

# Active Discriminative Text Representation Learning

**Ye Zhang**

Department of Computer Science  
University of Texas at Austin  
yezhang@utexas.edu

**Matthew Lease**

School of Information  
University of Texas at Austin  
ml@utexas.edu

**Byron C. Wallace**

College of Computer & Information Science  
Northeastern University  
byron@ccs.neu.edu

## Abstract

We propose a new *active learning* (AL) method for text classification with convolutional neural networks (CNNs). In AL, one selects the instances to be manually labeled with the aim of maximizing model performance with minimal effort. Neural models capitalize on word embeddings as representations (features), tuning these to the task at hand. We argue that AL strategies for multi-layered neural models should focus on selecting instances that most affect the embedding space (i.e., induce discriminative word representations). This is in contrast to traditional AL approaches (e.g., entropy-based uncertainty sampling), which specify higher level objectives.

We propose a simple approach for sentence classification that selects instances containing words whose embeddings are likely to be updated with the greatest magnitude, thereby rapidly learning discriminative, task-specific embeddings. We extend this approach to document classification by jointly considering: (1) the expected changes to the constituent word representations; and (2) the model's current overall uncertainty regarding the instance. The relative emphasis placed on these criteria is governed by a stochastic process that favors selecting instances likely to improve representations at the outset of learning, and then shifts toward general uncertainty sampling as AL progresses. Empirical results show that our method outperforms baseline AL approaches on both sentence and document classification tasks. We also show that, as expected, the method quickly learns discriminative word embeddings. To the best of our knowledge, this is the first work on AL addressing neural models for text classification.

## Introduction

In *active learning* (AL), the machine learning algorithm being trained is allowed to select the examples to be manually annotated by the teacher (Settles 2010). The idea is that by selecting training data cleverly, rather than at i.i.d. random, better models can be learned with less effort, and thus at lower cost. This approach is attractive in scenarios in which labels are expensive but unlabeled data is plentiful.

There has been a wealth of work on AL approaches for traditional machine learning methods in general (Settles 2010), and for text classification in particular (Tong and Koller 2002; McCallum and Nigam 1998; Wallace et al. 2010). However, almost no work has considered AL for text

classification using modern neural models. We posit that the importance of *representation learning* (Bengio 2009) with neural models motivates exploring a rather different approach to AL for neural models vs. classic techniques.

In this work, we propose an AL method for convolutional neural networks (CNNs), which have recently achieved strong performance across many diverse text classification tasks (Kim 2014; Zhang and Wallace 2015; Johnson and Zhang 2014; Zhang, Roller, and Wallace 2016; Zhang, Marshall, and Wallace 2016). These models first project words in texts into a low dimensional *embedding layer*, and then apply convolution operations on the resultant matrix.

While CNNs (and neural networks more generally) have demonstrated excellent performance when one has access to large amounts of training data, how can we make the best use of CNNs when annotation resources are scarce? Because word embedding estimation and tuning (for a specific text classification task) may be viewed as *representation learning*, it is reasonable to optimize feature vectors before expending effort to tune the parameters of a model that accepts these as input. Indeed, adjusting the former will render updates to the latter potentially useless. Thus, we argue that the objective in AL (at least at the outset) should primarily be to select instances that result in better representations.

More specifically, we propose a novel AL approach for sentence classification in which we select instances that contain words likely to most affect the embeddings. We achieve this by calculating the expected gradient length (EGL) *with respect to the embeddings* for each word comprising the remaining unlabeled sentences. We show that this approach allows us to rapidly learn discriminative, task-specific embeddings. For example, when classifying the sentiment of sentences, we find that selecting examples in this way quickly pushes the embeddings of ‘bad’ and ‘good’ apart (**Figure 3**, bottom row). Ultimately, results show our AL method improves accuracy over several baseline AL approaches, across sentence and document classification tasks considered.

This method selects instances based on a max operator over the gradients expected for the individual words in a text, and thus is less appropriate for longer texts such as documents. Therefore, we extend our approach for document classification by linearly combining two scores: one corresponding to individual word embeddings and one measuring the overall uncertainty regarding instances.

In summary, key contributions of this paper include:

- As far as we are aware, this is the first work to consider AL strategies explicitly for neural architectures in the context of text classification.
- We demonstrate that variants of our model outperform baseline AL approaches that do not consider embedding-level parameters: on both sentence and document classification tasks our method realizes better performance with fewer labels, compared to baseline sampling approaches.
- We also note that our approach substantially reduces the computational cost of AL, compared to previously proposed EGL approaches to AL.

