representations map raw data to more useful forms and the choice of representation is an important component of any application. Broadly speaking, there are two areas of research emphasizing different details of how to learn useful representations.

The supervised training of high-capacity models on large labeled datasets is critical to the recent success of deep learning techniques for a wide range of applications such as image classification (Krizhevsky et al. (2012)), speech recognition (Hinton et al. (2012)), and machine translation (Wu et al. (2016)). Analysis of the task specific representations learned by these models reveals many fascinating properties (Zhou et al. (2014)). Image classifiers learn a broadly useful hierarchy of feature detectors re-representing raw pixels as edges, textures, and objects (Zeiler & Fergus (2014)). In the field of computer vision, it is now commonplace to reuse these representations on a broad suite of related tasks - one of the most successful examples of transfer learning to date (Oquab et al. (2014)).

There is also a long history of unsupervised representation learning (Olshausen & Field (1997)). Much of the early research into modern deep learning was developed and validated via this approach (Hinton & Salakhutdinov (2006); Huang et al. (2007); Vincent et al. (2008); Coates et al. (2010); Le (2013)). Unsupervised learning is promising due to its ability to scale beyond the small subsets and subdomains of data that can be cleaned and labeled given resource, privacy, or other constraints. This advantage is also its difficulty. While supervised approaches have clear objectives that can be directly optimized, unsupervised approaches rely on proxy tasks such as reconstruction, density estimation, or generation, which do not directly encourage useful representations for specific tasks. As a result, much work has gone into designing objectives, priors, and architectures meant to encourage the learning of useful representations. We refer readers to Goodfellow et al. (2016) for a detailed review.

Despite these difficulties, there are notable applications of unsupervised learning. Pre-trained word vectors are a vital part of many modern NLP systems (Collobert et al. (2011)). These representations, learned by modeling word co-occurrences, increase the data efficiency and generalization capability of NLP systems (Pennington et al. (2014); Chen & Manning (2014)). Topic modelling can also discover factors within a corpus of text which align to human interpretable concepts such as "art" or "education" (Blei et al. (2003)).

How to learn representations of phrases, sentences, and documents is an open area of research. Inspired by the success of word vectors, Kiros et al. (2015) propose skip-thought vectors, a method of training a sentence encoder by predicting the preceding and following sentence. The representation learned by this objective performs competitively on a broad suite of evaluated tasks. More advanced training techniques such as layer normalization (Ba et al. (2016)) further improve results. However, skip-thought vectors are still outperformed by supervised models which directly optimize the desired performance metric on a specific dataset. This is the case for both text classification tasks, which measure whether a specific concept is well encoded in a representation, and more general semantic similarity tasks. This occurs even when the datasets are relatively small by modern standards, often consisting of only a few thousand labeled examples.

In contrast to learning a generic representation on one large dataset and then evaluating on other tasks/datasets, Dai & Le (2015) proposed using similar unsupervised objectives such as sequence autoencoding and language modeling to first pretrain a model on a dataset and then finetune it for a given task. This approach outperformed training the same model from random initialization and achieved state of the art on several text classification datasets. Combining word-level language modelling of a dataset with topic modelling and fitting a small neural network feature extractor on top has also achieved strong results on document level sentiment analysis (Dieng et al. (2016)).

Considering this, we hypothesize two effects may be combining to result in the weaker performance of purely unsupervised approaches. Skip-thought vectors were trained on a corpus of books. But some of the classification tasks they are evaluated on, such as sentiment analysis of reviews of consumer goods, do not have much overlap with the text of novels. We propose this distributional issue, combined with the limited capacity of current models, results in representational underfitting. Current generic distributed sentence representations may be very lossy - good at capturing the gist, but poor with the precise semantic or syntactic details which are critical for applications.

The experimental and evaluation protocols may be underestimating the quality of unsupervised representation learning for sentences and documents due to certain seemingly insignificant design decisions. Hill et al. (2016) also raises concern about current evaluation tasks in their recent work which provides a thorough survey of architectures and objectives for learning unsupervised sentence representations - including the above mentioned skip-thoughts.

