Deep Dirichlet Multinomial Regression

Adrian Benton  Mark Dredze
Center for Language and Speech Processing
Johns Hopkins University
Baltimore, MD 21218, USA
{adrian,mdredze}@cs.jhu.edu

Abstract

Dirichlet Multinomial Regression (DMR) and other supervised topic models can incorporate arbitrary document-level features to inform topic priors. However, their ability to model corpora are limited by the representation and selection of these features—a choice the topic modeler must make. Instead, we seek models that can learn the feature representations upon which to condition topic selection. We present deep Dirichlet Multinomial Regression (dDMR), a generative topic model that simultaneously learns document feature representations and topics. We evaluate dDMR on three datasets: New York Times articles with fine-grained tags, Amazon product reviews with product images, and Reddit posts with subreddit identity. dDMR learns representations that outperform DMR and LDA according to heldout perplexity and are more effective at downstream predictive tasks as the number of topics grows. Additionally, human subjects judge dDMR topics as being more representative of associated document features. Finally, we find that supervision leads to faster convergence as compared to an LDA baseline and that dDMR’s model fit is less sensitive to training parameters than DMR.

1 Introduction

Fifteen years of research on topic models, starting from Latent Dirichlet Allocation (LDA) (Blei et al., 2003), have led to a variety of models for numerous data settings. These models identify sets (distributions) of related words that reflect semantic topics in a large corpus of text data. Topic models are now routinely used in the social sciences and humanities to analyze text collections (Schmidt, 2012).

Document collections are often accompanied by metadata and annotations, such as a book’s author, an article’s topic descriptor tags, images associated with a product review, or structured patient information associated with clinical records. These document-level annotations can provide additional supervision for guiding topic model learning. Additional information can be integrated into topic models using either downstream or upstream models. Downstream models, such as supervised LDA (Mcauliffe and Blei, 2008), assume that these additional document features are generated from each document’s topic distribution. These models are most helpful when you desire topics that are predictive of the output, such as models for predicting the sentiment of product reviews. Upstream models, such as Dirichlet Multinomial Regression (DMR), condition each document’s topic distribution on document features, such as author (Rosen-Zvi et al., 2004), social network (McCallum et al., 2007), or document labels (Ramage et al., 2009). Previous work has demonstrated that upstream models tend to outperform downstream models in terms of model fit, as well as extracting topics that are useful in prediction of related tasks (Benton et al., 2016).

DMR is an upstream topic model with a particularly attractive method for incorporating arbitrary document features. Rather than defining specific random variables in the graphical model for each new document feature, DMR treats the document annotations as features in a log-linear model. The log-linear model parameterizes the Dirichlet prior for the document’s topic distribution, making the Dirichlet’s hyperparameter (typically \( \alpha \)) document-specific. By making no assumptions on model structure of new random variables, DMR is flexible to incorporating different types of features.

Despite this flexibility, DMR models are typically restricted to a small number of document features. Several reasons account for this restriction: 1) Many text corpora only have a small number of document-level features; 2) Model hyperparameters become less interpretable as the dimensionality grows; and 3) DMR is liable to overfit the hyperparameters when the dimensionality of document features is high. In practice, applications of DMR are limited to settings with a small number of features, or where the analyst selects a few meaningful features
A solution to this restriction is to learn low-dimensional representations of document features. Neural networks have shown wide-spread success at learning generalizable representations, often obviating the need for hand designed features (Collobert and Weston, 2008). A prime example is word embedding features in natural language processing, which supplant traditional lexical features (Brown et al., 1992; Mikolov et al., 2013; Pennington et al., 2014). Jointly learning networks that construct feature representations along with the parameters of a standard NLP model has become a common approach. For example, (Yu et al., 2015) used a tensor decomposition to jointly learn features from both word embeddings and traditional NLP features, along with the parameters of a relation extraction model. Additionally, neural networks can handle a variety of data types, including text, images and general metadata features. This makes them appropriate for addressing dimensionality reduction in DMR.