## Preliminaries

### CNNs for Text Classification

We briefly review CNNs for text classification. Specifically we summarize the model proposed by Kim (2014) and explored in depth by Zhang and Wallace (2015). We will denote the word *embedding matrix* by  $\mathbf{E} \in \mathbb{R}^{V \times d}$ , where  $V$  is the vocabulary size, and  $d$  is the dimension of the embedding layer. A specific instance (sentence) is then represented by stacking the vectors corresponding to the words it contains (stored in  $\mathbf{E}$ ), preserving word order. This results in an instance matrix  $\mathbf{A} \in \mathbb{R}^{S \times d}$ , where  $S$  is the text length.

Convolution operations are then applied to this matrix, using multiple linear filters. Each filter matrix  $\mathbf{W}_i \in \mathbb{R}^{h \times d}$  performs a convolution operation on  $\mathbf{A}$ , generating a feature map  $\mathbf{c}_i \in \mathbb{R}^{S-h+1}$ . One-max pooling can then be applied to each  $\mathbf{c}_i$  to obtain a feature value  $o_i$  for this filter. (We note that we use multiple filter heights and redundant filters of each height.) Finally, all  $o_i$  are concatenated to compose a final feature vector  $\mathbf{o}$  for each instance. This is run through a softmax layer to induce a probability distribution over the output space. Typically this model is trained by minimizing the cross-entropy (or some other) loss via back-propagation (Rumelhart, Hinton, and Williams 1988). **Figure 1** provides a schematic illustrating a toy realization of this model. For more details, see (Kim 2014; Zhang and Wallace 2015).

The above model for sentence classification can easily be generalized for document classification. In particular, we adopt the hierarchical approach described by Zhang et al. (2016), in which one first applies the above set of operations to each sentence comprising a document, and then sums these to induce a global representation, which is in turn fed through a softmax layer to obtain a final prediction.

### Active Learning

We consider a *pool-based* AL scenario (Zhu et al. 2008; Tong and Koller 2001), in which there exists a small set of labeled data  $\mathcal{L}$  and a large pool of available unlabeled data  $\mathcal{U}$ . The task for the learner is to draw examples to be labeled from  $\mathcal{U}$  cleverly, so as to maximize classifier performance. These selections, or *queries*, are typically made in a greedy fashion; an informativeness measure is used to score all candidate instances in the pool, and the instance maximizing this measure is selected.

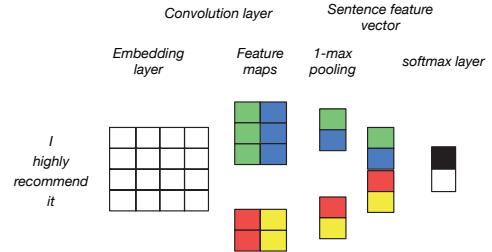


Figure 1: Illustrative schematic of a CNN for text classification. In this example, there are 2 feature maps with size 3, and 2 feature maps with size 2.

The key to developing AL strategies is designing a good informativeness measure. Let  $\mathbf{x}^*$  be the most informative instance according to a *query strategy*  $\phi(\mathbf{x}_i; \theta)$ , or function used to evaluate each instance  $\mathbf{x}_i$  in the unlabeled pool  $\mathcal{U}$  conditioned on the current set of parameter estimates  $\theta$ . We can define the following instance selection protocol:

$$\mathbf{x}^* = \operatorname{argmax}_{\mathbf{x}_i \in \mathcal{U}} \phi(\mathbf{x}_i; \theta) \quad (1)$$

For CNNs,  $\theta$  includes word embedding parameters  $\mathbf{E}$ , convolution layer parameters  $\mathbf{C}$ , and softmax layer parameters  $\mathbf{W}$ .

Many querying strategies have been proposed in the literature (Settles 2010). Our aim here is to ascertain whether AL works better in the case of neural models when one explicitly considers representation learning (i.e., focussing on  $\mathbf{E}$ ); we selected the following three general baseline approaches because they enable us to explore this question directly.

**Random sampling.** This strategy is equivalent to standard (or ‘passive’) learning; here the training data is simply an i.i.d. sample from  $\mathcal{U}$ .

**Uncertainty sampling.** Perhaps the most commonly used query strategy is *uncertainty sampling* (Lewis and Gale 1994; Tong and Koller 2002; Zhu et al. 2008; Ramirez-Loaiza et al. 2016), in which the learner requests labels for instances about which it is least certain wrt. categorization.

Uncertainty sampling can be instantiated in many ways, depending on the underlying classification model. A general uncertainty sampling variant uses entropy (Shannon 2001) as an uncertainty measure, defining  $\phi(\mathbf{x}_i; \theta)$  as:

$$-\sum_k P(y_i = k | \mathbf{x}_i; \theta) \log P(y_i = k | \mathbf{x}_i; \theta) \quad (2)$$

where  $k$  indexes all possible labels. Entropy-based uncertainty sampling often performs well (Settles 2010).