In this work, we test whether this is the case. We focus in on the task of sentiment analysis and attempt to learn an unsupervised representation that accurately contains this concept. Mikolov et al. (2013) showed that word-level recurrent language modelling supports the learning of useful word vectors. We are interested in pushing this line of work to learn representations of not just words but arbitrary scales of text with no distinction between sub-word, word, phrase, sentence, or document-level structure. Recent work has shown that traditional NLP task such as Named Entity Recognition and Part-of-Speech tagging can be performed this way by processing text as a byte sequence (Gillick et al. (2015)). Byte level language modelling is a natural choice due to its simplicity and generality. We are also interested in evaluating this approach as it is not immediately clear whether such a low-level training objective supports the learning of high-level representations. We train on a very large corpus picked to have a similar distribution as our task of interest. We also benchmark on a wider range of tasks to quantify the sensitivity of the learned representation to various degrees of out-of-domain data and tasks.

2 DATASET

Much previous work on language modeling has evaluated on relatively small but competitive datasets such as Penn Treebank (Marcus et al. (1993)) and Hutter Prize Wikipedia (Hutter (2006)). As discussed in Jozefowicz et al. (2016) performance on these datasets is primarily dominated by regularization. Since we are interested in high-quality sentiment representations, we chose the Amazon product review dataset introduced in McAuley et al. (2015) as a training corpus. In de-duplicated form, this dataset contains over 82 million product reviews from May 1996 to July 2014 amounting to over 38 billion training bytes. Due to the size of the dataset, we first split it into 1000 shards containing equal numbers of reviews and set aside 1 shard for validation and 1 shard for test.

3 MODEL AND TRAINING DETAILS

Many potential recurrent architectures and hyperparameter settings were considered in preliminary experiments on the dataset. Given the size of the dataset, searching the wide space of possible configurations is quite costly. To help alleviate this, we evaluated the generative performance in terms of log-likelihood of smaller candidate models after a single pass through the dataset and selected the best performing architecture according to this metric for large scale experiments. The model chosen is a single layer multiplicative LSTM (Krause et al. (2016)) with 4096 units.

We observed multiplicative LSTMs to converge faster than normal LSTMs for the hyperparameter settings that were explored both in terms of data and wall-clock time. The model was trained for a single epoch on mini-batches of 128 subsequences of length 256 for a total of 1 million weight updates. States were initialized to zero at the beginning of each shard and persisted across updates to simulate full-backpropagation and allow for the forward propagation of information outside of a given subsequence. Adam (Kingma & Ba (2014)) was used to accelerate learning with an initial 5e-4 learning rate that was decayed linearly to zero over the course of training. Weight normalization (Salimans & Kingma (2016)) was applied to the LSTM parameters. Data-parallelism was used across 4 Pascal Titan X gpus to speed up training and increase effective memory size. Training took approximately one month. The model is compact, containing approximately as many parameters as there are reviews in the training dataset. It also has a high ratio of compute to total parameters due to operating at a byte level. The model reaches 1.12 bits per byte.
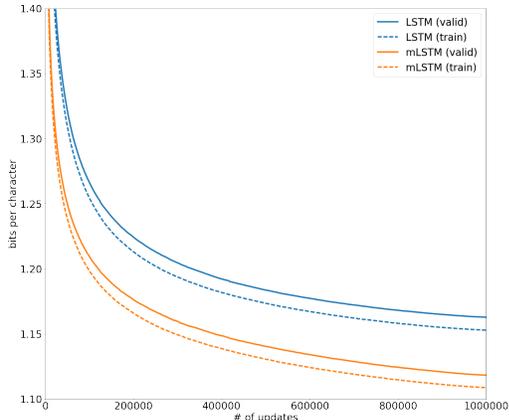

Figure 1: The mLSTM converges faster and achieves a better result within our time budget compared to a standard LSTM with the same hidden state size

4 EXPERIMENTAL SETUP AND RESULTS

Our model processes text as a sequence of UTF-8 encoded bytes (Yergeau (2003)). For each byte, the model updates its hidden state and predicts a probability distribution over the next possible byte. The hidden state of the model serves as an online summary of the sequence which encodes all information the model has learned to preserve that is relevant to predicting the future bytes of the sequence. We are interested in understanding the properties of the learned encoding. The process of extracting a feature representation is outlined as follows:

- Since newlines are used as review delimiters in the training dataset, all newline characters are replaced with spaces to avoid the model resetting state.
- Any leading whitespace is removed and replaced with a newline+space to simulate a start token. Any trailing whitespace is removed and replaced with a space to simulate an end token. HTML is unescaped and HTML tags are removed to mirror preprocessing applied to the training dataset. The text is encoded as a UTF-8 byte sequence.
- Model states are initialized to zeros. The model processes the sequence and the final cell states of the mLSTM are used as a feature representation.

