Feature discovery and visualization of robot mission data using convolutional autoencoders and Bayesian nonparametric topic models

Feature discovery and visualization of robot mission data using convolutional autoencoders and Bayesian nonparametric topic models

Genevieve Flaspohler[1,2], Nicholas Roy[1], and Yogesh Girdhar[2]

***Abstract*— The gap between our ability to collect interesting data and our ability to analyze these data is growing at an unprecedented rate. Recent algorithmic attempts to fill this gap have employed unsupervised tools to discover structure in data. Some of the most successful approaches have used probabilistic models to uncover latent thematic structure in discrete data. Despite the success of these models on textual data, they have not generalized as well to image data, in part because of the spatial and temporal structure that may exist in an image stream.**

We introduce a novel unsupervised machine learning framework that incorporates the ability of convolutional autoencoders to discover features from images that directly encode spatial information, within a Bayesian nonparametric topic model that discovers meaningful latent patterns within discrete data. By using this hybrid framework, we overcome the fundamental dependency of traditional topic models on rigidly hand-coded data representations, while simultaneously encoding spatial dependency in our topics without adding model complexity. We apply this model to the motivating application of high-level scene understanding and mission summarization for exploratory marine robots. Our experiments on a seafloor dataset collected by a marine robot show that the proposed hybrid framework outperforms current state-of-the-art approaches on the task of unsupervised seafloor terrain characterization.

I. Introduction

The benthic deep sea, the largest two-dimensional habitat on earth, is difficult to study and vastly unexplored. Autonomous underwater vehicles (AUVs) are filling observational gaps by collecting large datasets consisting of multiple sensor modalities, including seafloor imagery. This paper presents a novel unsupervised machine learning technique to discover and visualize structure in image datasets, enabling concise mission summarization and equipping exploratory robots with the capacity to describe their environment semantically, a precursor to adaptive real-time exploration. Although the focus of this work is the underwater domain, the proposed approach is applicable to any domains where there exist large volumes of unstructured image sequence data that would typically require manual analysis, such as remote sensing and long term monitoring.

Some of the most successful models for discovering structure within discrete data without supervision are Bayesian topic models, such as the latent Dirichlet allocation (LDA) [1] and its non-parametric extension, the Hierarchical Dirichlet process (HDP) [2]. Initially applied to text corpora, the modeling assumptions made by LDA and HDP allow them to discover useful latent structure that often corresponds to cohesive, human-understandable topics [3]. This property of topic models led to impressive results in areas such as text clustering [1], corpora summarization, and recommender systems [4].

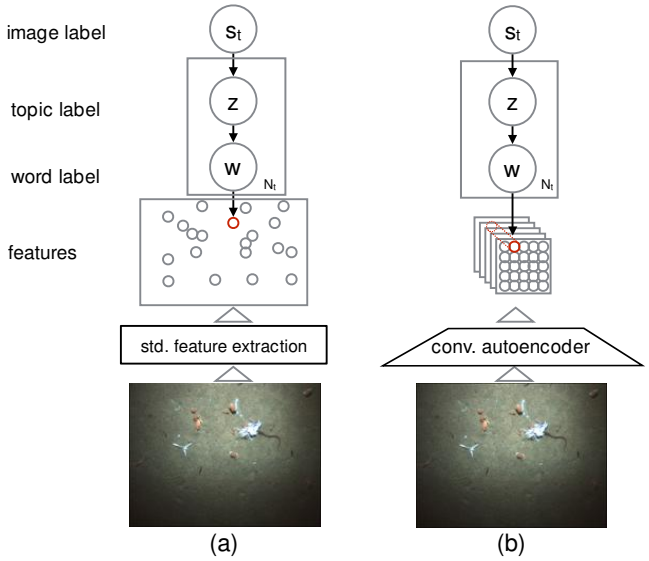

Fig. 1. Two scene modeling techniques evaluated in this paper. Here s_t is the image label, and z is the topic label of a visual word w in the input image. (a) A baseline spatiotemporal topic model using standard computer vision features as input. (b) The proposed spatio-temporal topic model using convolutional autoencoder-based features.

The success of Bayesian topic models for semantic understanding of text documents led to their adaptation to the computer vision domain. By replacing textural words with discrete visual features, LDA and HDP models can be applied directly to image data [5], [6], [7]. The most popular discretization of an image into visual words employs standard features such as SIFT [8], SURF [9], or Oriented BRIEF (ORB) [10]. Visual topic models have been used successfully in robotics applications for for unsupervised scene understanding [11] and adaptive mission planning [12].

However, the modeling capacity of topic models is fundamentally limited by the expressive power of the observed 'words'. Hand-crafted image features capture low-level patterns based on local image gradients. In contrast, deep neural models are often able to learn more complex, domain-specific features. Several papers have leveraged this property of neural networks to build more expressive models of textual data [13], [14]. In this paper, we make the natural extension

[1]Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology, Cambridge MA, 02139, USA {geflaspo, nickroy}@mit.edu

[2]Department of Applied Ocean Physics and Engineering, Woods Hole Oceanographic Institution, Woods Hole MA, 02543, USA yogi@whoi.edu

to unsupervised feature discovery for image data.