**Expected Gradient Length (EGL).** This AL strategy aims to select instances expected to result in the greatest change to the current model parameter estimates when their labels are revealed (or provided) (Settles and Craven 2008). The intuition is that one can view the magnitude of the resultant gradient as the value of purchasing a label; if this cost is small, then the label did not provide much new information. If the true class for a given instance were known, the gradient could be directly calculated under this assignment. But in practice this is unknown, and so the expectation is taken

by marginalizing over the gradients calculated conditioned on possible class assignments, scaled by current model estimates of the posterior probabilities of said assignments.

## AL with CNNs/Embeddings

We now introduce our proposed AL strategy for text classification with embeddings. This is based on the EGL method described above. In gradient-based optimization for neural models, the training gradient back-propagated to a set of model parameters given label  $y_i$  for instance  $\mathbf{x}_i$  may be viewed as a measure of change imparted by example  $i$  for those parameters. Thus the learner should request the label for an instance expected to produce a large magnitude training gradient. If this gradient is taken with respect to all model parameters (distributed over all layers), then this is a straight-forward instantiation of EGL. Past work on EGL (involving linear models) adopted exactly this approach: the expected change to model parameters was evaluated over the entire set of parameters in  $\theta$ . By contrast, we propose *explicitly selecting examples that are likely to affect the representation-level parameters (i.e., the embeddings)*.

Formally, let  $\nabla J(\mathcal{L}; \theta)$  be the gradient of the objective function  $J$  with respect to the model parameters  $\theta$ , where  $J$  is the cost function. Further, let  $\nabla J(\mathcal{L} \cup \langle \mathbf{x}_i, y_i \rangle; \theta)$  be the new gradient that would be obtained by adding the training tuple  $\langle \mathbf{x}_i, y_i \rangle$  to  $\mathcal{L}$ . Because the true label  $y_i$  will be unknown, we take an expectation over possible class assignments  $k$ . More precisely, we can calculate  $\phi(\mathbf{x}_i; \theta)$  as:

$$\sum_k P(y_i = k | \mathbf{x}_i; \theta) \|\nabla J(\mathcal{L} \cup \langle \mathbf{x}_i, y_i = k \rangle; \theta)\| \quad (3)$$

where  $\|\mathbf{r}\|$  denotes the Euclidean norm of  $\mathbf{r}$ . Note that at query time  $\|\nabla J(\mathcal{L}; \theta)\|$  should be near zero, assuming  $J$  converged during the previous iteration. Thus, we can approximate  $\nabla J(\mathcal{L} \cup \langle \mathbf{x}_i, y_i = k \rangle; \theta) \approx \nabla J(\langle \mathbf{x}_i, y_i = k \rangle; \theta)$  for efficiency.

This approach selects instances that are likely to most perturb all model parameters  $\theta$ . However, ‘deep’ neural architectures are distinguished by their multi-layered structure, which corresponds to a large set of features distributed across different layers in the architecture. This makes calculating the EGL computationally expensive. More importantly, it is arguably incoherent to jointly consider the expected change *at different layers* in the model. If we view lower levels in the model as learning to extract features, it makes little sense to jointly maximize expected change in these features *and* to the parameters of the final softmax layer that accepts these as input. Changes to the former will immediately change the implications of perturbing the latter.

Instead, we want to select unlabeled instances that can most improve the features learned by the model. Intuitively, it is paramount that the model learn good (discriminative) *representations*; these will feed forward through the network, in turn improving classification. In the context of sentence classification — in which instances comprise relatively few words — we propose a querying strategy that scores sentences using the maximum expected gradient over the words they contain. In the case of longer texts or documents (which contain many words), it is intuitive to strike

a balance between myopically selecting instances to maximize individual word gradients on the one hand, and considering the model’s overall uncertainty regarding the instance on the other. We next elaborate on the methods we propose for these two scenarios.

## Active Sentence Classification with CNNs

**EGL-word model.** For sentence classification, we adopt the following as our scoring function for sentence classification. For each instance (sentence) in  $\mathcal{U}$ , we take the expected gradient with respect to *only* the embeddings of its constituent words, selecting the example that maximizes this expected embedding gradient as our measure of informativeness. Intuitively, we use a max-over-words approach to adjust particular word embeddings that are discriminative for the task at hand. Formally, we define our  $\phi(\mathbf{x}_i; \theta)$  as:

$$\max_{j \in \mathbf{x}_i} \sum_k P(y_i = k | \mathbf{x}_i; \theta) \|\nabla J_{\mathbf{E}(j)}(\langle \mathbf{x}_i, y_i = k \rangle; \theta)\| \quad (4)$$

Where we denote by  $\nabla J_{\mathbf{E}(j)}$  the gradient of  $J$  with respect to the embedding of word  $j$  ( $j$  ranges over the words in  $\mathbf{x}_i$ ). Note that the gradient is only taken for each word in the instance  $\mathbf{x}_i$ . The gradients for embeddings corresponding to words not in  $\mathbf{x}_i$  are 0 and can thus be ignored; this is a computational boon because instances tend to be sparse. Another straightforward strategy to measure the informative of a sentence is to replace the ‘max’ operator in equation 4 with the average operation. That is, instead of choosing the word with the maximum expected gradient, we can average on the expected gradients of all the words in the sentence. But this method does not work as well as *EGL-word*. We attribute this to the fact that in a short sentence, most words are not relevant to the label of the sentence.

