

# Beyond Word Embeddings: Dense Representations for Multi-modal Data

Luis Armona<sup>1,2</sup>, Ralph Edezhath<sup>2</sup>, José P. González-Brenes,<sup>2</sup>

<sup>1</sup>Department of Economics, Stanford University,

<sup>2</sup>Chegg, Inc.

larmona@stanford.edu, {redezhath, jgonzalez}@chegg.com

## Abstract

Methods that calculate dense vector representations for text have proven to be very successful for knowledge representation. We study how to estimate dense representations for multi-modal data (e.g., text, continuous, categorical). We propose Feat2Vec as a novel model that supports supervised learning when explicit labels are available, and self-supervised learning when there are no labels. Feat2Vec calculates embeddings for data with multiple feature types, enforcing that all embeddings exist in a common space. We believe that we are the first to propose a method for learning self-supervised embeddings that leverage the structure of multiple feature types. Our experiments suggest that Feat2Vec outperforms previously published methods, and that it may be useful for avoiding the cold-start problem.

## 1 Introduction

Informally, in machine learning a *dense representation*, or *embedding* of a vector  $\vec{x} \in \mathbb{R}^n$  is another vector  $\vec{\beta} \in \mathbb{R}^r$  that has much lower dimensionality ( $r \ll n$ ) than the original representation. In general, we consider two kind of models that produce embeddings: **(i) supervised methods**, like matrix factorization, calculate embeddings that are highly tuned to a prediction task. For example, in the Netflix challenge, movie identifiers are embedded to predict user ratings. On the other hand, **(ii) self-supervised methods** (sometimes referred to as unsupervised methods) are not tuned for the prediction task they are ultimately used for. For example, word embedding algorithms such as Word2Vec (Mikolov et al. 2013b) are self-supervised. These algorithms are typically evaluated by analogy solving, or sentiment analysis (Le and Mikolov 2014), even though their loss functions are not tuned for either of these tasks.

Throughout this paper, we refer to any data that has different types of information as multi-modal. We propose Feat2Vec as a novel method that embeds multi-modal data, such as text, images, numerical or categorical data, in a common vector space—for both supervised and self-supervised scenarios. We believe that this is the first self-supervised algorithm that is able to calculate embeddings from data with multiple feature types. Consider the non-trivial work that

was required to extend a model like Word2Vec to support additional features. The authors of the seminal Doc2Vec (Le and Mikolov 2014) paper needed to design both a new neural network and a new sampling strategy to add a single feature (document ID). Feat2Vec is a general method that allows calculating embeddings for any number of features.

## 2 Feat2Vec

In this section we describe how Feat2Vec learns embeddings of feature groups.

### 2.1 Model

Feat2Vec operates on multi-modal data that naturally group into different types. For example, we may group individual words into a single “text” feature group; or we may group individual pixels into a single “image” feature group. We define an entity to be a particular value or realization of a feature group.

Feat2Vec predicts a target output  $\tilde{y}$  from observations that can be interpreted as a list of feature groups:

$$\vec{x} = \langle \vec{x}_{\vec{\kappa}_1}, \vec{x}_{\vec{\kappa}_2}, \dots, \vec{x}_{\vec{\kappa}_n} \rangle = \langle \vec{g}_1, \vec{g}_2, \dots, \vec{g}_n \rangle \quad (1)$$

Each of the  $n$  feature groups  $\vec{g}_i$  is defined a priori by  $\vec{\kappa}_i$ , which indexes the dimensions in  $\vec{x}$  that belong to the  $i$ -th group. Thus, each observation is constructed from a partition  $\vec{\kappa}$  of raw features  $\vec{x}$ : The number of dimensions of each group may vary, but all of the feature groups are embedded to the same space via their *feature extraction function*  $\phi_i$ . A feature extraction function  $\phi_i$  acts on the features in  $\kappa_i$ . Other features interact with  $\vec{x}_{\vec{\kappa}_i}$  only via the output of  $\phi_i$ . These functions learn how to transform each input (feature group) into an  $r$ -dimensional embedding in  $\mathbb{R}^r$ . More formally, we define our model for a target  $y$  as follows:

$$\tilde{y}(\vec{x}, \vec{\phi}, \vec{\theta}) = \omega \left( \sum_{i=1}^n \sum_{j=i}^n \overbrace{\phi_i(\vec{x}_{\vec{\kappa}_i}; \vec{\theta})}^{\text{group } i \text{ embedding}} \cdot \overbrace{\phi_j(\vec{x}_{\vec{\kappa}_j}; \vec{\theta})}^{\text{group } j \text{ embedding}} \right) \quad (2)$$