Much like textual data, image data show strong spatial correlations. These correlations are ignored by the simplifying bag of words (BOW) assumption made in most Bayesian topic models. Ideally, data features could encode these spatial correlations directly. Convolutional autoencoders (CAE) [15] preserve spatial relationships in data and hence are a powerful method for discovering useful features for image data. However, these features have not yet been incorporated into a topic modeling framework, in part because of the challenges of designing a CAE network architecture that produces useful features within a topic modeling context.

This work presents a CAE architecture that discovers feature representations directly from an image dataset and applies those features within an HDP-based topic modeling framework to discover cohesive visual topics. We explore the performance of this hybrid HDP-CAE model on the motivating application of autonomy and mission summarization for exploratory marine robots. We evaluate how well the topics discovered by the hybrid HDP-CAE model correspond to biologically distinct seafloor terrains and compare the hybrid HDP-CAE model to an HDP model using standard image features. Finally, we quantify the performance of the hybrid model on the secondary task of identifying anomalous images within a dataset. The probabilistic anomaly detection enabled by Bayesian topic models can inform more effective mission planning for exploratory marine robots and is more broadly useful for data summarization and visualization.

All models are evaluated on a realistic dataset that an individual biologist or data scientist could collect and wish to analyze, consisting of less than 4,000 seafloor images collected in-situ by a marine robot. Even in this small-data domain, we demonstrate that state-of-the-art performance can be achieved by applying neural feature discovery and nonparametric topic modeling to the task of unsupervised seafloor terrain characterization.

II. Related Work

Recent efforts have leveraged the power of neural models to discover data features within the context of topic modeling. Many, however, continue to rely on predefined data features at some level; frameworks that overcome this dependence have difficulty incorporating custom features within a completely unsupervised Bayesian topic model.

For textual data, Mikolvo et al. [13] propose the reverse of the architecture we present here; an LDA model is used to produce a contextual feature vector that is input into a recurrent neural network for contextually-aware language modeling. While powerful, this model does not address the model's dependence on standard word features and employs a BOW assumption. In [14], the BOW assumption is relaxed. Instead, a convolution operation maps variable length text sequences into a low-dimensional latent space. Unlike the work presented here, simple distance-based clustering is applied to discover semantically similar documents in place of a Bayesian topic model.

For image data, Wan et al. [16] introduce a hybrid neural-Bayesian topic model based on a Deep Boltzmann Machine (DBM). As in this work, the feature representation discovered by the DBM is fed directly into an HDP topic model to discover visual topics. However, instead of discovering features directly from image data, SIFT features are extracted from the image and the neural network learns an image representation based on these features, thereby not reducing the dependence on human-designed features.

The work most similar to our own is presented in [17]. The Hierarchical-Deep model introduced uses an HDP to learn priors over the activations of a DBM. In this way, the model is able to learn generic features from image data that enable learning image classes from very few examples. Each image in the model is annotated with a lower level class and the HDP discovers a hierarchy over these low-level classes. While suitable for the goal of one-shot learning, this supervision limits the generality of the Hierarchical-Deep model to a purely unsupervised problem. The model is also not convolutional, limiting the utility of the learned features. Convolutional autoencoders are directly able to model spatial correlations in image data and therefore are more suited to discover useful image representations.

Additionally, none of the aforementioned works evaluate their models on the small or medium-sized datasets that are prevalent in unsupervised learning applications. Instead, they use large (4 million+) standard image datasets [17] or 2D toy, simulated images [16].

Other works have incorporated neural feature learning for robotics applications outside of a topic modeling context. Naseer et al. [18] use up-convolutional networks to discover latent feature representations for the task of segmenting images. Rao et. al [19] use an autoencoder to learn features for classification of marine images. However, both of these works require human annotation and thus are not applicable in an unsupervised setting.

III. Methods

In the following sections, we provide a brief review of topic models and then discuss the two major components of the proposed hybrid HDP-CAE model: 1) a spatio-temporal HDP topic model and 2) a pipeline for discretizing an image into visual words using a convolutional autoencoder.

A. Bayesian topic models

Topic models [1], [20] seek to uncover semantic structure in a corpus of discrete data, segmented into documents. Topic models propose that each observed word in a document is generated by a latent topic and each document in a corpus has its own probability distribution over topics. Using word co-occurrences and distribution sparsity constraints, the distribution over topics z_i for each word w_i can be inferred. Under this model, the probability of the ith word w_i can be written as:

$$\mathbf{P}(w_i) = \sum_{k=1}^{K} \mathbf{P}(w_i|z_i = k)\mathbf{P}(z_i = k|d) \qquad (1)$$