**EGL-sm model.** Whereas *EGL-word* focuses on parameters associated with the lowest level in the model (Figure 1), we also consider the other extreme in sentence classification tasks: taking the gradient with respect to only the final softmax layer parameters  $\mathbf{W}$ . In this case  $\phi(\mathbf{x}_i; \theta)$  becomes:

$$\sum_k P(y_i = k | \mathbf{x}_i; \theta) \|\nabla J_{\mathbf{W}}(\langle \mathbf{x}_i, y_i = k \rangle; \theta)\| \quad (5)$$

where  $J_{\mathbf{W}}$  denotes the gradient wrt. the softmax layer.

## Active Document Classification with CNNs

**EGL-word-doc model.** For longer text classification tasks, we modify the above *EGL-word* variant in a few key ways. First, we normalize the gradient of each word by dividing it by its frequency in the document. This is because in longer texts there exist many ‘stop words’ such as ‘the’, and their gradients dominate if occurrence counts are ignored, since there are more branches flowing back to these words during back-propagation. Accounting for term frequencies in the gradient calculation mitigates this issue. Second, rather than exclusively relying on the single word with the largest gradient to score documents, we sum over the (frequency-normalized) gradients corresponding to the top  $k$  words. The number of top words ( $k$ ) is a hyper-parameter and will depend on the average document length in a given corpus. We

	CR	MR	Subj
Positive	2406	5331	5000
Negative	1367	5331	5000
Avg. word count	19	20	23

Table 1: Statistics of sentence datasets.

refer to this method as *EGL-word-doc* for document classification.<sup>1</sup>

**EGL-Entropy-Beta model.** In addition to the above modifications, we extend our approach for longer text classification to jointly consider: (1) the expected updates to word gradients (for words in the instance); and (2) the current uncertainty regarding the instance. For the former, we use *EGL-word-doc* (modified as described above), and for the latter we use entropy (Equation 2). We denote the entropy score by  $\phi_{\text{Entropy}}$  and the *EGL-word-doc* score by  $\phi_{\text{EGL-word-doc}}$ . We interpolate these to form a composite document score.

These scores are on incomparable scales, so we normalize them by transforming them into percentiles.  $\mathcal{P}(i, \mathcal{U})$  is used to denote the percentile of the score of a given instance among a pool of instances  $\mathcal{U}$ . For example,  $\mathcal{P}(i, \mathcal{U})=87\%$  indicates that 87% of the instances in  $\mathcal{U}$  are smaller than  $i$ . To encode the relative entropy score of a given instance in  $\mathcal{U}$ , we use  $\mathcal{P}(\phi_{\text{Entropy}}(i), \{\phi_{\text{Entropy}}(j) : j \in \mathcal{U}\})$ . We can now define our composite, interpolated scoring function which considers feature learning and output certainty jointly:

$$\phi_t(i) = \gamma_t \cdot \mathcal{P}(\phi_{\text{Entropy}}(i), \{\phi_{\text{Entropy}}(j) : j \in \mathcal{U}\}) + (1 - \gamma_t) \cdot \mathcal{P}(\phi_{\text{EGL-word-doc}}(i), \{\phi_{\text{EGL-word-doc}}(j) : j \in \mathcal{U}\})$$

We treat the interpolation parameter  $\gamma_t$  — constrained to be between 0 and 1 — as a random variable with a temporal dependence ( $t$  indexes time, or AL iteration). Intuitively, we assume that at the outset of AL, the model should pay relatively more attention to learning discriminative representations of words. As learning progresses, focus should shift toward the higher-level uncertainty-based score. To realize this intuition, we assume  $\gamma_t \sim \text{Beta}(\alpha, \beta_t)$ . We decrease  $\beta_t$  linearly over time (AL iterations), which has the desired effect of increasing the expectation of  $\gamma_t$ , in turn increasing the attention paid to the document level entropy score. We found that drawing  $\gamma_t$  from a distribution yields smoother performance compared to setting it deterministically.

## Experimental Setup

We report results on three sentence datasets and three document datasets. **Tables 1** and **2** provide key statistics for each dataset. We briefly describe each dataset below and refer the reader to the source citations for additional details.