Here,  $\omega$  is an activation (link) function. Intuitively, the dot product ( $\cdot$ ) returns a scalar that measures the (dis)similarity between the embeddings of the feature groups. A simple implementation of  $\phi_i$  is a linear fully-connected layer, where

the output of the  $r$ -th entry is:

$$\phi_{i,r}(\vec{x}_i; \vec{\theta}_r)_r = \sum_{a=1}^{d_i} \theta_{r_a} x_{i_a} \quad (3)$$

where  $\vec{\theta}_r$  are learned weights. Other functional forms of  $\phi_i$  we have explored in our experiments include:

- A convolutional neural network that transforms a text sequence into an  $r$ -dimensional embedding. This function is able to preserve information in the order of a text sequence, unlike other methods for transforming text to numerical values such as bag-of-words.
- A deep, fully connected neural network that projects a scalar, such as average rating, into an  $r$ -dimensional embedding space. This extraction function, by treating the input as numerical instead of categorical, requires inputs close in value (e.g. 4.01 and 4.02 star rating) to have similar embeddings.

Any neural function that outputs an  $r$ -dimensional vector could be plugged into our framework as an extraction function  $\phi$ . Although in this work we have only used datasets with natural language, categorical and numerical features, it would be straightforward to incorporate other types of data—e.g. images or audio—with the appropriate feature extraction functions, enabling the embedding of multi-modal features in the same space.

Figure 1 compares existing factorization methods with our novel model. Figure 1a shows a Factorization Machine (Rendle 2010) that embeds each feature. Figure 1b shows Feat2Vec with two feature groups: the first group only has a single feature which is projected to an embedding (just like matrix factorization); but the second group has multiple features, which are together projected to a single embedding. Figure 1c shows an approach of using neural networks within factorization machines that has been proposed multiple times (Dziugaite and Roy 2015; Guo et al. 2017). It replaces the dot product of factors with a learned neural function, which has been shown to improve predictive accuracy. The caveat of this architecture is that is no longer possible to interpret the embeddings as latent factors related to the target task.

Our work in Feat2Vec significantly expands previously published models. For example, Principal Component Analysis (PCA) is a common method to learn embeddings of individual dimensions of data. Although structured regularization techniques exist (Jenatton, Obozinski, and Bach 2010), it is not obvious how to combine different dimensions in PCA when we are interested in treating a subset of dimensions as a single group, such as words in an item’s description. In traditional factorization methods, item identifiers are embedded and thus need to be observed during training (i.e., cold-start problem). Feat2Vec can learn an embedding from an alternative characterization of an item, such as a textual description. Feat2Vec extends Factorization Machine (Rendle 2010), in that we allow calculating embeddings for feature groups and we use feature extraction functions. StarSpace (Wu et al. 2018) introduces feature groups (called entities in their paper), but constrains them

to be “bags of features,” is supervised, and limited to two feature types. In contrast, Feat2Vec allows continuous features, and does not require the user to assign one feature as a specific target. Additionally, we propose a self-supervised learning algorithm.

## 2.2 Supervised Learning from Data

We can learn the parameters of a Feat2Vec model  $\theta$  by minimizing a supervised loss function  $\mathcal{L}_{\text{sup}}$ :

$$\arg \min_{\theta} \sum_{\vec{x} \in \mathbf{X}} \mathcal{L}_{\text{sup}}(y(\vec{x}), \tilde{y}(\vec{x}; \Phi, \theta)) \quad (4)$$

Here,  $\mathbf{X}$  is the training dataset,  $y(\vec{x})$  is the true target value for an observation  $\vec{x} \in \mathbf{X}$ , and  $\tilde{y}(\vec{x})$  is its estimated value (using Equation 2);  $\theta$  represents the parameters learned during training (i.e. the parameters associated with the extraction functions  $\phi_i$ ). For labeling and classification tasks, we optimize the logistic loss.