We propose deep Dirichlet Multinomial Regression (dDMR), a model that extends DMR by introducing a deep neural network that learns a transformation of the input metadata into features used to form the Dirichlet hyperparameter. Whereas DMR parameterizes the document-topic priors as a log-linear function of document features, dDMR jointly learns a feature representation for each document along with a log-linear function that best captures the distribution over topics. Since the function mapping document features to topic prior is a neural network, we can jointly optimize the topic model and the neural network parameters by gradient ascent with back-propagation. We show that dDMR can use network architectures to better fit text corpora with high-dimensional document features as compared to other supervised topic models. The topics learned by dDMR are judged as being more representative of document features by human subjects. We also find that dDMR tends to converge in many fewer iterations than LDA, and also does not suffer from tuning difficulties that DMR encounters when applied to high-dimensional document features.

2 Model

Our model builds on the generative model of DMR: an LDA-style topic model that replaces the hyperparameter (vector) of the topic distribution Dirichlet prior with a hyperparameter that is output from a log-linear model given the document features. Our model deep DMR (dDMR) replaces this log-linear model with an arbitrary function $f$ that maps a real-valued vector of dimension $F$ to a representation of dimension $K$. For simplicity we make no assumptions on the choice of this function, only that it can be optimized to minimize a cost on its output by gradient ascent. In practice, we define this function as a neural network, where the architecture of this network is informed by the type of document features, e.g. a convolutional neural network for images. We use neural networks since they are expressive, generalize well to unseen data, and can be jointly trained using straightforward gradient ascent with back-propagation.

The generative story for dDMR is as follows:

1. Representation function $f \in \mathbb{R}^F \to \mathbb{R}^K$
2. Topic-word prior parameters: $\omega^{bias} \in \mathbb{R}^V$
3. For each document $m$ with features $\alpha_m \in \mathbb{R}^F$, generate document prior:
   (a) $\tilde{\theta}_m = exp(f(\alpha_m))$
   (b) $\theta_m \sim Dirichlet(\tilde{\theta}_m)$
4. For each topic $k$, generate word distribution:
   (a) $\tilde{\phi}_k = exp(\omega^{bias})$
   (b) $\phi_k \sim Dirichlet(\tilde{\phi}_k)$
5. For each token $(m,n)$, generate data:
   (a) Topic (unobserved): $z_{m,n} \sim \theta_m$
   (b) Word (observed): $w_{m,n} \sim \phi_{z_{m,n}}$

where $V$ is the vocabulary size and $K$ are the number of topics. In practice, the document features need not be restricted to fixed-length feature vectors, e.g. $f$ may be an RNN that maps from a sequence of characters to a fixed-length vector in $\mathbb{R}^k$. DMR is a special case of dDMR with the choice of a linear function for $f$. Figure 1 displays the graphical model diagram for dDMR.

2.1 Inference and Parameter Estimation

We infer the random variables of the topic model using collapsed Gibbs sampling, and estimate the model parameters using gradient ascent with back-propagation. We use alternating optimization: one
iteration of collapsed Gibbs sampling (sample topics for each word) and then an update of the parameters of \( f \) by gradient ascent to maximize the log-likelihood of the tokens and topic assignments. Given the parameters, the sampling step remains unchanged from LDA (Griffiths and Steyvers, 2004). The network parameters are estimated via backpropagation through the network for a fixed sample. Eq. 1 shows the gradient of the data log-likelihood, \( \mathcal{L} \), with respect to \( \theta_{m,k} = \exp(f(\alpha_{m,k})) \), the prior weight of topic \( k \) for document \( m \). \( \psi \) is the digamma function (derivative of the log-gamma function), \( n_{m} \) is the number of tokens in document \( m \), and \( n_{m,k} \) is the count of how many tokens topic \( k \) was assigned to in document \( m \).

\[
\frac{\delta \mathcal{L}}{\delta \theta_{m,k}} = \psi(\sum_{k=1}^{K} \theta_{m,k}) - \psi(\sum_{k=1}^{K} \theta_{m,k} + n_{m}) + \psi(\theta_{m,k} + n_{m,k}) - \psi(\theta_{m,k}) \tag{1}
\]

3 Data

We explore the flexibility of our model by considering three different datasets that include different types of metadata associated with each document. For each dataset, we describe the documents and metadata.