<sup>1</sup>Experiments applying the same variant of *EGL-word* used for sentence classification does not perform as well for longer texts. *EGL-sm* model also performs much worse than the other methods in the document classification tasks, so we do not report their results.

	MR	MuR	DR
Positive	1000	1000	23649
Negative	1000	1000	30254
$l_{\text{sen}}$	21.2	16.8	15.2
$l_{\text{doc}}$	32.6	7.5	4.1

Table 2: Statistics of document datasets.  $l_{\text{doc}}$  denotes the average sentence length in words, and  $l_{\text{doc}}$  denotes the average number of sentences per document.

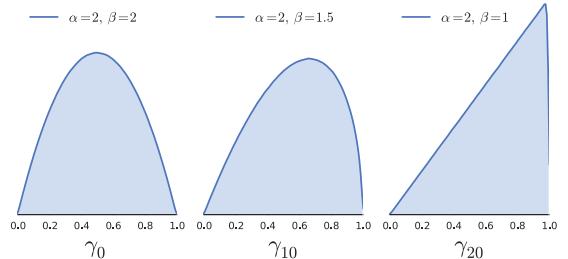


Figure 2: Beta distributions over  $\gamma_t$  at  $t=0, t=10, t=20$ .

## Sentence Datasets

**CR:** positive / negative product reviews (Hu and Liu 2004).<sup>2</sup>

**MR:** positive / negative movie reviews (Pang and Lee 2005).

**Subj:** subjective / objective sentences (Pang and Lee 2004).<sup>3</sup>

**Document Datasets** positive / negative classification tasks  
**MR:** (Longer) movie reviews (Pang and Lee 2004)<sup>4</sup>.

**MuR:** Music reviews (Blitzer et al. 2007).<sup>5</sup>

**DR:** Doctor reviews (Wallace et al. 2014).

## Model Configuration

We used standard pre-trained word2vec-induced vectors<sup>6</sup> to initialize  $\mathbf{E}$ . As per Zhang and Wallace (2015), we used three filter heights (3, 4, 5). For sentence and document classification tasks, we used 50 and 100 filters of each size, respectively.<sup>7</sup> Given our goal to explore AL strategies appropriate for neural architectures (particularly CNNs), rather than to maximize absolute CNN performance for new state-of-art results, we did not tune these hyperparameters.

We performed 20 rounds of batch active learning. At the outset, we provided all learners with the same 25 instances (sampled i.i.d. at random). In subsequent rounds, each learner was allowed to select 25 instances from  $\mathcal{U}$  ac-

<sup>2</sup>[www.cs.uic.edu/liub/FBS/sentiment-analysis.html](http://www.cs.uic.edu/liub/FBS/sentiment-analysis.html)

<sup>3</sup>MR and Subj datasets are available at: <http://www.cs.cornell.edu/people/pabo/movie-review-data/>.

<sup>4</sup>Both MR datasets can be found online at the same URL.

<sup>5</sup><http://www.cs.jhu.edu/~mdredze/datasets/sentiment/>

<sup>6</sup><https://code.google.com/archive/p/word2vec/>

<sup>7</sup>We used more filters for document classification tasks because we expect more diversity in longer pieces of text, but we found that the performance was not sensitive to this choice in any case.

cording to their respective querying strategies. These examples were added to  $\mathcal{L}$ , and the models were retrained.

For *EGL-word-doc* and *EGL-Entropy-Beta* in document classification, the number of top words  $k$  used to calculate the score for each document was set to 3, 2 and 30 respectively for MuR, DR, and MR datasets. For *EGL-Entropy-Beta*, we fixed  $\alpha = 2$  and initialized  $\beta = 2$  as well, which implies a roughly equal weight on embedding and uncertainty scores. We then decreased  $\beta_t$  linearly with iterations  $t$ . Thus  $\gamma_t$  is expected to increase over time, ascribing more weight to the entropy score. For reference, **Figure 2** provides illustrative empirical distributions used for  $\gamma$  at three time points during AL. To reiterate, our goal was to shift from initially paying equal attention to the representation learning and instance uncertainty criteria, to increasingly focusing on the latter (document-level uncertainty) as time progresses.

We evaluated performance by calculating accuracy (classes are fairly balanced) on a held-out test set after each round. For all but one dataset we repeated this entire AL process 10 times, using test sets generated via 10-fold CV. The exception was the doctor reviews ('DR') dataset, which is comparatively large; we therefore used a single big test set in this case. We replicated all experiments 5 times for all train/test splits, for all datasets, to account for variance. We estimated parameters by Adadelta (Zeiler 2012), tuning  $\mathbf{E}$  in back-propagation to induce discriminative embeddings.