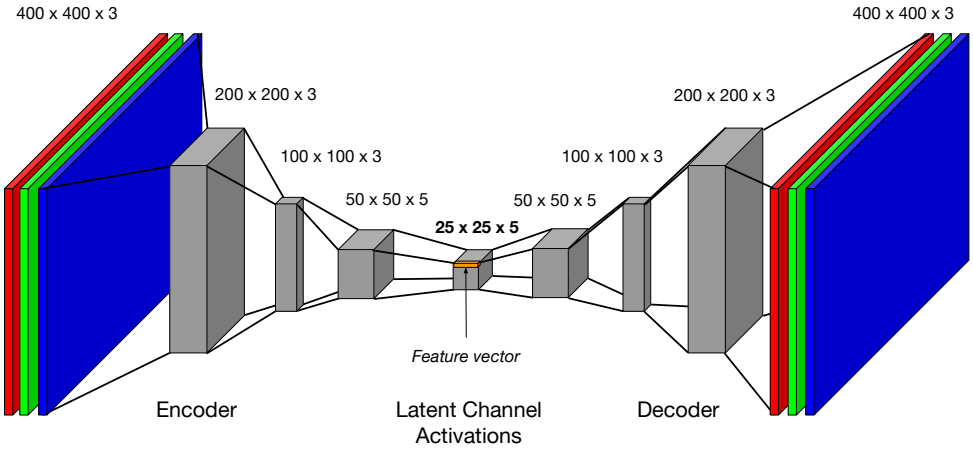

Fig. 2. Network architecture for the convolutional autoencoder (CAE) used to extract low level visual features from the image datasets. Network specific parameters were set as: training epochs (400 epochs), output channels (ordered by encoding layer 3-3-3-5-5 channels), stride (2), and convolutional filter size (ordered by encoding layer 10-10-3-3 pixels),

where K is total number of topics, $\mathbf{P}(w_i|z = k)$ is the probability of word i under topic k, and $\mathbf{P}(z_i = k|d)$ is the probability of topic z_i in the current document d.

B. Realtime spatio-temporal HDP model

Traditional topic modeling frameworks treat each word in a document as exchangeable. We instead adapt the ROST HDP model presented in [12], which relaxes the BOW assumption and explicitly models the correlation between spatio-temporal neighborhoods in a continuous image stream. ROST uses a Dirichlet process to model the growth in number of topics with the size and complexity of the data. A biased Gibbs sampler [21] enables online computation of the posterior distribution over topics for observed visual words.

The ROST model factors the probability of observing the visual word w_i at location x and time t in terms of the topic label variables z_i.

$$\mathbf{P}(w_i|x,t) = \sum_{k=1}^{K} \mathbf{P}(w_i|z_i = k)\mathbf{P}(z_i = k|x,t). \qquad (2)$$

The distribution $\mathbf{P}(w|z = k)$ is invariant to the spatio-temporal location of the observation, while $\mathbf{P}(z = k|x,t)$ models the distribution of topic labels in the spatio-temporal neighborhood of location (x,t). K is total number of topics that have at least at least one or more words assigned to them, plus one more to encode the possibility of creating a new topic for word w_i. ROST uses the Dirichlet distribution to model $\mathbf{P}(w|z)$, allowing for control of the sparseness of the topic model, whereas $\mathbf{P}(z|x,t)$ is modeled using the Chinese Restaurant Process [22], removing the need to predetermine the total number of unique topics.

C. Convolutional autoencoder architecture and training

To extract a discrete list of words from an image, we exploit the ability of neural models to discover useful abstract representations of data without supervision. We train a CAE architecture following the encoder/decoder paradigm described in [15]. The input image is first transformed into a lower dimensional bottleneck layer using successive convolution operations and rectified linear unit (ReLU) activations and then expanded back to its original size using a deconvolution operation with tied weight matrices. We call the channels in the bottleneck layer the latent channel activations (LCA).

The squared error between the original image and the image reconstruction provides an unsupervised loss function which allows the weight matrices for each layer to be learned. The network is trained using stochastic gradient descent with L2 regularization on the weight matrices. Because our goal is to treat the nodes in the bottleneck layer as non-overlapping features of the image, the neuron redundancy encouraged by dropout regularization is actually undesirable [23], so we do not include dropout. In this unsupervised setting, the entire dataset is used to train and test the model, so generalization is much less of a concern than in supervised learning problems. In this paper, all models are also trained without max-pooling/unpooling layers; dimensionality reduction is achieved using a stride greater than one and overlapping convolutional filter windows. During experimentation, we found that the inclusion of max-pooling and unpooling layers decreased the expressive power of the LCA, contrary to [15], so the final models were purely convolutional with ReLU nonlinearity.

The CAE network architecture used to produce the results in this paper is shown in Figure 2. The network consists of four encoding layers and four associated decoding layers. Each sequential encoding layer increases the number of output channels while decreasing the height and width of each individual channel, following a pyramid architecture. The architecture-specific parameters, such as number of training epochs (400 epochs), output channels (ordered by encoding layer 3-3-3-5-5 channels), stride (2), and convolutional filter size (ordered by encoding layer 10-10-3-

3 pixels), were determined empirically. After training, we remove the decoding layers of the network and use the LCA to generate low dimensional image features. The Tensorflow [24] implementation is available online[1].

D. Generating a visual vocabulary for topic models

HDPs require discrete data drawn from a vocabulary $\mathcal{V}$. To produce a vocabulary for CAE features, the CAE is trained on several example ocean mission datasets and the LCA for each image are extracted as described in Section III-C. This $25 \times 25 \times 5$ tensor is segmented into 625 feature vectors of length 5 by taking slices across LCA channels, as shown in Figure 2. These features are then clustered using the k-means algorithm into $|\mathcal{V}|$ clusters, where $|\mathcal{V}|$ is the desired vocabulary size. The centroid of each cluster represents a visual vocabulary word. Because the LCA are low dimensional (5×1 pixels using the architecture in Figure 2), as compared to 128-dimensional SIFT/SURF features, this clustering is relatively efficient.

Given a new image, visual features are extracted and mapped to the visual word $v_i \in \mathcal{V}$ corresponding to the nearest neighbor in the space of cluster centroids.