We follow the methodology established in Kiros et al. (2015) by training a logistic regression classifier on top of our model's representation on datasets for tasks including semantic relatedness, text classification, and paraphrase detection. For the details on these comparison experiments, we refer the reader to their work. One exception is that we use an L1 penalty for text classification results instead of L2 as we found this performed better in the very low data regime.
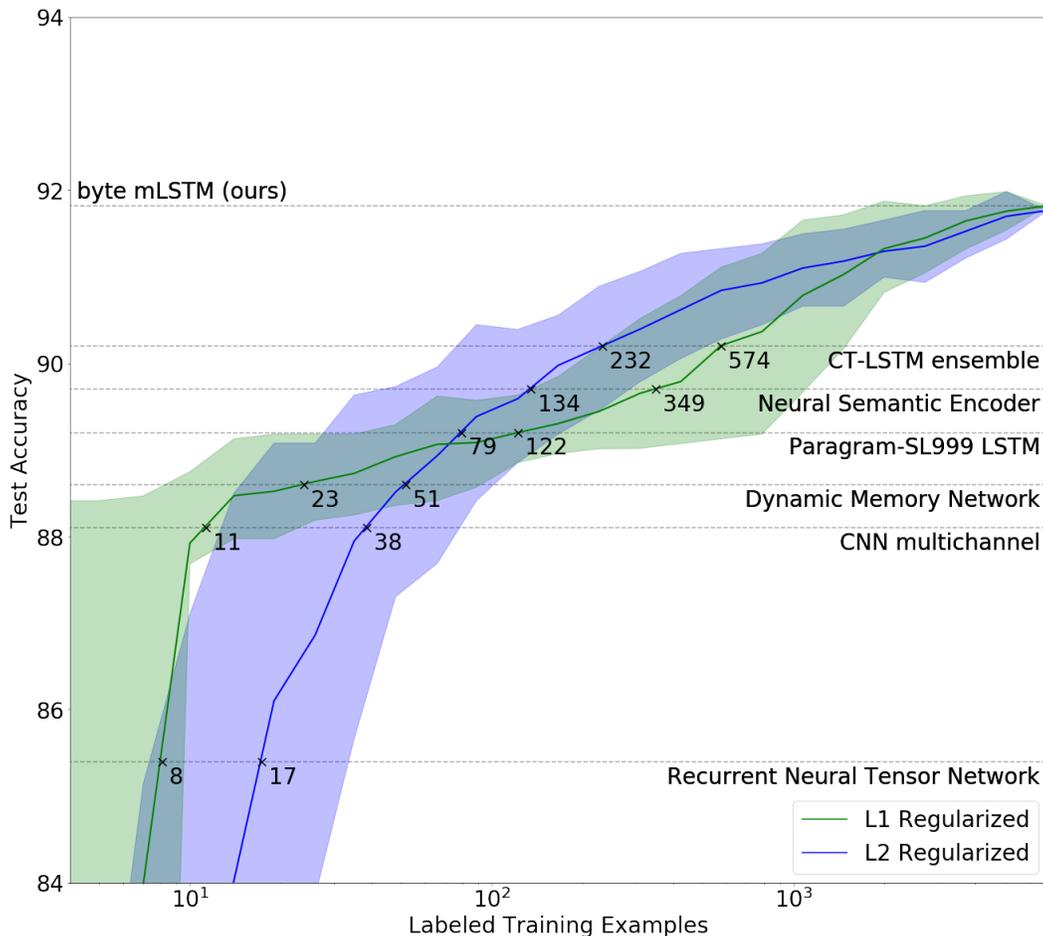
Figure 2: Performance on the binary version of SST as a function of labeled training examples. The solid lines indicate the average of 100 runs while the sharded regions indicate the 10th and 90th percentiles. Previous results on the dataset are plotted as dashed lines with the numbers indicating the amount of examples required for logistic regression on the byte mLSTM representation to match their performance. RNTN (Socher et al. (2013)) CNN (Kim (2014)) DMN (Kumar et al. (2015)) P-SL999 LSTM (Wieting et al. (2015)) NSE (Munkhdalai & Yu (2016)) CT-LSTM (Looks et al. (2017))