It is straightforward to optimize Equation 4 directly. Typically, this loss will be for a prediction task, and involve at least two feature groups—one of the feature groups is the target label that we want to predict, such as rating, and the other group(s) is the input from which we want to make the prediction, such as review text.

Consider a target feature group  $g_i$  that exists in a high dimensional space. This could be a one-hot encoding of a categorical variable with a large number of possible values. In such a scenario, it is overwhelmingly costly to feed the model all possible negative labels. In these cases, we may instead use a binary classifier with implicit sampling (Dyer 2014). In implicit sampling, instead of using all of the possible negative labels, one simply samples a fixed number ( $k$ ) from the set of possible negative labels for each positively labelled record.

## 2.3 Self-Supervised Learning From Data

We now discuss how Feat2Vec can be used to learn embeddings in a self-supervised setting with no explicit target for prediction.

The training dataset for a Feat2Vec model consists of the observed data. In natural language, these would be documents written by humans, along with document metadata. Since self-supervised Feat2Vec will require positive and negative examples during training, we supply unobserved data as negative examples. This sampling procedure is analogous to the Word2Vec algorithm, which samples a negative observation from a noise distribution  $\mathcal{Q}_{w2v}$  that is proportional to the empirical frequency of a word in the training data. Unlike Word2Vec, we do not constrain features types to be words. Instead, by leveraging the feature groups  $\mathbf{K}$  of the data (Equation 2), the model can reason on more abstract entities in the data. For example, in our experiments on a movie dataset, we define a “genre” feature group, where we group non-mutually exclusive indicators for movie genres, including comedy, action, and drama films.

We start with a training dataset  $\mathbf{X}^+$  of records with  $n$  feature groups that we assume to have a positive label. For each

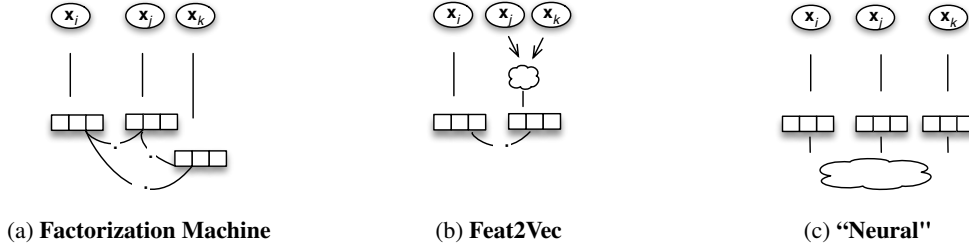


Figure 1: **Network architectures for factorization models.** The white clouds (☁) represent deep layers, for example a convolutional network for text features, while the dots (·) denote dot products.

observed (positive) datapoint  $\vec{x}^+$ , we generate  $k$  observations with negative labels using the 2-step algorithm documented in Algorithm 1. We illustrate the intuition of this algorithm with an example. Consider that we have a dataset with three feature groups—entities for (i) a passage of text (bag of words), (ii) a restaurant identifier (categorical), and a sentiment (a continuous value). The positive examples are observations of the values seen in the training set; while negative examples are generated by modifying observations by substituting one of the entities randomly from another observation. For example, by substituting the passage of text, a negative example may result as: *(Ewww, gross!!!, Le Fancy Restaurant, 100)*.