E. CAE feature visualization

The LCA discovered by our CAE model correspond to a low-dimensional, abstract representation of the image. In later sections we will apply these features within a topic modeling framework and attempt to visualize the properties of the latent channels constructed in this manner.

To quantify the strength of a latent channel's response to a particular input image, we consider the magnitude of each latent channel at a particular pixel location p_{ij} in the LCA. For each of the 5×5 pixels, we assign the pixel p_{ij} to the channel with the maximum activation at location (i, j). The magnitude of a latent channel M is equal to number of pixels for which it had the maximal value. This approch produces a clearer segmentation between channels than directly plotting channel magnitudes. Different channels have different baseline activations. To compare between channels, we normalize each channel's activation between their minimum value and maximum value before plotting.

IV. Experiments

We evaluate our hybrid HDP-CAE model against a ROST HDP baseline using SURF [9] and ORB [10] features. The two models are visualized in Figure 1. We apply each model to image streams from two marine robot missions [25] and present experimental results.

Mission I contains 1,117 images sampled every four seconds from a downwards facing camera mounted to the bottom of the robot. During Mission I, the robot passes over several seafloor terrains, including images of the water column, a rocky seafloor, and a porous sandy seafloor. This mission tests the model's ability to discover visually distinct terrain types.

[1] http://warp.whoi.edu/code/

Mission II consists of 2,296 images sampled in a similar manner. Mission II contains mostly images of a sandy seafloor, interrupted several times by large, biologically interesting phenomena, such as crab congregations, seafloor carnage, and geothermal vents. In addition to terrain discovery, this mission also tests the model's ability to accurately identify anomalous images within a mission dataset.

To evaluate how well the image topic labels discovered by the unsupervised HDP-CAE model correspond to visually meaningful seafloor terrains, we hand-labeled each image in both missions with one of thirteen possible terrain labels, including: 'water column', 'sparse boulders', 'smooth sand', 'biological congregation', etc. These labels are used exclusively for model evaluation.

We apply the HDP-ROST model described in Section III-B to the two mission datasets, using standard image features and CAE-derived LCA features for the standard HDP model and the hybrid HDP-CAE model respectively. New images are incorporated into the model in a streaming fashion; visual words are extracted from a new image and added to the model at regular intervals (200 ms). The ROST hyperparameters for Mission I and Mission II respectively were set to maximize mutual information between discovered topics and hand annotated labels: $\alpha = 0.1, 0.1; \beta = 25, 50; \gamma = 10^{-7}, 10^{-7}$. After Gibbs sampling, we have an approximation of the posterior over topics z_i for an observed visual world w_i, $P(z_i | w_i = v)$. The predicted topic label for each visual word is assigned as the *maximum a posteriori* (MAP) topic label given by the posterior, and the predicted scene label s_t for each image is calculated as the majority consensus of the visual words in the image.

Taking the MAP is a standard way of reducing a probabilistic distribution over a latent variable to a single point estimate. However, by using only the MAP topic label for each image, we are not allowing for important visual constructs to be represented by a mixture of topics. This approximation may be suitable for our experimental domain. The majority of images in our marine datasets consist of a homogeneous visual terrain, and an ideal topic model would discover topics rich enough to have nearly a one-to-one correspondence with semantically distinct visual constructs. Having to build a heuristic on top of a topic model to extract meaningful topics from mixtures of the discovered topics adds an unnecessary layer of complication to the model.

V. Results

To quantify the accuracy of the topic labels discovered by each of our models, we use normalized mutual information between the annotated topic distribution and the computed topic distribution. Mutual information captures the reduction in entropy of a random variable X after observing random variable Y (Eq. 3). A normalized mutual information score of one indicates that X and Y are completely dependent (i.e. the discovered topics are completely correlated with the true labels), whereas a mutual information score of zero indicates independence (i.e. the discovered topics are unrelated to the true labels).