## Results and Discussion

We now report results. For sentence classification, we use the simple variant of our method (*EGL-word*) which is more appropriate for short texts (since it is ultimately a max-operator over expected gradients for individual words). For document classification, we also use the interpolated method, which considers expected gains both with respect to feature learning and in terms of instance-level uncertainty reduction. This method is more appropriate for longer texts.

### Sentence Classification Results

**Figure 3** reports learning curves on the three sentence datasets. The proposed *EGL-word* active learning method outperforms baseline approaches, performing especially well on sentiment analysis tasks (MR and CR). We believe this is due to our model rapidly learning more discriminative representations of words with opposing polarities.

To further illustrate this point, Figure 3's bottom row provides plots displaying the Euclidean distances between selected pairs of word embeddings induced using different AL strategies. In the customer review (CR) dataset, for example, we consider the embeddings of words 'good' vs. 'bad' and see that *EGL-word* quickly pushes these embeddings apart. Similarly, on the movie review (MR) dataset, 'fun' and 'boring' are rapidly separated in embedding space. The subjectivity (Subj) detection task is less clear-cut. Here we picked words 'amusing' and 'their', because 'amusing' strongly indicates subjectivity, while 'their' is plainly neutral. As expected, *EGL-word* quickly pushes these apart, though less rapidly than with the sentiment tasks.

**Table 3** reports Area Under Curve (AUC) scores for each

	EGL-word	Entropy	Random	EGL-sm
MR	<b>0.707</b>	0.690	0.681	0.667
CR	<b>0.743</b>	0.732	0.720	0.674
Subj	<b>0.856</b>	0.840	0.839	0.785

Table 3: Area Under (learning) Curve (AUC) scores on sentence classification datasets; bold indicates best results.

	E-E-B	EGL-word-doc	Entropy	Random
MR	<b>0.725</b>	0.719	0.719	0.704
DR	<b>0.893</b>	0.889	0.877	0.878
MuR	<b>0.736</b>	0.718	0.725	0.726

Table 4: Area Under (learning) Curves (AUC) scores on the three document datasets. E-E-B refers to *EGL-Entropy-Beta*.

learning curve from 25-500 labeled instances using trapezoidal rule (Süli and Mayers 2003). We normalize AUC by the maximum possible for the range:  $(500 - 25) * 1 = 475$ .

### Document classification results

**Figure 4** displays learning curves achieved on the document classification datasets, and **Table 4** reports the corresponding AUC scores achieved by each method on each dataset. Overall, the *EGL-Entropy-Beta* outperforms other methods, demonstrating the value of explicitly selecting examples likely to improve representation level parameters.

Results using the simple variant of *EGL-word-doc* are mixed. In general it outperforms baselines only during the first several iterations of AL, but is later outperformed by entropy-based sampling. Our intuition here is that narrowly focusing on improving feature representations provides early gains, but longer texts require attention to be shifted to instance-level uncertainty. And indeed, the proposed *EGL-Entropy-Beta* method consistently performs more robustly, and tends to realize the best of both worlds, achieving rapid gains but also generally maintaining dominance over all AL iterations.

Similar to **Figure 3**'s bottom row for sentence tasks, **Figure 4**'s bottom row shows for document tasks how distances between selected word embeddings grow as more examples are collected. *EGL-word-doc* and *EGL-Entropy-Beta* consistently push the representations for the selected polar word-pairs apart more rapidly than other methods. However, recall that the *EGL-Entropy-Beta* method differs from *EGL-word-doc* in interpolating entropy along with expected updates to word gradients. As a result, we observe that *EGL-Entropy-Beta* method tends to shift from rising with *EGL-word-doc* at the start of learning, while later merging with the distances achieved by the *Entropy* method as learning progresses. *EGL-Entropy-Beta* thus strikes a balance between this and refining the parameters at higher levels in the model, as evidenced by the superior classification performance seen in the top row of **Figure 4**. Maintaining a narrow focus on embeddings only ultimately results in comparatively poor performance in the case of document classification.