**Algorithm 1** Implicit sampling algorithm for self-supervised Feat2Vec

```

1: function  $Q(\vec{x}^+; k, \alpha_1, \alpha_2)$ 
2:    $S^- \leftarrow \emptyset$ 
3:   for  $j \in \{1, \dots, k\}$  do
4:     Draw which feature group  $i \sim Q_1(\Phi, \alpha_1)$  to sample
5:     Draw a random entity  $\vec{g} \sim Q_2(X_{\kappa_i}, \alpha_2)$ 
6:      $\vec{x}^- \leftarrow \vec{x}^+$   $\triangleright$  Set identical to the positive sample
7:      $\vec{x}_{\kappa_i}^- \leftarrow \vec{g}$   $\triangleright$  substitute the  $i$ -th feature group with  $\vec{g}$ 
8:      $\mathbf{X}^- \leftarrow \mathbf{X}^- + \{\vec{x}^-\}$   $\triangleright$  append
9:   end for
10:  return  $\mathbf{X}^-$   $\triangleright$  Sampled negative observations from  $\vec{x}^+$ 
11: end function

```

More formally, our negative sampling generates observations with a negative label from an observation in the training dataset (that is assumed to have a positive label). First, it randomly selects the  $i$ -th feature group from a noise distribution  $Q_1(\cdot)$ . It then creates a negative observation that is identical to  $\vec{x}^+$ , except that its  $i$ -th feature group value is replaced by a value sampled from a noise distribution  $Q_2(\cdot)$ . In our application, we use the same class of noise distributions (flattened multinomial) for both levels of sampling, but this need not be the case. We now describe the noise distributions that we use.

**Sampling Feature Groups.** The function  $\text{params}$  calculates the complexity of a feature extraction function  $\phi_i$ . To sample a feature group, we choose a feature group  $\kappa_i$  from a multinomial distribution with probabilities proportional to a feature’s complexity. By complexity, we mean the number of parameters we learn that are associated with a feature group’s extraction function  $\phi_i$ . This sampling procedure

places more weight on features that have more parameters and thus are going to require more training iterations to properly learn. The sampling probabilities of each feature group are:

$$P_{Q_1}(i|\Phi, \alpha_1) = \frac{\text{params}(\phi_i)^{\alpha_1}}{\sum_{j=1}^{|\kappa|} \text{params}(\phi_j)^{\alpha_1}} \quad (5)$$

For categorical variables using a linear fully-connected layer, the complexity is simply proportional to the number of categories in the feature group. However, if we have multiple intermediate layers for some feature extraction functions (e.g., convolutional layers), these parameters should also be counted towards a feature group’s complexity. The hyper-parameter  $\alpha_1 \in [0, 1]$  flattens the distribution. When  $\alpha_1 = 0$ , the feature groups are sampled uniformly, and when  $\alpha_1 = 1$ , they are sampled exactly proportional to their complexity.

**Sampling Feature Group Values.** To sample an entity within a feature group  $\kappa_i$ , we use a similar strategy to Word2Vec and use the empirical distribution of values:

$$P_{Q_2}(\vec{g}|\mathbf{X}_{\kappa_i}, \alpha_2) = \frac{\text{count}(\vec{g})^{\alpha_2}}{\sum_{\vec{g}' \in \mathbf{X}_{\kappa_i}^+} \text{count}(\vec{g}')^{\alpha_2}}, \quad \alpha_2 \in [0, 1] \quad (6)$$

Here,  $\text{count}(\vec{g})$  is the number of times a value  $\vec{g}$  of feature group  $i$  appeared in the training dataset  $\mathbf{X}^+$ . Thus, an entity is sampled according to its empirical distribution—whether it is a single word, a continuous number, or a passage of text. Again,  $\alpha_2$  is simply a flattening hyperparameter.

This method will sometimes by chance generate negatively labeled samples that *do* exist in our sample of observed records. The literature offers two solutions: in the Negative Sampling of Word2Vec, duplicate negative samples are ignored (Dyer 2014). Instead, we account for the probability of random negative labels being identical to positively labeled data using *Noise Contrastive Estimation* (NCE) (Gutmann and Hyvärinen 2010).

**The Loss Function for Self-Supervised Learning** For our self-supervised learning of embeddings, we optimize a NCE loss function. We need to make sure that our loss function is well-behaved distribution that integrates to 1. This often requires a partition function  $Z_{\vec{x}}$  for each unique record type  $\vec{x}$ , which can be computationally costly and greatly increase the complexity of our model. Instead, we appeal to

the work of Mnih and Teh (2012), who show that in models with many parameters setting the  $Z_{\vec{x}} = 1$  does not change the performance of the model. The intuition is that if the underlying model has enough free parameters, it will effectively learn the probabilities itself, since systemic under/over prediction of probabilities will result in penalties on the loss function.