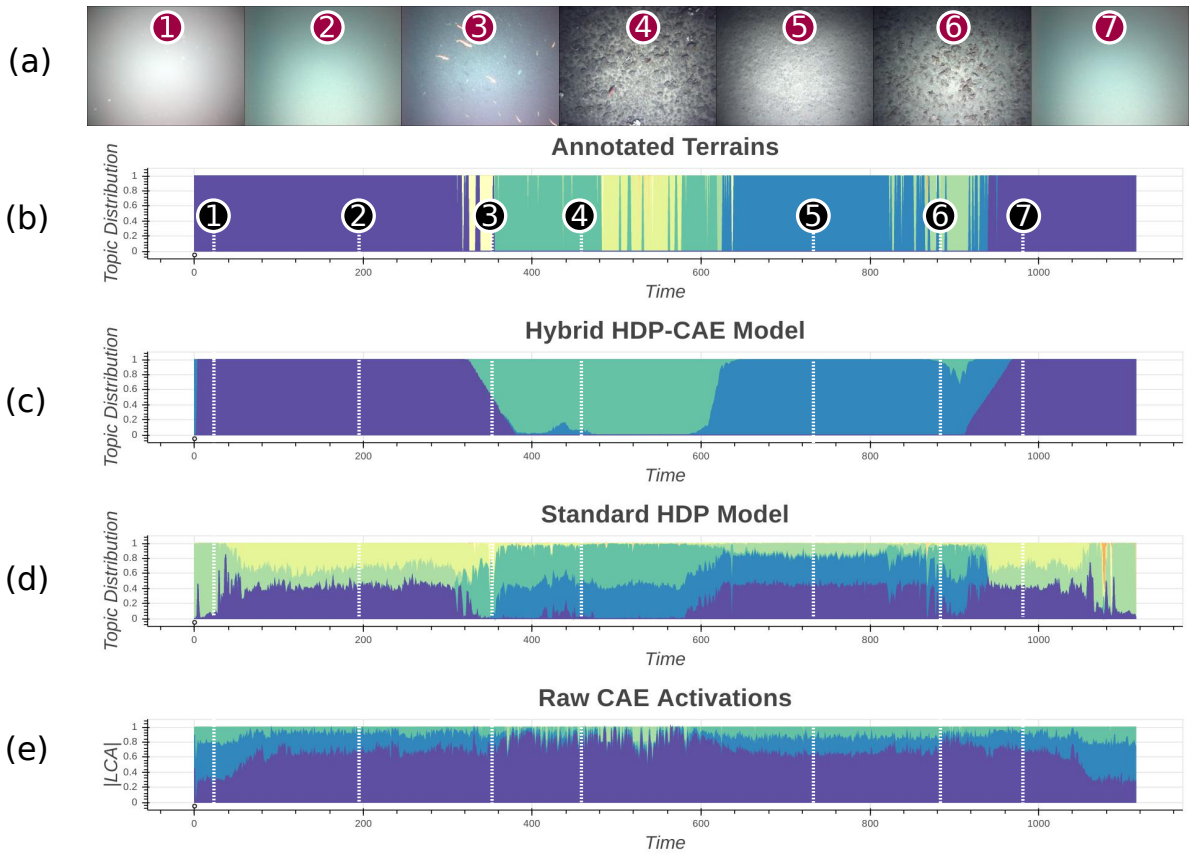

Fig. 3. Results for unsupervised topic models versus hand-annotated terrain labels (b) for the Mission I dataset. Example images from the dataset are shown in (a). To generate plots (c,d), visual words are extracted from an image at time t and assigned a topic label z_i as described in the text. The proportion of words in the image at time t assigned to each topic label is shown on the y-axis, where different topics are represented by colors. Colors are unrelated across plots. The hybrid HDP-CAE model (c), using more abstract features, is able to define topics that correspond more directly to useful visual phenomena than the HDP model using standard image features (d). The learned feature representation is visualized as described in the text in (e).

$$I(X,Y) = H(X) - H(X|Y) = \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \quad (3)$$

Interactive visualizations of the experimental results presented here are available online[2].

A. *Mission I - Seafloor terrain discovery*

Mission I tests the model's ability to uncover topics corresponding to meaningful visual terrains. Table I shows that the topics discovered by the hybrid HDP-CAE model are highly predictive of the ground-truth seafloor terrains. The raw topic distribution (before MAP reduction) for the two models is plotted in Figure 3 along with example images from the major terrain types. To generate the plots in Figure 3, visual words are extracted from an image at time t and assigned a topic label z_i as described in Section III. The proportion of words in the image at time t assigned to each topic label is shown on the y-axis, where different topics are represented by colors. Colors are unrelated across plots.

[2] http://warp.whoi.edu/iros2017/

Although the hybrid HDP-CAE model differs from the human annotated terrains by, for example, not modeling the transient topic at (3), the major terrain transitions are captured faithfully. The rocky seafloor terrain that dominates at (4) and then partially appears again at (6) is assigned to the same topic. The moment that the robot first observes the seafloor through the water column in (3) is captured as a mixture of the 'water column' topic (indigo) and the 'rocky seafloor' topic (green).

TABLE I
MUTUAL INFORMATION BETWEEN DISCOVERED TOPICS AND ANNOTATIONS

	Model	$I(X,Y)$
Mission I	Standard HDP	0.185
	Hybrid HDP-CAE	**0.535**
Mission II	Standard HDP	0.123
	Hybrid HDP-CAE	**0.441**

Despite its low mutual information scores, the standard HDP models does capture some of these transitions as changes in the topic distributions. However, it is not clear how to distill this information into meaningful topics. Be-