**New York Times** The New York Times Annotated Corpus (Sandhaus, 2008) contains articles with extensive metadata used for indexing by the newspaper. For supervision, we used the “descriptor” tags associated with each article assigned by archivists. These tags reflect the topic of an article, as well as organizations or people mentioned in the article. We selected all articles published in 1998, and kept those tags that were associated with at least 3 articles in that year – 2424 unique tags. 20 of the 200 most frequent tags were held out from training for validation purposes: \{“education and schools”, “law and legislation”, “advertising”, “budgets and budgeting”, “freedom and human rights”, “telephones and telecommunications”, “bombs and explosives”, “sexual harassment”, “reform and reorganization”, “teachers and school employees”, “tests and testing”, “futures and options trading”, “boxing”, “firearms”, “company reports”, “embargoes and economic sanctions”, “hospitals”, “states (us)”, “bridge (card game)”, and “auctions”\}. Articles contained a mean of 2.1 tags, with 738 articles not containing any of these tags. Tags were represented using a one-hot encoding.

Articles were tokenized by non-alphanumeric characters and numerals were replaced by a special token. Words occurring in more than 40% of documents were removed, and only the 15,000 most frequent types were retained. There were a total of 89,397 articles with an average length of 158 tokens per article.

**Amazon Reviews** The Amazon product reviews corpus (McAuley and Yang, 2016) contains reviews of products as well as images of the product. We sampled 100,000 Amazon product reviews: 20,000 reviews sampled uniformly from the Musical Instruments, Patio, Lawn, & Garden, Grocery & Gourmet Food, Automotive, and Pet Supplies product categories. We hypothesize that knowing information about the product’s appearance will indicate which words appear in the review, especially for product images occurring in these categories. 66 of the reviews we sampled contained only highly infrequent tokens, and were therefore removed from our data, leaving 99,934 product reviews. Articles were preprocessed identically to the New York Times data.

We include images as supervision by using the 4096-dimensional second fully-connected layer of the Caffe convolutional neural network reference model, trained to predict ImageNet object categories\(^1\). Using these features as supervision to dDMR is similar to fine-tuning a pre-trained CNN to predict a new set of labels. Since the Caffe reference model is already trained on a large corpus of images, we chose to fine-tune only the final layers so as to learn a transformation of the already learned representation.

**Reddit** We selected a sample of Reddit posts made in January 2016. A standard stop list was used to remove frequent function words and we restricted the vocabulary to the 30,000 most frequent types. We restricted posts made to subreddits, collections of topically-related threads, with at least ten comments in this month (26,830 subreddits), and made by users with at least five comments across these subreddits (total of 1,351,283 million users). We then sampled 10,000 users uniformly at random and used all their comments as a corpus, for a total of 389,234 comments over 7,866 subreddits (token length mean: 16.3, median: 9)\(^2\).

This corpus differs from the others in two ways. First, Reddit documents are very short, which is problematic for topic models that rely on detecting correlations in token use. Second, the Reddit metadata that may be useful for topic modeling is necessarily high-dimensional (e.g. subreddit identity, a proxy for topical content). DMR may have trouble exploiting high-dimensional supervision.

4 Experiments

**Model Estimation** We used the same procedure for training topic models on each dataset. Hyperparameter gradient updates were performed after

\(^1\)Features used directly from [http://jmcauley.ucsd.edu/data/amazon/](http://jmcauley.ucsd.edu/data/amazon/)

\(^2\)The sampled comment IDs can be found here: [https://github.com/abenton/deep-dmr/blob/master/resources/reddit_comment_ids.txt](https://github.com/abenton/deep-dmr/blob/master/resources/reddit_comment_ids.txt)
a burnin period of 100 Gibbs sampling iterations. Hyperparameters were updated with the adaptive learning rate algorithm Adadelta (Zeiler, 2012), with a tuned base learning rate and fixed $\rho = 0.95$\textsuperscript{3}. All models were trained for a maximum of 15,000 epochs, with early stopping if heldout perplexity showed no improvements after 200 epochs (evaluated once every 20 epochs). Hyperparameters were fit on every other token in the corpus, and (held-out) log-likelihood/perplexity was calculated on the remaining tokens.

For the architecture of the $dDMR$ model we used single-hidden-layer multi-layer perceptrons (MLPs), with rectified linear unit (ReLU) activations on the hidden layer, and linear activation on the output layer. We sampled three architectures for each dataset, by drawing layer widths independently at random from $\{10, 50\}$, and also included two architectures with $\{50, 10\}$ and $\{100, 50\}$, (hidden, output) layers \textsuperscript{4}. We compare the performance of $dDMR$ to $D MR$ trained on the same feature set as well as $L DA$.