We adjust the structural statistical model of Equation 2 to handle self-supervised learning. Because we only deal with a dichotomous label (positive or negative example), we restrict our attention to the logistic link function. Written explicitly, this is:

$$\tilde{p}(y = 1|\vec{x}, \vec{\Phi}, \vec{\Theta}) = \frac{\exp(\tilde{y}(\vec{x}|\vec{\Phi}, \vec{\Theta}))}{\exp(\tilde{y}(\vec{x}, \vec{\Phi}, \vec{\Theta})) + P_{\mathcal{Q}}(\vec{x}|\alpha_1, \alpha_2)} \triangleq \hat{s}(\vec{x}; \vec{\Theta}) \quad (7)$$

For notational convenience, we refer to this structural probability as  $\hat{s}(\cdot; \vec{\Theta})$ .  $P_{\mathcal{Q}}(x)$  denotes the probability of  $x$  under distribution  $\mathcal{Q}$ ; it accounts for the possibility of sampling an observation with a negative label that appears identical to one in the training data (with a positive label):

$$P_{\mathcal{Q}}(\vec{x}|\alpha_1, \alpha_2, X, \vec{x}^+) = P_{\mathcal{Q}_2}(\vec{x}_{\kappa_i}|X_{\kappa_i}, \alpha_2) \times P_{\mathcal{Q}_1}(\kappa_i | \text{params}(\phi_i)\}_{i=1}^n, \alpha_1) \quad (8)$$

We seek to optimize the parameters  $\vec{\Theta}$  of the feature extraction functions  $\vec{\phi}$ :

$$\arg \min_{\vec{\Theta}} \sum_{\vec{x}^+ \in X^+} \mathcal{L}_{\text{self}}(\vec{x}^+; \vec{\Theta}) \quad (9)$$

For this, we define our self-supervised loss as  $\mathcal{L}_{\text{self}}$  as:

$$- \sum_{\vec{x}^+ \in X^+} \left( \log(\hat{s}(\vec{x}^+; \vec{\Theta})) + \sum_{\vec{x}^- \sim \mathcal{Q}(\cdot|\vec{x}^+)}^k \log(1 - \hat{s}(\vec{x}^-; \vec{\Theta})) \right) \quad (10)$$

Note that the observations with positive labels come directly from the training data, and the observations with negative labels are sampled using Algorithm 1. With  $n$  feature groups, the loss function of self-supervised Feat2Vec can be shown to be a weighted average of the losses from  $n$  distinct supervised Feat2Vec models,<sup>1</sup> one for each feature group. In other words, it optimizes  $n$  multi-label classifiers, where each classifier is optimized for a different feature group. The intuition is as follows. In Algorithm 1, we first sample a target feature group  $g$  according to  $P_{\mathcal{Q}_1}$ , and then add the loss from a supervised Feat2Vec model for  $g$  to the total self-supervised loss. Thus,  $P_{\mathcal{Q}_1}$  determines the weights the self-supervised loss function assigns to each feature group.

### 3 Empirical Results

For all our experiments, we define a development set and a single test set which is 10% of the dataset, and a part of the development set is used for early stopping or validating hyper-parameters.

<sup>1</sup>The proof can be found in supplementary material at [http://web.stanford.edu/~larmona/feat2vec/supplemental\\_theorem.pdf](http://web.stanford.edu/~larmona/feat2vec/supplemental_theorem.pdf)

Table 1: Supervised Yelp rating prediction

	MSE	Improvement over MF
Feat2Vec	<b>0.480</b>	<b>69.2 %</b>
DeepCoNN	1.441	19.6 %
MF (Matrix Factorization)	1.561	-

### 3.1 Supervised Embeddings

We now address our working hypotheses for evaluating supervised embeddings. Since we only observe positive labels during supervised training, for each positive label in the test set we sample negative labels according to the label frequency.