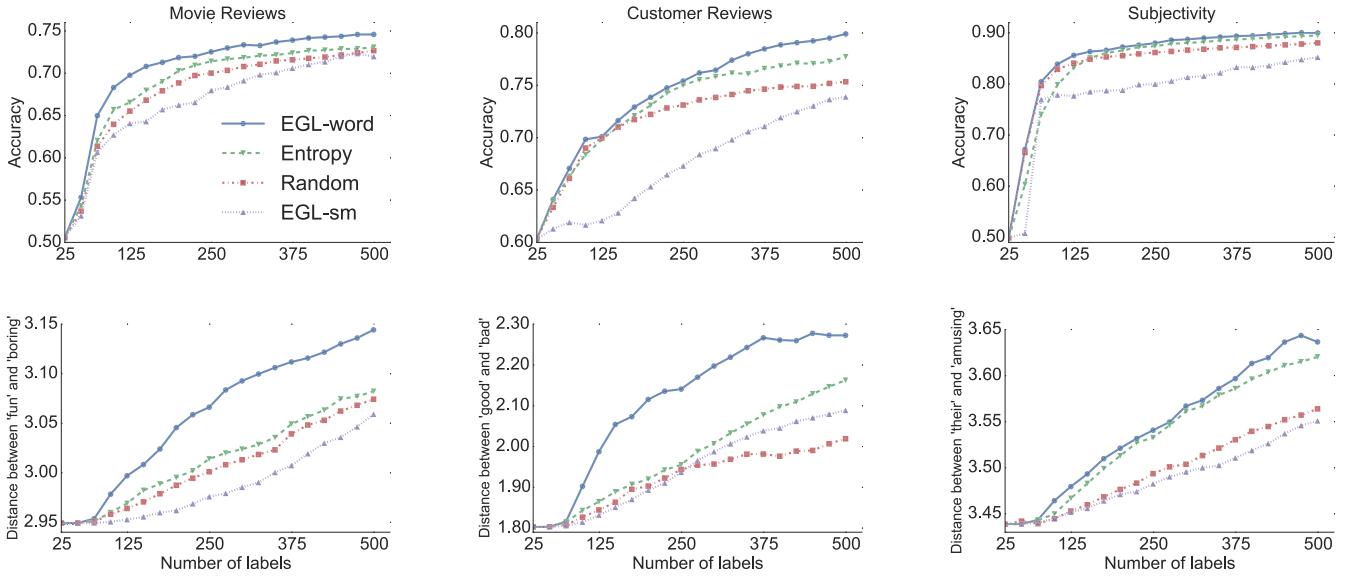


Figure 3: Results on the three sentence classification datasets. **Top row:** number of labels versus accuracy. **Bottom row:** number of labels versus the distance between tuned embeddings for selected pairs of informative words (with opposite polarity) for each dataset. The scale in this case, which captures the Euclidean distance in the embedding space, has only relative meaning.

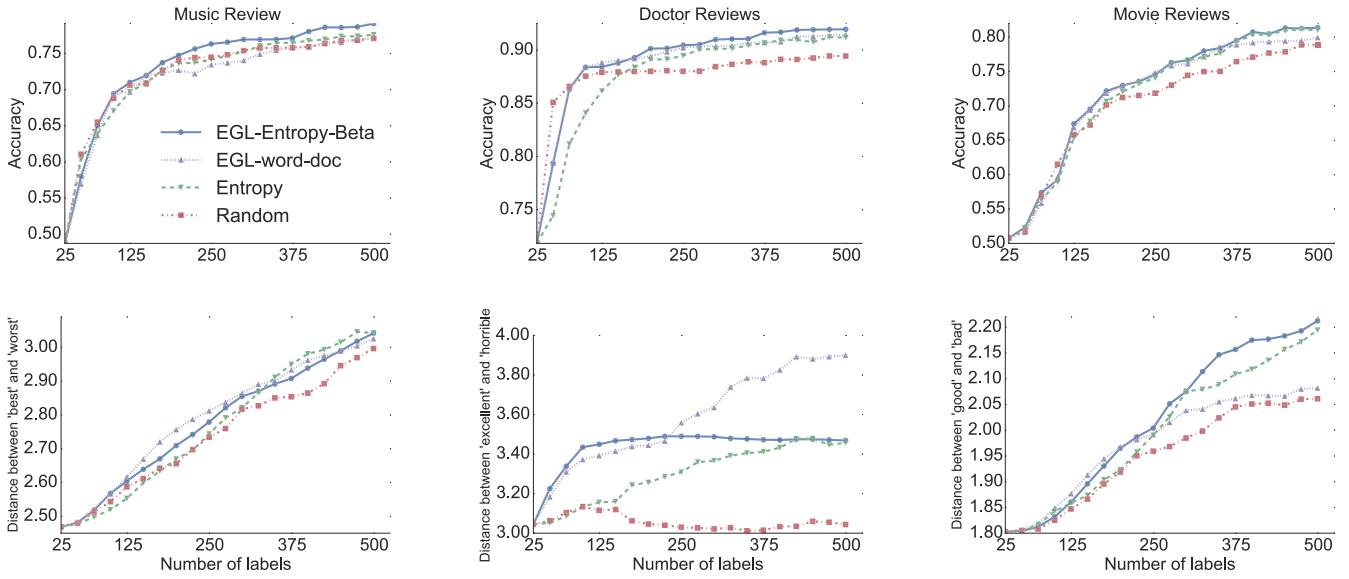


Figure 4: Results on the three document datasets. **Top row:** number of labels versus accuracy. **Bottom row:** number of labels versus the distance between tuned embeddings for selected pairs of informative words (with opposite polarity) in each task.