For the New York Times dataset, we also compare $dDMR$ to $D MR$ trained on features after applying principal components analysis (PCA) to reduce the dimensionality of descriptor feature supervision, sweeping over PCA projection width in $\{10, 50, 100, 250, 500, 1000\}$. Comparing performance of $dDMR$ to PCA-reduced $D MR$ tests two modeling choices. First, it tests the hypothesis that explicitly learning a representation for document annotations to maximize data likelihood produces a “better-fit” topic model than learning this annotation representation in unsupervised fashion – a two-step process. It also lets us determine if a linear dimensionality reduction technique is sufficient to learning a good feature representation for topic modeling, as opposed to learning a non-linear transformation of the document supervision. Note that we cannot apply PCA to reduce the dimensionality for subreddit id in Reddit since it is a one-hot feature.

Documents in each dataset were partitioned into ten equally-sized folds. Model training parameters of $L 1$ and $L 2$ regularization penalties on feature weights for $D MR$ and $dDMR$ and the base learning rate for each model class were tuned to minimize heldout perplexity on the first fold. These were tuned independently for each model, with number of topics fixed to 10, and $dDMR$ architecture fixed to narrow layer widths ($50, 10$). Model selection was based on the macro-averaged performance on the next eight folds, and we report performance on the remaining fold. We selected models separately for each evaluation metric. For $dDMR$, model selection amounts to selecting the document prior architecture, and for $D MR$ with PCA-reduced feature supervision, model selection involved selecting the PCA projection width.

**Evaluation** Each model was evaluated according to heldout perplexity, topic coherence by normalized pointwise mutual information (NPMI) (Lau et al., 2014), and a dataset-specific predictive task.

Heldout perplexity was computed by only aggregating document-topic and topic-word counts from every other token in the corpus, and evaluating perplexity on the remaining heldout tokens. This corresponds to the “document completion” evaluation method as described in (Wallach et al., 2009), where instead of holding out the words in the second half of a document, every other word is held out.

NPMI (Lau et al., 2014) computes a an automatic measure of topic quality, the sum of pointwise mutual information between pairs of $m$ most likely words normalized by the negative log of each pair jointly occurring within a document (Eq. 2). We calculated this topic quality metric on the top 20 most probable words in each topic, and averaged over the most coherent 1, 5, 10, and over all learned topics. However, models were selected to only maximize average NPMI over all topics.

$$\text{NPMI} = \sum_{i=1}^{m} \sum_{j=m+1}^{m} \frac{\log P(w_i, w_j)}{-\log P(w_i)P(w_j)} \quad (2)$$

For prediction tasks, we used the sampled topic distribution associated with a document, averaged over the last 100 iterations, as features to predict a document-level label. For New York Times articles we predicted 10 of the 200 most frequent descriptor tags restricting to articles with exactly one of these descriptors. For Amazon, we predicted the product category a document belonged to (one of five), and for Reddit we predicted a heldout set of document subreddit IDs. In the case of Reddit, these heldout subreddits were 10 out of the 100 most prevalent in our data, and were held out similar to the New York Times evaluation. SVM models were fit on inferred topic distribution features and were then evaluated according to accuracy, F1-score, and area under the ROC curve. The SVM slack parameter was tuned by 4-fold cross-validation on 60% of the documents, and evaluated on the remaining 40%.

We also collected human topic judgments using Amazon Mechanical Turk (Callison-Burch and
Table 1: Test fold heldout perplexity for each dataset and model for number of topics \( Z \). Standard error of mean heldout perplexity over all cross-validation folds in parentheses.