For the regression task, we use mean squared error (MSE) as the evaluation metric. For our feature extraction function  $\phi$  for text, we use a Convolutional Neural Network (CNN) that has been shown to be effective for natural language tasks (Kalchbrenner, Grefenstette, and Blunsom 2014). Instead of tuning the hyper-parameters, we follow previously published guidelines (Zhang and Wallace 2017).

**Comparison with alternative CNN-based text factorization** We now compare with a method called DeepCoNN, a deep network specifically designed for incorporating text into matrix factorization (Zheng, Noroozi, and Yu 2017)—which, reportedly, is the state of the art for predicting customer ratings when textual reviews are available. To make results comparable, the Feat2Vec experiments use the same feature extraction function used by DeepCoNN. We evaluate on the Yelp dataset<sup>2</sup>, which consists of 4.7 million reviews of restaurants. For each user-item pair, DeepCoNN concatenates the text from all reviews for that item and all reviews by that user. The concatenated text is fed into a feature extraction function followed by a factorization machine. For Feat2Vec, we build 3 feature groups: item (restaurant) identifiers, user identifiers, and review text.

Table 1 compares our methods to DeepCoNN’s published results because a public implementation is not available. Feat2Vec provides a large performance increase when comparing the reported improvement in the mean squared error (MSE) over Matrix Factorization. Our approach is more general, and we claim that it is also more efficient. Since DeepCoNN concatenates text, when the average reviews per user is  $\bar{n}_u$  and reviews per item is  $\bar{n}_i$ , each text is duplicated on average  $\bar{n}_i \times \bar{n}_u$  times per training epoch. In contrast, for Feat2Vec each review is seen only once per epoch. Thus it can be 1-2 orders of magnitude more efficient for datasets where  $\bar{n}_i \times \bar{n}_u$  is large.

### 3.2 Self-Supervised Embeddings

The prediction of similar or related words is a commonly used method for evaluating self-supervised word embeddings (Bojanowski et al. 2017; Mikolov et al. 2013a). We generalize this task for feature types beyond text as follows.

<sup>2</sup><https://www.yelp.com/dataset/challenge>

Given a dataset with  $n$  feature groups  $g_1, g_2, \dots, g_n$  we have trained a self-supervised Feat2Vec model on, we choose a single group  $g_i$  as the “target” and a group  $g_j$  as the predictor. Given a held-out instance, we rank all possible values of the target group by cosine similarity to the source group’s embedding  $\phi_j(\vec{x}_{\vec{\kappa}_j})$  for that instance. We then retrieve the ranking of the actual target entity in the test dataset relative to cosine similarity scores for all other possible entities in  $\vec{\kappa}_i$ . Rankings are evaluated according to their mean percentile rank (MPR):  $MPR = \frac{1}{N} \sum_{i=1}^N R_i / (\max_i R_i)$ , where  $R_i$  is the rank of the entity for observation  $i$  under our evaluation procedure, and  $\max_i R_i$  is the worst ranking that can be received. A score of 0 would indicate perfect performance (i.e. top rank every test sample given), so lower is better under this metric.

Under this evaluation, we will compare the performance of Feat2Vec algorithm to Word2Vec’s CBOW algorithm for learning embeddings. To create a corpus for training Word2Vec, each record in the data is treated as a separate document. Specifically, we tokenize all feature values in a record and treat them as words (e.g. a movie record flagged as belonging to the horror genre has the token `genre_horror` in its document). We set the Word2Vec context window wide enough so that training is invariant to the token order. We use hyperparameters  $\alpha_2 = .75$  and  $k = 5$  from (Mikolov et al. 2013b) and set the embeddings dimension to  $r = 50$ . These hyperparameters are used in each of our embedding models during evaluation. The novel feature group weighting parameter in Feat2Vec,  $\alpha_1$ , was set ex-ante to a value  $\alpha_1 = 0.25$  that would showcase Feat2Vec’s ability to oversample low-complexity feature groups. It is not tuned to the data. We evaluate self-supervised Feat2Vec on 3 datasets:

**Movies** The Internet Movie Database (IMDB) is a publicly available dataset<sup>3</sup> of information related to films, television programs and video games. We limit our experiments to data on its 465,136 movies. Table 2 summarizes our feature groups. For groups with multiple categories, we use a “Bag of categories” approach. For example, we sum embeddings of individual actors such as Tom Cruise to produce a single “actor” feature group embedding associated with a title.