4.1 TEXT CLASSIFICATION AND SENTIMENT ANALYSIS

Table 1 shows the results of our model on 4 standard text classification datasets. The performance of our model is noticeably lopsided. On the MR (Pang & Lee (2005)) and CR (Hu & Liu (2004)) sentiment analysis datasets we improve the state of the art by a significant margin. The MR and CR datasets are sentences extracted from Rotten Tomatoes, a movie review website, and Amazon product reviews (which almost certainly overlaps with our training corpus). This suggests that our model has learned a rich representation of text from a similar domain. On the other two datasets, SUBJ's subjectivity/objectivity detection (Pang & Lee (2004)) and MPQA's opinion polarity (Wiebe et al. (2005)) our model has no noticeable advantage over other unsupervised representation learning approaches and is still outperformed by a supervised approach.

To better quantify the learned representation, we also test on a wider set of sentiment analysis datasets with different properties. The Stanford Sentiment Treebank (SST) (Socher et al. (2013)) was created specifically to evaluate more complex compositional models of language. It is derived from the same base dataset as MR but was relabeled via Amazon Mechanical Turk and includes dense labeling of the phrases of parse trees computed for all sentences. For the binary subtask, this amounts to

Method	MR	CR	SUBJ	MPQA
NBSVM Wang & Manning (2012)	79.4	81.8	93.2	86.3
ST Kiros et al. (2015)	77.3	81.8	92.6	87.9
STLN Ba et al. (2016)	79.5	83.1	93.7	89.3
SDAE Hill et al. (2016)	74.6	78.0	90.8	86.9
CNN Kim (2014)	81.5	85.0	93.4	89.6
Adasent Zhao et al. (2015)	83.1	86.3	**95.5**	**93.3**
byte mLSTM	**86.9**	**91.4**	94.6	88.5

Table 1: Small dataset classification accuracies

Method	Err
FullBoW Maas et al. (2011)	11.1
NBSVM Mesnil et al. (2014)	8.1
Sentiment unit (ours)	7.7
SA-LSTM Dai & Le (2015)	7.2
byte mLSTM (ours)	7.1
TopicRNN Dieng et al. (2016)	6.2
Virt Adv Miyato et al. (2016)	5.9

Table 2: IMDB sentiment analysis

76961 total labels compared to the 6920 sentence level labels. As a demonstration of the capability of unsupervised representation learning to simplify data collection and remove steps from a traditional NLP pipeline, our reported results ignore these dense labels and computed parse trees, using only the raw text and sentence level labels.

The representation learned by our model achieves 91.8% significantly outperforming the state of the art of 90.2% by a 30 model ensemble (Looks et al. (2017)). As visualized in Figure 2, our model is very data efficient. It matches the performance of baselines using as few as a dozen labeled examples and outperforms all previous results with only a few hundred labeled examples. This is under 10% of the total sentences in the dataset. Confusingly, despite a 16% relative error reduction on the binary subtask, it does not reach the state of the art of 53.6% on the fine-grained subtask, achieving 52.9%.

4.2 SENTIMENT UNIT

We conducted further analysis to understand what representations our model learned and how they achieve the observed data efficiency. The benefit of an L1 penalty in the low data regime (see Figure 2) is a clue. L1 regularization is known to reduce sample complexity when there are many irrelevant features (Ng (2004)). This is likely to be the case for our model since it is trained as a language model and not as a supervised feature extractor. By inspecting the relative contributions of features on various datasets, we discovered a single unit within the mLSTM *that directly corresponds to sentiment*. In Figure 3 we show the histogram of the final activations of this unit after processing IMDB reviews (Maas et al. (2011)) which shows a bimodal distribution with a clear separation between positive and negative reviews. In Figure 5 we visualize the activations of this unit on 6 randomly

Figure 3: Histogram of cell values for the sentiment unit on IMDB reviews.