| \( Z \) | Model | NYT | Amazon | Reddit |
|-------|-------|-----|--------|--------|
| 10    | LDA   | 3422 (5) | 2500 (7) | 3811 (15) |
|       | DMR   | 3805 (6) | 2475 (9) | 3753 (10) |
|       | DMR-PCA | 3417 (8) |        |        |
|       | dDMR  | 3395 (7) | 2272 (68) | 3024 (13) |
| 20    | LDA   | 3031 (6) | 2255 (17) | 3055 (19) |
|       | DMR   | 3018 (4) | 2556 (48) | 3650 (8) |
|       | DMR-PCA | 3082 (8) |        |        |
|       | dDMR  | 3023 (7) | 2222 (7) | 3581 (16) |
| 50    | LDA   | 2766 (8) | 2269 (9) | 3095 (17) |
|       | DMR   | 2797 (34) | 2407 (20) | 3640 (40) |
|       | DMR-PCA | 2773 (9) |        |        |
|       | dDMR  | 2657 (8) | 2197 (13) | 3597 (17) |
| 100   | LDA   | 2491 (8) | 2206 (10) | 3676 (19) |
|       | DMR   | 2491 (27) | 2410 (75) | 3832 (30) |
|       | DMR-PCA | 2644 (52) |        |        |
|       | dDMR  | 2483 (10) | 2215 (6) | 3642 (18) |
| 200   | LDA   | 2513 (8) | 2217 (7) | 3653 (19) |
|       | DMR   | 2630 (13) | 2480 (65) | 3905 (15) |
|       | DMR-PCA | 2525 (14) |        |        |
|       | dDMR  | 2394 (9) | 2214 (12) | 3587 (11) |

It is also worth noting that finding a low-dimensional linear projection of the supervision features with PCA does not improve model fit as well as dDMR. dDMR benefits both from joint learning to maximize corpus log-likelihood and possibly by the flexibility of learning non-linear projection (through the hidden layer ReLU activations).

Another striking result is the difference in speed of convergence between the supervised models and LDA (Figure 2). Even supervision that provides a weak signal for topic modeling, such as Amazon product image features, can speed convergence over LDA. In certain cases (Figure 2 left), training dDMR for 1,000 iterations results in a lower perplexity model than LDA trained for over 10,000 iterations.

In terms of actual run time, parallelization of model training differs between the supervised model and LDA. Gradient updates necessary for learning the representation can be trivially distributed across multiple cores using optimized linear algebra libraries (e.g., BLAS), mitigating the additional cost incurred by hyperparameter updates in supervised models. In contrast, the Gibbs sampling iterations can also be parallelized, but not as easily, ultimately making resampling topics the most expensive step in model training. Because of this, the potential difference in runtime for a single iteration between dDMR and LDA is small, with the former converging in far fewer iterations. In our experiments, per iteration time taken by DMR or dDMR was at most twice as long as LDA across all experiments.

dDMR performance is also insensitive to training parameters relative to DMR. While DMR requires heavy L1 and L2 regularization and a very small step size to achieve low heldout perplexity, dDMR is relatively insensitive to the penalty on regularization and benefits from a higher base learning rate (Figure 3). We found that dDMR is easier to tune than DMR, requiring less exploration of the training parameters. This is also corroborated by higher variance in perplexity achieved by DMR across different cross-validation folds (Table 1).

**Topic Quality** Results for the automatic topic quality evaluation, NPMI, are mixed across datasets. In many cases, LDA and DMR score highly according to NPMI, despite achieving higher heldout perplexity than dDMR (Table 2). This may not be surprising as previous work has found that perplexity does not correlate well with human judgments of topic coherence (Lau et al., 2014).

However, in the human evaluation, subjects find that dDMR-learned topics are more representative of document annotations than DMR (Table 3). While subjects only statistically significantly favored dDMR models over LDA on the Reddit data, they favored dDMR topics over LDA across all datasets, and significantly preferred dDMR top-
Figure 2: Heldout perplexity as a function of iteration for lowest-perplexity models with $Z = 100$. The vertical dashed line indicates when models are burned in and hyperparameter optimization begins.

Figure 3: Heldout perplexity on the Amazon data tuning fold for $DMR$ (orange) and $dDMR$ (purple) with a (50, 10) layer architecture as a function of training parameters: L1, L2 feature weight regularization, and base learning rate. All models were trained for a fixed 5,000 iterations, with horizontal jitter added to each point.