**Education** We use a dataset from a leading technology company that provides educational services. In this proprietary dataset, we have 57 million observations and 9 categorical feature types which include textbook identifier, user identifier, school identifier, along with other proprietary features. Each observation is an interaction a user had with a textbook.

**Yelp** We use the Yelp dataset from our supervised experiments to evaluate the efficacy of self-supervised embeddings in ratings prediction. We train embeddings for the following feature groups: user ID, restaurant ID, review text, number

Table 2: IMDB Feature Groups

Feature Type Name	Type	# of features
Runtime (minutes)	Real-valued	1
IMDB rating (0-10)	Real-valued	1
# of rating votes	Real-valued	1
Is adult film?	Boolean	2
Movie release year	Categorical	271
Movie title	Text	165,471
Directors	Bag of categories	174,382
Genres	Bag of categories	28
Writers	Bag of categories	244,241
Principal actors	Bag of categories	1,104,280

of times a review is flagged as funny, and 0-5 star rating of the review.

**Results** After training IMDB embeddings, we use cast members in a held out set of movies to predict the actual director of the film. We rank directors by cosine similarity of their embedding to the title’s cast member feature group embedding. For the educational dataset, we directly retrieve the most similar textbooks to the user embedding. Table 3 presents our evaluation results. Feat2Vec outperforms CBOW in the MPR metric. Additionally, while CBOW predicts the actual director 1.26% of the times, Feat2Vec does so 2.43% of the time, a 92% improvement in Top-1 Precision over CBOW.

Table 3: Mean percentile rank

Dataset	Feat2Vec	CBOW
IMDB	19.36%	24.15%
Educational	25.2%	29.2%

**Self-Supervised Feat2Vec Performance with Continuous Inputs** We now focus on how well our estimated self-supervised Feat2Vec model performs when predicting a real-valued feature. We expect this task to highlight Feat2Vec’s advantage over token-based embedding learning algorithms, such as Word2Vec. For IMDB ratings, we use a 3-layer DNN extraction function that will require embeddings of numerically similar ratings to be close, while Word2Vec will treat two numerically different ratings as completely different entities. We evaluate the prediction of IMDB ratings by choosing the rating embedding most similar<sup>4</sup> to the embedding of the movie’s director, and compute the MSE of the predicted rating in the test dataset. Feat2Vec scores an MSE of 6.6, while Word2Vec (CBOW) scores 9.3.

We also use the Yelp dataset, predicting the most similar rating embedding to the review text embeddings produced by Feat2Vec, Word2Vec (CBOW), and Doc2Vec (DM). Doc2Vec is only used for the Yelp dataset because there is no obvious “base document” in the IMDB dataset for Doc2Vec to learn on, while in the Yelp dataset the review text is a natural candidate. For Word2Vec and Doc2Vec, the review text embedding is the average of word embeddings, analogous to

<sup>3</sup><http://www.imdb.com/interfaces/>

<sup>4</sup>As before, the metric is cosine similarity.

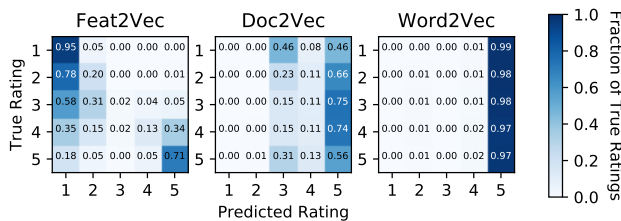


Figure 2: Confusion Matrices of Self-Supervised Yelp Ratings Predictions

the context vector used for learning in these algorithms. Figure 2 reports the confusion matrices for each model from this experiment. Word2Vec is poor in its predictive power, predicting 5 stars for 97% of the test sample. Though Feat2Vec and Doc2Vec yield comparable MSE (2.94 vs. 2.92, respectively), Feat2Vec outperforms Doc2Vec by a substantial margin in classification error rate (55% vs. 73%) and mean absolute error (1.13 vs. 1.31). In general, Doc2Vec is unable to identify low rating reviews: only 0.4% of Doc2Vec predictions are  $\leq 2$  stars, despite this comprising 20% of the data. In contrast, Feat2Vec is more diverse in its predictions, and better able to identify extreme reviews.