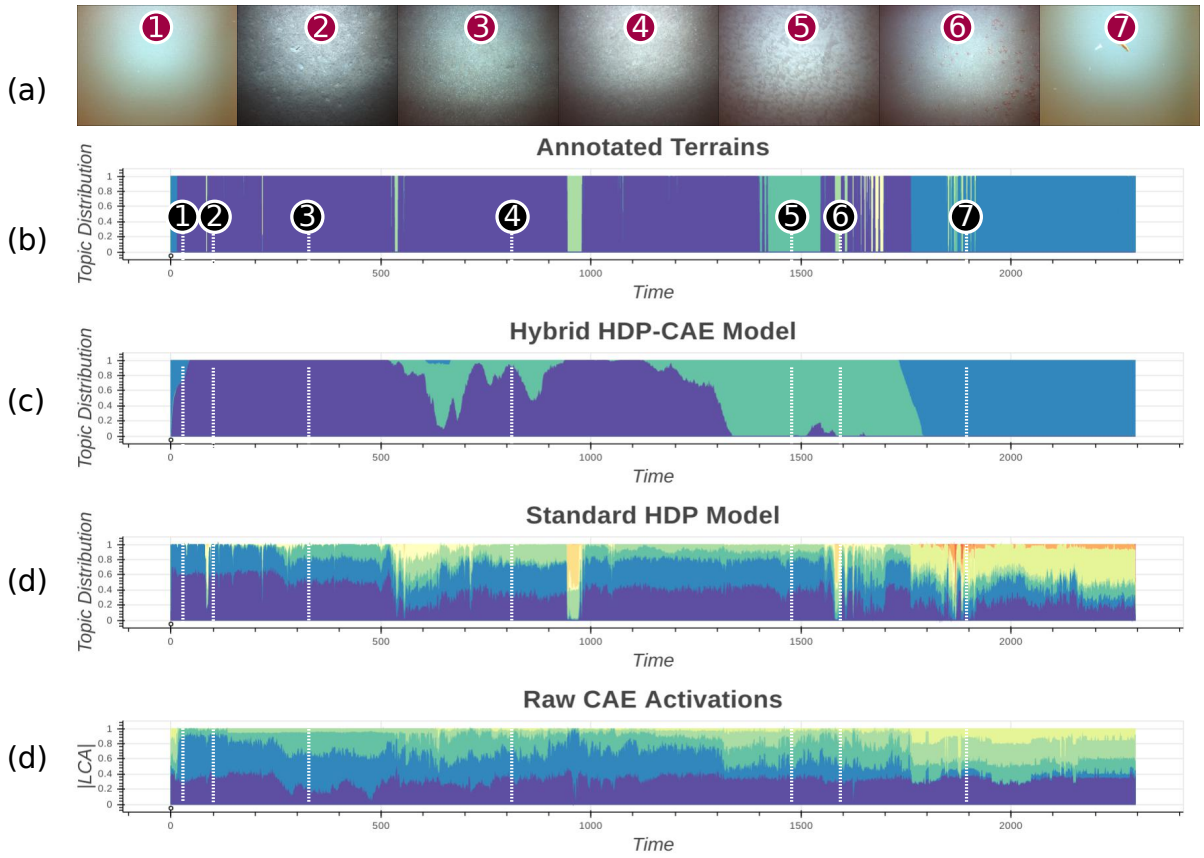

Fig. 4. Results for two unsupervised topic models versus annotated labels in (b) for the Mission II dataset. Example images from the dataset are shown in (a). The hybrid HDP-CAE model (c) again outperforms the HDP model using standard image features (d). However, the hybrid model fails recognize some of the more transient topics, such as the crustacean swarm at (7). The learned feature representation is visualized as described in the text in (e).

cause the hybrid HDP-CAE model uses much fewer, more abstract features, it is able to define topics that correspond more directly to useful visual phenomena. The discovered feature space is also visualized in Figure 3.

B. Mission II - Biological anomaly detection

Characterizing seafloor terrains is a vital task for an exploratory marine robot. Another complementary skill is the ability to identify images that are anomalous under the robot's current model of the world and flag these as interesting, one of the behaviors demonstrated by the standard HDP model presented in [12]. The Mission II dataset was designed to test both of these abilities.

By comparing mutual information between terrain labels, the hybrid HDP-CAE model again outperforms the baseline model. Figure 4 shows the raw topic distribution for the two models, along with example images from the major terrain types. Mission II is a more visually homogeneous dataset, consisting almost entirely of sandy seafloor images. The hybrid HDP-CAE model discovers segmentations within the sandy seafloor topic that the human annotators do not, but otherwise captures major terrain transitions. However, the hybrid HDP-CAE model does not capture some of the more transient topics, such as the crustaceans at (7).

To quantify this result further, we introduce the notion of perplexity. Because HDP is a probabilistic model, it is straightforward to quantify the average word perplexity (Eq. 4) of a new image X_t under the model.

$$Perplexity(X_t) = \exp\left(-\frac{\sum_{w \in W_t} \log p(w|X_t)}{|W_t|}\right) \quad (4)$$

where the set W_t consists of the visual words in X_t. High perplexity indicates that the image is not well modeled by the topic model, whereas low perplexity indicates an image that is well explained by the topic model. Figure 5 compares the perplexity response of each model when presented with biologically interesting images; the hybrid model does not have an obvious increase in perplexity when presented with the images of seafloor carnage (2), crab congregations (3), or submerged tree (4).

To compute how well each model's perplexity score corresponds to some interesting visual phenomena, we annotated each image in the dataset as either high, medium, or low 'interest' and computed the mutual information between the annotated and computed perplexity. Although perplexity scores do not necessarily correspond well with human intuition about semantic coherency [3], perplexity has been used successfully for anomaly detection in previous work [12]. The CAE is not a probabilistic model, so there is no