Table 2: Top-1, 5, 10, and overall topic NPMI across all datasets. Models that maximized overall NPMI across dev folds were chosen and the best-performing model is in bold.
| New York Times | LDA | DMR |
|---------------|-----|-----|
|                | 51.1% | 61.9% |
| Amazon        | 51.9% | 61.4%* |
| Reddit        | 55.5%* | 57.6%* |

Table 3: % HITs where humans preferred dDMR topics as more representative of document supervision than the competing model. * denotes statistical significance according to a one-tailed binomial test at the $p = 0.05$ level.

ics over DMR on two of the three datasets. This is contrary to themodel rankings according to NPMI, which suggest that DMR topics are often higher quality when it comes to human interpretability.

We also qualitatively explored the product image representations DMR and dDMR learned on the Amazon data. To do so, we computed and normalized the prior document distribution for a sample of documents for lowest perplexity DMR and dDMR $Z = 200$ topic models: $p(k|m) = \frac{\theta_{m,k}}{\sum_{k=1}^{Z} \theta_{m,k}}$, the prior probability of sampling topic $k$, conditioned on the features for document $m$. We then marginalize over topics to yield the conditional probability of a word $w$ given document $m$: $p(w|m) = \sum_{k=1}^{Z} p(w|k)p(k|m)$. Table 4 contains a sample of these probable words given document supervision. We find that dDMR identifies words likely to appear in a review of the product pictured. However, some images lead dDMR down a garden path. For example, a bottle of “Turtle Food” should not be associated with words for human consumables like “coffee” and “chocolate”, despite the container resembling some of these products. However, the image-specific document priors DMR learned are not as sensitive to the actual product image as those learned by dDMR. The prior conditional probabilities $p(w|m)$ for “Turtle Food”, “Slushy Magic Cup”, and “Rawhide Dog Bones” product images are all ranked identically by DMR.

**Predictive Performance** Finally, we consider the utility of the learned topic distributions for downstream prediction tasks, a common use of topic models. Although token perplexity is a standard measure of topic model fit, it has no direct relationship with how topic models are typically used: to identify consistent themes or reduce the dimensionality of a document corpus. We found that features based on topic distributions from dDMR outperform LDA and DMR on the Amazon and Reddit data when the number of topics fit is large, although they fail to outperform DMR on New York Times (Table 5). Heldout perplexity is strongly correlated with predictive performance, with a Pearson correlation coefficient, $\rho = 0.898$ between F1-score and heldout perplexity on the Amazon data. This strong correlation is likely due to the tight relationship between words used in product reviews and product category: a model that assigns high likelihood to a words in a product review corpus should also be informative of the product categories. Prior work showed that upstream supervised topic models, such as DMR, learn topic distributions that are effective at downstream prediction tasks (Benton et al., 2016). We find that topic distributions learned by dDMR improve over DMR in certain cases, particularly as the number of topics increases.

### 6 Related Work

With the widespread adoption of neural networks, others have sought to combine topic and neural models. One line of work replaces generative, LDA-based, topic models with discriminatively-trained models based on neural networks. (Cao et al., 2015) model $\theta$ and $\phi$ using neural networks with softmax output layers and learn network parameters that maximize data likelihood. They also learn n-gram embeddings to identify topics whose elements are not restricted to unigrams. (Chen et al., 2015) similarly expresses the (smoothed) supervised LDA (Mcauliffe and Blei, 2008) generative model as a neural network, and give an algorithm to discriminatively train it. (Wan et al., 2012) take a similar approach to dDMR where they use a neural network to extract image representations that maximize the probability of SIFT descriptors extracted from the image. However, this model is used for image classification, not for exploring a corpus of documents as is typical of topic models. These models are computationally attractive in that they avoid approximating the posterior distribution of topic assignments given tokens by dropping the assumption that $\theta$ and $\phi$ are drawn from Dirichlet priors. Model fitting is performed by back-propagation of a max-margin cost. In contrast, we use neural networks to learn feature representations for documents, not as a replacement for the LDA generative story. This is similar to variants of SPRITE (Paul and Dredze, 2015), where many document-level factors are combined to generate a document-topic prior. In contrast to several of these models, the core of our topic model remains unchanged, meaning that dDMR is agnostic to many other extensions of LDA.

There has been extensive work in modeling both textual and visual topics. Models such as Corr-LDA (Blei and Jordan, 2003) suppose that a text document and associated image features are generated by a shared latent topic. This property is shared by other topic models over images, such as STM-TwitterLDA (Cai et al., 2015) and (Zhang et al., 2015). While these models try to model images, we instead use images in the Amazon data to better estimate topic distributions.