selected reviews from a set of 100 high contrast reviews which shows it acts as an online estimate of the local sentiment of the review. Fitting a threshold to this *single* unit achieves a test accuracy of 92.30% which outperforms a strong supervised results on the dataset, the 91.87% of NB-SVM trigram (Mesnil et al. (2014)), but is still below the semi-supervised state of the art of 94.09% (Miyato et al. (2016)). Using the full 4096 unit representation achieves 92.88%. This is an improvement of only 0.58% over the sentiment unit suggesting that almost all information the model retains that is relevant to sentiment analysis is represented in the very compact form of a single scalar. Table 2 has a full list of results on the IMDB dataset.
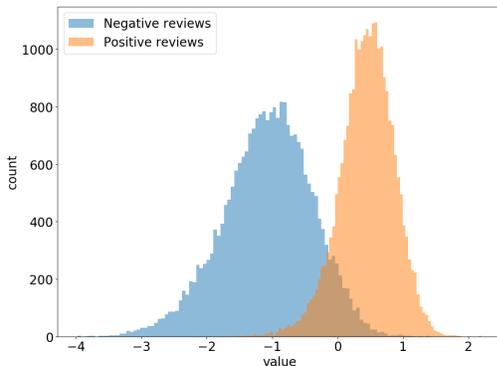
Method	r	ρ	MSE
ST Kiros et al. (2015)	0.858	0.792	0.269
STLN Ba et al. (2016)	0.858	0.788	0.270
Tree-LSTM Tai et al. (2015)	**0.868**	**0.808**	**0.253**
byte mLSTM	0.792	0.725	0.390

Table 3: SICK semantic relatedness subtask

Method	Acc	F1
ST Kiros et al. (2015)	73.0	82.0
SDAE Hill et al. (2016)	76.4	83.4
MTMETRICS Madnani et al. (2012)	**77.4**	**84.1**
byte mLSTM	75.0	82.8

Table 4: Microsoft Paraphrase Corpus

4.3 CAPACITY CEILING

Encouraged by these results, we were curious how well the model's representation scales to larger datasets. We try our approach on the binary version of the Yelp Dataset Challenge in 2015 as introduced in Zhang et al. (2015). This dataset contains 598,000 examples which is an order of magnitude larger than any other datasets we tested on. When visualizing performance as a function of number of training examples in Figure 4, we observe a "capacity ceiling" where the test accuracy of our approach only improves by a little over 1% across a four order of magnitude increase in training data. Using the full dataset, we achieve 95.22% test accuracy. This better than a BoW TFIDF baseline at 93.66% but slightly worse than the 95.64% of a linear classifier on top of the 500,000 most frequent n-grams up to length 5.

The observed capacity ceiling is an interesting phenomena and stumbling point for scaling our unsupervised representations. We think a variety

Figure 4: Performance on the binary version of the Yelp reviews dataset as a function of labeled training examples. The model's performance plateaus after about ten labeled examples and only slow improves with additional data.

of factors are contributing to cause this. Since our model is trained only on Amazon reviews, it is does not appear to be sensitive to concepts specific to other domains. For instance, Yelp reviews are of businesses, where details like hospitality, location, and atmosphere are important. But these ideas are not present in reviews of products. Additionally, there is a notable drop in the relative performance of our approach transitioning from sentence to document datasets. This is likely due to our model working on the byte level which leads to it focusing on the content of the last few sentences instead of the whole document. Finally, as the amount of labeled data increases, the performance of a simple linear model trained on top of a static representation will eventually saturate. Complex models explicitly trained for a task can continue to improve and eventually outperform our approach when provided with enough labeled data.

With this context, the observed results reveal a natural ordering. On a small sentence level dataset from a very similar domain (the movie reviews of Stanford Sentiment Treebank) our model sets a new state of the art. But on a large, document level dataset from a slightly different domain (the Yelp reviews) it is only competitive with standard baselines.

4.4 REPRESENTATION STABILITY

The features of our model are learned via an unsupervised objective. A potential concern with this training methodology is that it is unreliable. There is no guarantee that the representations learned will be useful for a desired task because the model is not optimizing directly for this. This could lead to high variance of performance on desired tasks. This concern is further amplified for the properties of individual features and not just the representation as whole. While the work of Li et al. (2015) has shown that networks differing in only random initialization can learn the same features, this was
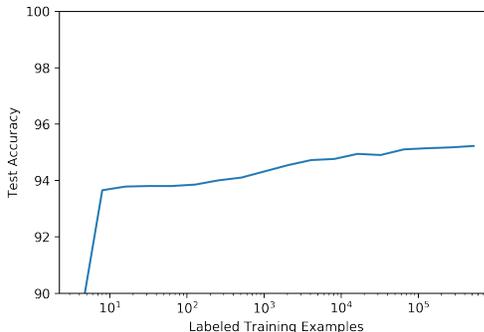
Sentiment fixed to positive	Sentiment fixed to negative
Just what I was looking for. Nice fitted pants, exactly matched seam to color contrast with other pants I own. Highly recommended and also very happy!	The package received was blank and has no barcode. A waste of time and money.
This product does 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.
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.