Methods that explicitly consider representation/embedding parameters more quickly push discriminative word vectors apart. Intuitively, the distances between the contrasting word-pairs increases quickly with both of the proposed EGL methods. However, recall that the *EGL-Entropy-Beta* method differs from *EGL-word-doc* in interpolating entropy along with expected updates to word gradients. As a result, we observe that *EGL-Entropy-Beta* method tends to shift from rising with *EGL-word-doc* at the start of learning, while later merging with the distances achieved by the *Entropy* method as learning progresses. This transition corresponds to first focusing on embeddings, and then later shifting emphasis to the entropy criterion.

## Conclusions

The importance of *representation learning* (Bengio 2009) with neural models motivates exploring new, representation-based active learning (AL) approaches with neural models. To this end, we proposed a new AL strategy for CNNs that is specifically designed to quickly induce discriminative, task-specific representations (word embeddings), thus improving classification. We showed that this approach outperforms baseline AL strategies across sentence and document classification datasets considered, and that such discriminative word embeddings can be rapidly induced.

We believe that these encouraging results will help to stimulate further research on active learning tailored to deep/hierarchical architectures. Our own future work will include generalize the similar AL strategies to other neural models such as recurrent neural network and improving the modeling strategy for  $\gamma_t$  (the parameter governing relative emphasis on representation vs. instance-level uncertainty), perhaps based on reinforcement learning. We also envision augmenting the model to optimize instance selection in terms of refining additional intermediate layer representations in deeper networks.

## Acknowledgments

This research was supported in part by IMLS grant RE-04-13-0042-13 and the Foundation for Science and Technology, Portugal (FCT), through contract UTAP-EXPL/EEIESS/0031/2014. Any opinions, findings, and conclusions or recommendations expressed by the authors do not express the views of any of the supporting funding agencies.

## References

- Bengio, Y. 2009. Learning deep architectures for ai. *Foundations and trends in Machine Learning* 2(1):1–127.
- Blitzer, J.; Dredze, M.; Pereira, F.; et al. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, volume 7, 440–447.
- Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 168–177. ACM.
- Johnson, R., and Zhang, T. 2014. Effective use of word order for text categorization with convolutional neural networks. *arXiv preprint arXiv:1412.1058*.
- Kim, Y. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.
- Lewis, D. D., and Gale, W. A. 1994. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, 3–12. Springer-Verlag New York, Inc.
- McCallumzy, A. K., and Nigam, K. 1998. Employing em and pool-based active learning for text classification. In *Proc. International Conference on Machine Learning (ICML)*, 359–367. Cite-seer.
- Pang, B., and Lee, L. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, 271. Association for Computational Linguistics.
- Pang, B., and Lee, L. 2005. 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*, 115–124. Association for Computational Linguistics.
- Ramirez-Loaiza, M. E.; Sharma, M.; Kumar, G.; and Bilgic, M. 2016. Active learning: an empirical study of common baselines. *Data Mining and Knowledge Discovery* 1–27.
- Rumelhart, D. E.; Hinton, G. E.; and Williams, R. J. 1988. Learning representations by back-propagating errors. *Cognitive modeling* 5(3):1.
- Settles, B., and Craven, M. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the conference on empirical methods in natural language processing*, 1070–1079. Association for Computational Linguistics.
- Settles, B. 2010. Active learning literature survey. *University of Wisconsin, Madison* 52(55-66):11.
- Shannon, C. E. 2001. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review* 5(1):3–55.
- Süli, E., and Mayers, D. F. 2003. *An introduction to numerical analysis*. Cambridge university press.
- Tong, S., and Koller, D. 2001. Support vector machine active learning with applications to text classification. *Journal of machine learning research* 2(Nov):45–66.
- Tong, S., and Koller, D. 2002. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research* 2:45–66.
- Wallace, B. C.; Small, K.; Brodley, C. E.; and Trikalinos, T. A. 2010. Active learning for biomedical citation screening. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 173–182. ACM.
- Wallace, B. C.; Paul, M. J.; Sarkar, U.; Trikalinos, T. A.; and Dredze, M. 2014. A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. *Journal of the American Medical Informatics Association* 21(6):1098–1103.
- Zeiler, M. D. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.
- Zhang, Y., and Wallace, B. 2015. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*.
- Zhang, Y.; Marshall, I.; and Wallace, B. C. 2016. Rationale-augmented convolutional neural networks for text classification. *arXiv preprint arXiv:1605.04469*.
- Zhang, Y.; Roller, S.; and Wallace, B. 2016. Mgnc-cnn: A simple approach to exploiting multiple word embeddings for sentence classification. *arXiv preprint arXiv:1603.00968*.
- Zhu, J.; Wang, H.; Yao, T.; and Tsou, B. K. 2008. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, 1137–1144. Association for Computational Linguistics.