## 4 Conclusion

Embeddings have proven useful in a wide variety of contexts, but they are typically built from datasets with a single feature type as in the case of Word2Vec, or tuned for a single prediction task as in the case of Factorization Machine. We believe Feat2Vec is an important step towards general-purpose embedding methods. It decouples feature extraction from prediction for datasets with multiple feature types, it can be self-supervised, and its embeddings are easily interpretable.

In the supervised setting, Feat2Vec outperforms an algorithm specifically designed for text—even when using the same feature extraction CNN. In the self-supervised setting, Feat2Vec exploits the structure of a dataset to learn embeddings in a more sensible way than existing methods. This yields performance improvements in our ranking and prediction tasks. To the extent of our knowledge, self-supervised Feat2Vec is the first method to calculate continuous representations of data with multi-modal feature types and no explicit labels.

Future work could study how to reduce the amount of human knowledge our approach requires; for example by automatically grouping features into entities, or by automatically choosing a feature extraction function. These ideas can extend to our codebase that we make available<sup>5</sup>. Though further experimentation is necessary, we believe that our results

<sup>5</sup>The Feat2Vec model code is available here: <https://www.dropbox.com/sh/wdr2sgt0z9gj6kb/AABlzw7QhteTYViSoMk3CDZpa?dl=0>. The self-supervised experiments using non-proprietary data are available here: [https://www.dropbox.com/sh/uc07ng403i4ss9a/AAAPshFzdOug\\_ooeLn4SGR3Ua?dl=0](https://www.dropbox.com/sh/uc07ng403i4ss9a/AAAPshFzdOug_ooeLn4SGR3Ua?dl=0) here. The supervised experiments are available here: <https://goo.gl/zEQBiA>.

are an encouraging step forward towards general-purpose embedding models.

## References

- Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5:135–146.
- Dyer, C. 2014. Notes on noise contrastive estimation and negative sampling. *arXiv preprint arXiv:1410.8251*.
- Dziugaite, G. K., and Roy, D. M. 2015. Neural network matrix factorization. *CoRR* abs/1511.06443.
- Guo, H.; TANG, R.; Ye, Y.; Li, Z.; and He, X. 2017. Deepfm: A factorization-machine based neural network for ctr prediction. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 1725–1731.
- Gutmann, M., and Hyvärinen, A. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 297–304.
- Jenatton, R.; Obozinski, G.; and Bach, F. 2010. Structured sparse principal component analysis. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 366–373.
- Kalchbrenner, N.; Grefenstette, E.; and Blunsom, P. 2014. A convolutional neural network for modelling sentences. *CoRR* abs/1404.2188.
- Le, Q., and Mikolov, T. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 1188–1196.
- Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.
- Mnih, A., and Teh, Y. W. 2012. A fast and simple algorithm for training neural probabilistic language models. In *Proceedings of the International Conference on Machine Learning*.
- Rendle, S. 2010. Factorization machines. In *Proceedings of the 2010 IEEE International Conference on Data Mining, ICDM '10*, 995–1000. Washington, DC, USA: IEEE Computer Society.
- Wu, L. Y.; Fisch, A.; Chopra, S.; Adams, K.; Bordes, A.; and Weston, J. 2018. Starspace: Embed all the things! In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Zhang, Y., and Wallace, B. 2017. A sensitivity analysis of (and practitioners’ guide to) convolutional neural networks for sentence classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 253–263. Asian Federation of Natural Language Processing.
- Zheng, L.; Noroozi, V.; and Yu, P. S. 2017. Joint deep modeling of users and items using reviews for recommendation. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, WSDM '17*, 425–434. New York, NY, USA: ACM.