Table 5: Random samples from the model generated when the value of sentiment unit hidden state is fixed to either -1 or 1 for all steps. The sentiment unit has a strong influence on the model's generative process.

tested for CNNs trained for supervised image classification. To check whether this is the case for our approach, we trained a second model using the same hyperparameters differing only in weight parameterization and compared to the model discussed in this paper. The best single unit in this model (selected via validation set) achieves 92.42% test accuracy on IMDB compared to the 92.30% of the unit visualized in this paper. This suggests that the sentiment unit is an example of a convergent representation.

4.5 OTHER TASKS

Besides classification, we also evaluate on two other standard tasks: semantic relatedness and paraphrase detection. While our model performs competitively on Microsoft Research Paraphrase Corpus (Dolan et al. (2004)) in Table 3, it performs poorly on the SICK semantic relatedness task (Marelli et al. (2014)) in Table 4. It is likely that the form and content of the semantic relatedness task, which is built on top of descriptions of images and videos and contains sentences such as "A sea turtle is hunting for fish" is effectively out-of-domain for our model which has only been trained on the text of product reviews.

4.6 GENERATIVE ANALYSIS

Although the focus of our analysis has been on the properties of our model's representation, it is trained as a generative model and we are also interested in its generative capabilities. Hu et al. (2017) and Dong et al. (2017) both designed conditional generative models to disentangle the content of text from various attributes like sentiment or tense. We were curious whether a similar result could be achieved using the sentiment unit. In Table 5 we show that by simply setting the sentiment unit to be positive or negative, the model generates corresponding positive or negative reviews. While all sampled negative reviews contain sentences with negative sentiment, they sometimes contain sentences with positive sentiment as well. This might be reflective of the bias of the training corpus which contains over 5x as many five star reviews as one star reviews. Nevertheless, it is interesting to see that such a simple manipulation of the model's representation has a noticeable effect on its behavior. The samples are also high quality for a byte level language model and often include valid sentences.

25 August 2003 League of Extraordinary Gentlemen: Sean Connery is one of the all time greats and I have been a fan of his since the 1950's. I went to this movie because Sean Connery was the main actor. I had not read reviews or had any prior knowledge of the movie. The movie surprised me quite a bit. The scenery and sights were spectacular, but the plot was unreal to the point of being ridiculous. In my mind this was not one of his better movies it could be the worst. Why he chose to be in this movie is a mystery. For me, going to this movie was a waste of my time. I will continue to go to his movies and add his movies to my video collection. But I can't see wasting money to put this movie in my collection

Judy Holliday struck gold in 1950 with George Cukor's film version of "Born Yesterday," and from that point forward, her career consisted of trying to find material good enough to allow her to strike gold again. It never happened. In "It Should Happen to You" (I can't think of a blander title, by the way), Holliday does yet one more variation on the dumb blonde who's maybe not so dumb after all, but everything about this movie feels warmed over and half hearted. Even Jack Lemmon, in what I believe was his first film role, can't muster up enough energy to enliven this recycled comedy. The audience knows how the movie will end virtually from the beginning, so mostly it just sits around waiting for the film to catch up. Maybe if you're enamored of Holliday you'll enjoy this; otherwise I wouldn't bother. Grade: C

Team Spirit is maybe made by the best intentions, but it misses the warmth of "All Stars" (1997) by Jean van de Velde. Most scenes are identic, just not that funny and not that well done. The actors repeat the same lines as in "All Stars" but without much feeling.

I found this to be a charming adaptation, very lively and full of fun. With the exception of a couple of major errors, the cast is wonderful. I have to echo some of the earlier comments -- Chynna Phillips is horribly miscast as a teenager. At 27, she's just too old (and, yes, it DOES show), and lacks the singing "chops" for Broadway-style music. Vanessa Williams is a decent-enough singer and, for a non-dancer, she's adequate. However, she is NOT Latina, and her character definitely is. She's also very STRIDENT throughout, which gets tiresome. The girls of Sweet Apple's Conrad Birdie fan club really sparkle -- with special kudos to Brigitta Dau and Chiara Zanni. I also enjoyed Tyne Daly's performance, though I'm not generally a fan of her work. Finally, the dancing Shriners are a riot, especially the dorky three in the bar. The movie is suitable for the whole family, and I highly recommend it.