Our experiment on using images to model Ama-
Table 4: Top twenty words associated with each of the product images – learned by dDMR vs. DMR ($Z = 200$). These images were drawn at random from the Amazon corpus (no cherry-picking involved). Word lists were generated by marginalizing over the prior topic distribution associated with that image, and then normalizing each word’s probability by subtracting off its mean marginal probability across all images in the corpus. This is done to avoid displaying highly frequent words. Words that differ between each model’s ranked list are in bold.

---

**7 Conclusion**

We present deep Dirichlet Multinomial Regression, a supervised topic model which both learns a representation of document-level features and how to use that representation for informing a topic distribution. We demonstrate the flexibility of our model on three corpora with different types of metadata: topic descriptor tags, images, and subreddit IDs. dDMR is better able to fit text corpora with high-dimensional supervision compared to LDA or DMR. Furthermore, we find that document supervision greatly reduces the number of Gibbs sampling iterations for a topic model to converge, and that the dDMR prior architecture makes it more robust to training parameters than DMR. We also find that the topic distributions learned by dDMR are more predictive of external Amazon product reviews resembles work on image caption generation, yet the similarity is superficial. The relationship between an image and its caption is relatively tight (Fang et al., 2015) – objects in the image will likely be referenced in the caption. For Amazon product reviews, visual features of the product, like color, may be explicitly mentioned in the review, but then again, they may not. Also, the aim of topic models is to extract common themes of co-occurring words, and how those themes are distributed across each document. The similarity between our work and captioning lies only in the fact that we extract image features from a CNN trained as an object recognizer to inform document-topic distributions.
| Z | Model | New York Times | Amazon | Reddit |
|---|-------|----------------|--------|--------|
| 10 | LDA | 0.208 | 0.380 | 0.767 | 0.662 | 0.667 | 0.891 | 0.130 | 0.276 | 0.565 |
|   | DMR | 0.236 | 0.367 | 0.781 | 0.311 | 0.407 | 0.619 | 0.092 | 0.229 | 0.597 |
|   | DMR-PCA | 0.280 | 0.347 | 0.758 | 0.154 | 0.347 | 0.790 | 0.130 | 0.276 | 0.565 |
|   | dDMR | 0.208 | 0.380 | 0.767 | 0.608 | 0.656 | 0.864 | 0.170 | 0.300 | 0.596 |
| 20 | LDA | 0.315 | 0.463 | 0.784 | 0.657 | 0.659 | 0.887 | 0.121 | 0.258 | 0.579 |
|   | DMR | 0.319 | 0.477 | 0.805 | 0.294 | 0.405 | 0.647 | 0.057 | 0.245 | 0.520 |
|   | DMR-PCA | 0.343 | 0.540 | 0.831 | 0.706 | 0.711 | 0.911 | 0.071 | 0.274 | 0.566 |
|   | dDMR | 0.315 | 0.463 | 0.784 | 0.650 | 0.644 | 0.870 | 0.130 | 0.276 | 0.565 |
|   | LDA | 0.455 | 0.613 | 0.849 | 0.630 | 0.634 | 0.870 | 0.131 | 0.199 | 0.542 |
|   | DMR | 0.478 | 0.650 | 0.877 | 0.396 | 0.499 | 0.619 | 0.113 | 0.199 | 0.542 |
|   | DMR-PCA | 0.505 | 0.667 | 0.887 | 0.716 | 0.726 | 0.916 | 0.118 | 0.272 | 0.551 |
|   | dDMR | 0.507 | 0.657 | 0.856 | 0.716 | 0.726 | 0.916 | 0.118 | 0.272 | 0.551 |
| 50 | LDA | 0.531 | 0.657 | 0.874 | 0.646 | 0.649 | 0.874 | 0.148 | 0.201 | 0.538 |
|   | DMR | 0.552 | 0.683 | 0.898 | 0.392 | 0.463 | 0.688 | 0.107 | 0.233 | 0.512 |
|   | DMR-PCA | 0.602 | 0.687 | 0.917 | 0.650 | 0.660 | 0.893 | 0.172 | 0.316 | 0.614 |
|   | dDMR | 0.514 | 0.653 | 0.893 | 0.650 | 0.660 | 0.893 | 0.172 | 0.316 | 0.614 |
| 200 | LDA | 0.566 | 0.683 | 0.903 | 0.646 | 0.651 | 0.882 | 0.111 | 0.227 | 0.547 |
|   | DMR | 0.576 | 0.670 | 0.917 | 0.288 | 0.401 | 0.697 | 0.089 | 0.229 | 0.499 |
|   | DMR-PCA | 0.648 | 0.762 | 0.915 | 0.716 | 0.721 | 0.909 | 0.198 | 0.323 | 0.580 |
|   | dDMR | 0.605 | 0.730 | 0.903 | 0.716 | 0.721 | 0.909 | 0.198 | 0.323 | 0.580 |