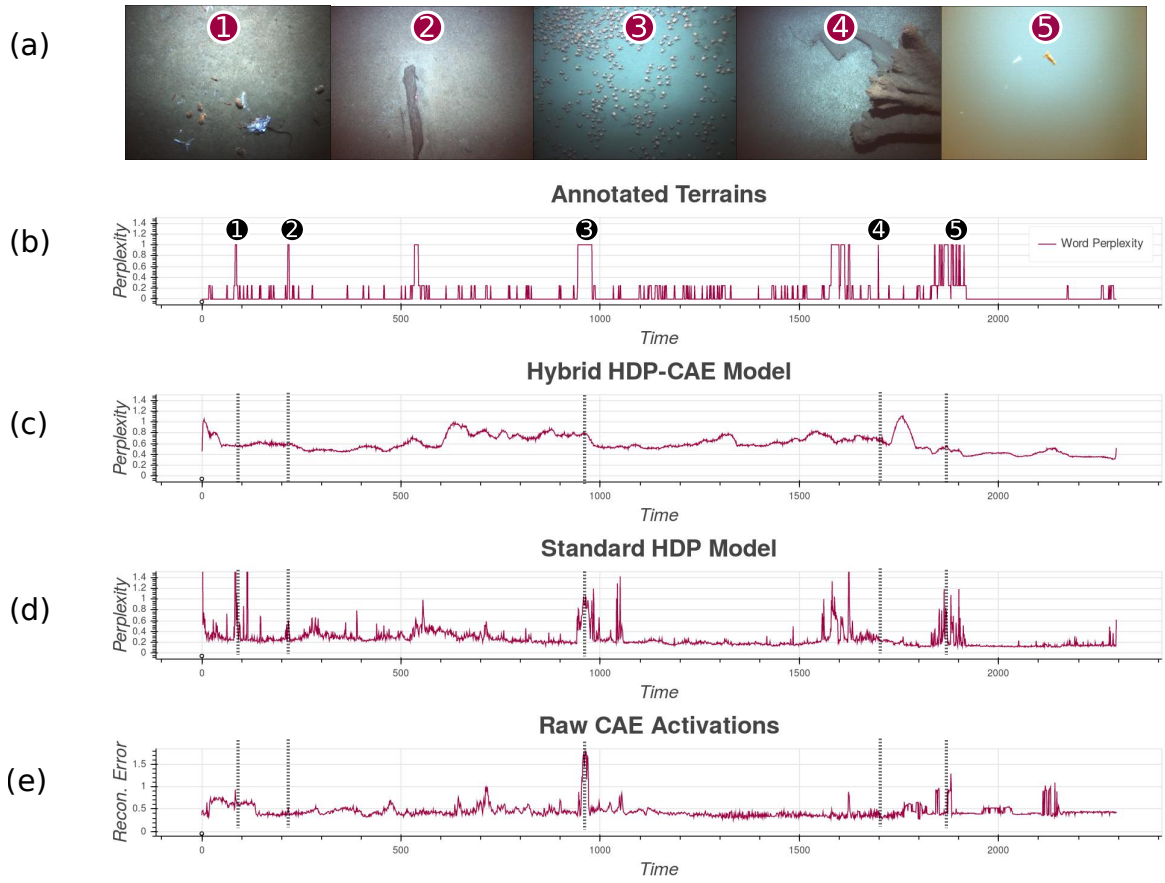

Fig. 5. Correlation of perplexity score of the models (c,d,e) with annotated biological anomalies (b) for the Mission II dataset. Example images of biological anomalies are shown in (a). Each image in the dataset was labeled with high, medium, or low perplexity (b). Although all three models do have differential responses in areas of high perplexity, the HDP model using standard features (d) outperforms the hybrid HDP-CAE model (c) and raw reconstruction error from the CAE (e).

well-defined operation to compute perplexity. We instead use squared image reconstruction error as a proxy for perplexity. The mean μ and standard deviation σ^2 of each model's perplexity distribution are used to bin the perplexity into low ($0 \leq x \leq \mu + \sigma^2$), medium($\mu + \sigma^2 \leq x \leq \mu + 2\sigma^2$), and high ($x > \mu + 2\sigma^2$) perplexity. The results of this analysis are shown in Table II. Although all three models do have differential responses in areas of high annotated perplexity, the standard HDP model's perplexity has higher mutual information with annotated perplexity. We will address this discrepancy in the Discussion.

TABLE II
MUTUAL INFORMATION BETWEEN PERPLEXITY AND ANNOTATIONS

	Model	$I(X,Y)$
Mission II	**Standard HDP**	**0.153**
	Hybrid HDP-CAE	0.006
	Raw CAE	0.033

VI. DISCUSSION

The proposed hybrid HDP-CAE model significantly outperforms alternative models on the task of seafloor terrain discovery. The hybrid model, however, did not perform as well on the secondary task of anomaly detection, as quantified by image perplexity. Our hypothesis is that this limitation stems from the inherently imbalanced nature of anomalies within a dataset. In our anomaly detection experiments, the CAE was trained on over 2000 images of sandy seafloor and only 80 images of crab congregations. Many other anomalous events, such as the seafloor carnage in Figure 5(a) appear for even shorter spans. Neural models have been shown to struggle when presented with imbalanced training data [26]. There may not be enough images of anomalous biological events for the CAE to learn a meaningful feature representation. Although the CAE image reconstruction error, plotted in Figure 5(d), does capture the inability of the features to represent the anomalous images, our current hybrid HDP-CAE model does not incorporate this uncertainty within the topic modeling stage. An interesting extension of this work would be to use image reconstruction error directly as a metric of feature quality. Alternatively, there are methods within the machine learning community for dealing with imbalanced datasets that could improve CAE training.

For the specific biological anomalies tested here, such as the crab congregations or seafloor carnage, it may be difficult to outperform standard image features, which are designed to

detect areas of high image gradient. However, there are other reasons to prefer a CAE-based anomaly detector. Interesting anomalies may not always manifest themselves as complex visual structure; a smooth sandy seafloor is anomalous within a rocky mission. Standard image features may struggle to represent these visually simple anomalies. Additionally, CAE-based anomaly detectors can use reconstruction error to not only detect when an image is anomalous, but also which part of the image is particularly difficult to resolve. The ability to spatially localize anomalies could be very useful in robot scene understanding and real-time planning.