God bless Randy Quaid...his leachorous Cousin Eddie in Vacation and Christmas Vacation hilariously stole the show. He even made the awful Vegas Vacation at least worth a look. I will say that he tries hard in this made for TV sequel, but that the script is so NON funny that the movie never really gets anywhere. Quaid and the rest of the returning Vacation vets (including the orginal Audrey, Dana Barron) are wasted here. Even European Vacation's Eric Idle cannot save the show in a brief cameo.... Pathetic and sad...actually painful to watch....Christmas Vacation 2 is the worst of the Vacation franchise.

Once in a while you get amazed over how BAD a film can be, and how in the world anybody could raise money to make this kind of crap. There is absolutely No talent included in this film - from a crappy script, to a crappy story to crappy acting. Amazing...
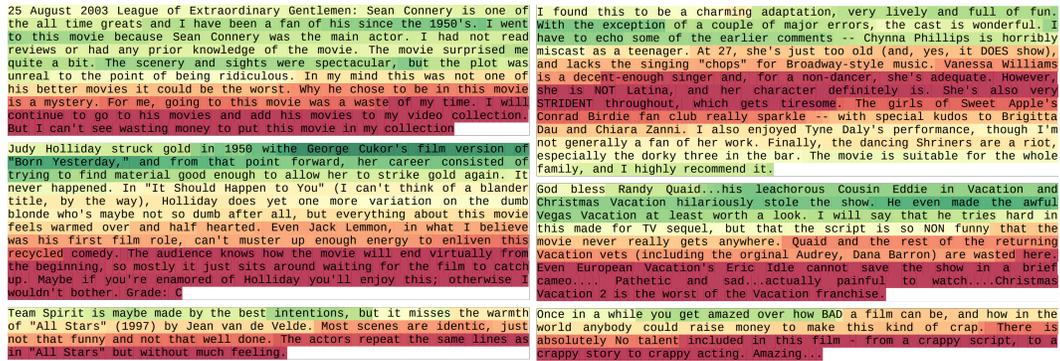
Figure 5: Visualizing the value of the sentiment unit's cell state as it processes six randomly selected high contrast IMDB reviews. Red indicates negative sentiment while green indicates positive sentiment. Best seen in color.

5 DISCUSSION AND FUTURE WORK

It is an open question why our model recovers the concept of sentiment in such a precise, disentangled, interpretable, and manipulable way. It is possible that sentiment as a conditioning feature has strong predictive capability for language modelling. This is likely since sentiment is such an important component of a review. Previous work analyzing LSTM language models showed the existence of interpretable units that indicate position within a line or presence inside a quotation (Karpathy et al. (2015)). In many ways, the sentiment unit in this model is just a scaled up example of the same phenomena. The update equation of an LSTM could play a role. The element-wise operation of its gates may encourage axis-aligned representations. Models such as word2vec have also been observed to have small subsets of dimensions strongly associated with specific tasks (Li et al. (2016)).

Our work highlights the sensitivity of learned representations to the data distribution they are trained on. The results make clear that it is unrealistic to expect a model trained on a corpus of books, where the two most common genres are Romance and Fantasy, to learn an encoding which preserves the exact sentiment of a review. Likewise, it is unrealistic to expect a model trained on Amazon product reviews to represent the precise semantic content of a caption of an image or a video.

There are several promising directions for future work highlighted by our results. The observed performance plateau, even on relatively similar domains, suggests improving the representation model both in terms of architecture and size. Since our model operates at the byte-level, hierarchical/multi-timescale extensions could improve the quality of representations for longer documents. The sensitivity of learned representations to their training domain could be addressed by training on a wider mix of datasets with better coverage of target tasks. Finally, our work encourages further research into language modelling as it demonstrates that the standard language modelling objective with no modifications is sufficient to learn high-quality representations.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Danqi Chen and Christopher D Manning. A fast and accurate dependency parser using neural networks. In *EMNLP*, pp. 740–750, 2014.

Adam Coates, Honglak Lee, and Andrew Y Ng. An analysis of single-layer networks in unsupervised feature learning. *Ann Arbor*, 1001(48109):2, 2010.