Table 5: Top F-score, accuracy, and AUC on prediction tasks for all datasets.

References

Adrian Benton, Michael J Paul, Braden Hancock, and Mark Dredze. 2016. Collective supervision of topic models for predicting surveys with social media. In Proceedings of the AAAI Conference on Artificial Intelligence. pages 2892–2898.

David M Blei and Michael I Jordan. 2003. Modeling annotated data. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, pages 127–134.

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

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. Computational linguistics 18(4):467–479.

Hongyun Cai, Yang Yang, Xuefei Li, and Zi Huang. 2015. What are popular: exploring twitter features for event detection, tracking and visualization. In Proceedings of the 23rd ACM International Conference on Multimedia. ACM, pages 89–98.

Chris Callison-Burch and Mark Dredze. 2010. Creating speech and language data with amazon’s mechanical turk. In NAACL-HLT Workshop on Creating Speech and Language Data With Mechanical Turk. pages 1–12.

Ziqiang Cao, Sujian Li, Yang Liu, Wenjie Li, and Heng Ji. 2015. A novel neural topic model and its supervised extension. In Proceedings of the AAAI conference on Artificial Intelligence. pages 2210–2216.

Jianshu Chen, Ji He, Yelong Shen, Lin Xiao, Xiaodong He, Jianfeng Gao, Xinying Song, and Li Deng. 2015. End-to-end learning of LDA by mirror-descent back propagation over a deep architecture. In Advances in Neural Information Processing Systems. pages 1765–1773.

Roman Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th International Conference on Machine Learning. ACM, pages 160–167.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. From captions to visual concepts and back. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pages 1473–1482.

Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. Proceedings of the National Academy of Sciences 101(suppl 1):5228–5235.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. pages 530–539.
Julian McAuley and Alex Yang. 2016. Addressing complex and subjective product-related queries with customer reviews. In Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, pages 625–635.

Jon D McAuliffe and David M Blei. 2008. Supervised topic models. In Advances in Neural Information Processing Systems. pages 121–128.

Andrew McCallum, Xuernui Wang, and Andres Corrada-Emmanuel. 2007. Topic and role discovery in social networks with experiments on enron and academic email. Journal of Artificial Intelligence Research.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems. pages 3111–3119.

Michael J Paul and Mark Dredze. 2015. Sprite: Generalizing topic models with structured priors. Transactions of the Association for Computational Linguistics 3:43–57.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods on Natural Language Processing. volume 14, pages 1532–1543.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pages 248–256.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. AUAI Press, pages 487–494.

Evan Sandhaus. 2008. The new york times annotated corpus. Linguistic Data Consortium, Philadelphia 6(12):e26752.

Benjamin M Schmidt. 2012. Words alone: Dismantling topic models in the humanities. Journal of Digital Humanities 2(1):49–65.

Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In Proceedings of the 26th Annual International Conference on Machine Learning. ACM, pages 1105–1112.

Li Wan, Leo Zhu, and Rob Fergus. 2012. A hybrid neural network-latent topic model. In Proceedings of the 15th International Conference on Artificial Intelligence and Statistics. volume 12, pages 1287–1294.

Mo Yu, Matthew R Gormley, and Mark Dredze. 2015. Combining word embeddings and feature embeddings for fine-grained relation extraction. In Proceedings of the 14th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pages 1374–1379.

Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701.

Hao Zhang, Gunhee Kim, and Eric P Xing. 2015. Dynamic topic modeling for monitoring market competition from online text and image data. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pages 1425–1434.