Another important extension to the hybrid HDP-CAE model presented in this work is the adaptation of convolutional feature discovery for realtime, streaming applications. Bayesian nonparametric models are well suited for the life-long learning required in streaming and robotics applications; this is one compelling reason to use them over purely neural models. However, for simplicity, the CAE-based feature discovery training in this work was done offline on complete datasets. Exploring methods for efficient, life-long training of convolutional models is an important area of future work for applying hybrid HDP-CAE models to realtime applications.

VII. Conclusions

Bayesian topic models have achieved impressive performance by learning both model parameters and useful structure directly from data. However, these nonparametric models still fundamentally rely on predefined feature representations of data. We present a novel model that overcomes this limitation using convolutional autoencoders, allowing unsupervised discovery of *both* a feature representation and thematic structure in image data.

Our proposed hybrid model incorporates a convolutional autoencoder for data-driven feature discovery within a Bayesian topic modeling framework. We apply this model to the problem of high-level scene understanding and mission visualization for exploratory marine robots. On complex mission datasets, the hybrid model discovers a rich latent visual structure that has over four times the mutual information with biologically meaningful seafloor terrains when compared to a Bayesian nonparametric topic model with standard, hand-designed features. This work defines a paradigm for including the ability of unsupervised neural models to discover useful, low-dimensional data representations within a Bayesian nonparametric topic modeling framework and demonstrates state of the art performance on a challenging problem from the marine robotics community. Our future work will focus on adapting the convolutional model to run in real-time on a marine robot and improving the model's ability to detect visual anomalies in an image dataset.

Acknowledgements

This work was supported in part by an NSF Graduate Research Fellowship Program award and The John P. Chase Memorial Endowed Fund.

References

[1] D. M. Blei, B. B. Edu, A. Y. Ng, A. S. Edu, M. I. Jordan, and J. B. Edu, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[2] Y. W. Teh and M. I. Jordan, "Hierarchical Bayesian Nonparametric Models with Applications," *Bayesian nonparametrics*, pp. 158–207, 2010.

[3] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei, "Reading Tea Leaves: How Humans Interpret Topic Models." in *NIPS*, 2009, pp. 1–9.

[4] R. Krestel, P. Fankhauser, and W. Nejdl, "Latent Dirichlet Allocation for Tag Recommendation."

[5] A. Bosch, A. Zisserman, and X. Muñoz, "Scene Classification Via pLSA," in *Computer Vision – ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds. Springer Berlin / Heidelberg, 2006.

[6] L. Fei-Fei and P. Perona, "A Bayesian Hierarchical Model for Learning Natural Scene Categories," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2. IEEE, 6 2005, pp. 524–531.

[7] X. Wang and E. Grimson, "Spatial Latent Dirichlet Allocation," in *Advances in Neural Information Processing Systems*, vol. 20, 2007, p. 15771584.

[8] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, 2004.

[9] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded Up Robust Features," in *European Conference on Computer Vision*, 2006.

[10] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *2011 International Conference on Computer Vision*. Barcelona: IEEE, 11 2011, pp. 2564–2571.

[11] D. M. Steinberg, O. Pizarro, and S. B. Williams, "Hierarchical Bayesian Models for Unsupervised Scene Understanding."

[12] Y. Girdhar, Walter Cho, M. Campbell, J. Pineda, E. Clarke, and H. Singh, "Anomaly detection in unstructured environments using Bayesian nonparametric scene modeling," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 5 2016, pp. 2651–2656.

[13] T. Mikolov and G. Zweig, "Context Dependent Recurrent Neural Network Language Model," 2012.

[14] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, "A Latent Semantic Model with Convolutional-Pooling Structure for Information Retrieval," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management - CIKM '14*. New York, New York, USA: ACM Press, 2014, pp. 101–110.

[15] J. Masci, U. Meier, D. Cirean, and J. Schmidhuber, "Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction."

[16] L. Wan, L. Zhu, and R. Fergus, "A Hybrid Neural Network-Latent Topic Model."

[17] R. Salakhutdinov, J. B. Tenenbaum, and A. Torralba, "Learning to Learn with Compound HD Models."

[18] T. Naseer, G. L. Oliveira, T. Brox, and W. Burgard, "Semantics-aware Visual Localization under Challenging Perceptual Conditions."

[19] D. Rao, M. De Deuge, N. Nourani-Vatani, B. Douillard, S. B. Williams, and O. Pizarro, "Multimodal learning for autonomous underwater vehicles from visual and bathymetric data," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, may 2014, pp. 3819–3825.

[20] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. Suppl 1, pp. 5228–5235, 2004.

[21] Y. Girdhar and G. Dudek, "Gibbs Sampling Strategies for Semantic Perception of Streaming Video Data," *ArXiv e-prints*, p. 7, 2015.

[22] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 12 2006.

[23] N. Srivastava *et al.*, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[24] M. Abadi *et al.*, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems."

[25] J. Pineda *et al.*, "A crab swarm at an ecological hotspot: patchiness and population density from AUV observations at a coastal, tropical seamount," *PeerJ*, vol. 4, p. e1770, apr 2016.

[26] N. V. Chawla, N. Japkowicz, and A. Ko, "Editorial: Special Issue on Learning from Imbalanced Data Sets."