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

Andrew M Dai and Quoc V Le. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems*, pp. 3079–3087, 2015.

Adji B Dieng, Chong Wang, Jianfeng Gao, and John Paisley. Topicrnn: A recurrent neural network with long-range semantic dependency. *arXiv preprint arXiv:1611.01702*, 2016.

Bill Dolan, Chris Quirk, and Chris Brockett. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th international conference on Computational Linguistics*, pp. 350. Association for Computational Linguistics, 2004.

Li Dong, Shaohan Huang, Furu Wei, Mirella Lapata, Ming Zhou, and Xu Ke. Learning to generate product reviews from attributes. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 623–632. Association for Computational Linguistics, 2017.

Dan Gillick, Cliff Brunk, Oriol Vinyals, and Amarnag Subramanya. Multilingual language processing from bytes. *arXiv preprint arXiv:1512.00103*, 2015.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. 2016.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. Learning distributed representations of sentences from unlabelled data. *arXiv preprint arXiv:1602.03483*, 2016.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 168–177. ACM, 2004.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. Controllable text generation. *arXiv preprint arXiv:1703.00955*, 2017.

Fu Jie Huang, Y-Lan Boureau, Yann LeCun, et al. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pp. 1–8. IEEE, 2007.

Marcus Hutter. The human knowledge compression contest. 2006. *URL http://prize. hutter1. net*, 2006.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016.

Andrej Karpathy, Justin Johnson, and Li Fei-Fei. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*, 2015.

Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

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

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pp. 3294–3302, 2015.

Ben Krause, Liang Lu, Iain Murray, and Steve Renals. Multiplicative lstm for sequence modelling. *arXiv preprint arXiv:1609.07959*, 2016.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Ankit Kumar, Ozan Irsoy, Jonathan Su, James Bradbury, Robert English, Brian Pierce, Peter Ondruska, Ishaan Gulrajani, and Richard Socher. Ask me anything: Dynamic memory networks for natural language processing. *CoRR, abs/1506.07285*, 2015.

Quoc V Le. Building high-level features using large scale unsupervised learning. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 8595–8598. IEEE, 2013.

Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*, 2016.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? In *Feature Extraction: Modern Questions and Challenges*, pp. 196–212, 2015.

Moshe Looks, Marcello Herreshoff, DeLesley Hutchins, and Peter Norvig. Deep learning with dynamic computation graphs. *arXiv preprint arXiv:1702.02181*, 2017.

Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 142–150. Association for Computational Linguistics, 2011.

Nitin Madnani, Joel Tetreault, and Martin Chodorow. Re-examining machine translation metrics for paraphrase identification. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 182–190. Association for Computational Linguistics, 2012.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330, 1993.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. *SemEval-2014*, 2014.

Julian McAuley, Rahul Pandey, and Jure Leskovec. Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. ACM, 2015.

Grégoire Mesnil, Tomas Mikolov, Marc'Aurelio Ranzato, and Yoshua Bengio. Ensemble of generative and discriminative techniques for sentiment analysis of movie reviews. *arXiv preprint arXiv:1412.5335*, 2014.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. 2013.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*, 2016.

Tsendsuren Munkhdalai and Hong Yu. Neural semantic encoders. *arXiv preprint arXiv:1607.04315*, 2016.

Andrew Y Ng. Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 78. ACM, 2004.

Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1717–1724, 2014.

Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, pp. 271. Association for Computational Linguistics, 2004.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pp. 115–124. Association for Computational Linguistics, 2005.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pp. 1532–1543, 2014.

Tim Salimans and Diederik P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 901–901, 2016.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, Christopher Potts, et al. Recursive deep models for semantic compositionality over a sentiment treebank. Citeseer, 2013.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*, 2015.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103. ACM, 2008.

Sida Wang and Christopher D Manning. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pp. 90–94. Association for Computational Linguistics, 2012.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2):165–210, 2005.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. Towards universal paraphrastic sentence embeddings. *arXiv preprint arXiv:1511.08198*, 2015.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

Francois Yergeau. Utf-8, a transformation format of iso 10646. 2003.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pp. 649–657, 2015.

Han Zhao, Zhengdong Lu, and Pascal Poupart. Self-adaptive hierarchical sentence model. *arXiv preprint arXiv:1504.05070*, 2015.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014.