HIERARCHICAL PROBABILISTIC EMBEDDINGS FOR MULTI-VIEW IMAGE CLASSIFICATION

Benjamin Brodie, Subash Khanal, Muhammad Usman Rafique, Connor Greenwell, Nathan Jacobs

Department of Computer Science, University of Kentucky

ABSTRACT

We address the task of image classification, when the available spectral bands can vary from image to image. We propose a model that learns to represent uncertainty over latent features in a way that is conditioned on the available bands. We expect that images with fewer bands will generally be more difficult to classify and hence have higher uncertainty. We compare two strategies for training such a model, one which uses explicit hierarchical constraints and one which relies on implicit constraints. We evaluate both using RGB and multispectral imagery from the EuroSat dataset and find that the hierarchical approach improves the compatibility of the resulting distributions without sacrificing accuracy.

1. INTRODUCTION

In many scenarios, we may have incomplete information about a given scene. For example, if trying to identify landscape features, we may have RGB imagery available for some areas, but only IR imagery for others, or vice versa. We show that we can take advantage of the cases where we have complete sets of information to make more informed predictions when presented with partial information. In most such work, the features of a scene are represented by a vector in a latent space. We move to a probabilistic setting, and capture the inherent uncertainty in measurement by representing features with a Gaussian probability distribution over the latent vector space.

We define a *view* of a scene to be a collection of image channels. A *sub-view* has a subset of the channels of a given view, and a *super-view* contains more channels than the view. We employ a hierarchical training scheme, which enforces a relationship between feature distributions corresponding to sub-views and super-views. We demonstrate that this training scheme provides us with a natural progression of feature distributions in which uncertainty decreases as the the amount of information increases. In addition, we show that enforcing the hierarchy leads to more accurate predictions on downstream classification tasks.

We demonstrate our method on RGB images and 13 channel multispectral bands from the EuroSat land classification dataset [1]. To represent partial information, we divide the overhead RGB images into separate R, G, and B channels, as well as the combinations RG, RB, and GB. In this way, for each RGB image, we have seven different views fitting into a sub-view/super-view hierarchy. The split for the full multispectral version is discussed in Section 4.3.

Our contributions include: (1) a method using Gaussian distributions to capture uncertainty in feature representations given incomplete information; (2) a hierarchical training scheme, based on the relationship between views and subviews of scenes; and (3) evaluation of our method on accuracy and uncertainty quantification.

2. RELATED WORK

Multi-view Feature Embedding: Image feature embedding aims aims to learn a mapping from a set of images to a low-dimensional metric space such that the mappings of similar images are close together and those of dissimilar images are far apart [2, 3]. Multi-view embedding has been considered in various contexts, including action recognition [4], image geolocalization [5, 6, 7, 8, 9], and 3D geometry [10]. Methods based on auto-encoders [4, 10] learn a common representation space where examples are recoverable through view-specific decoders. In contrast, retrieval based methods [5, 6, 7, 8, 9, 11] are optimized for finding matching examples from other viewpoints through a nearest neighbors lookup. In settings where there is a known geometric relationship between views, geometric operations can be embedded into the model or the loss to improve performance [6, 7, 8].

Probabilistic Embedding: Reasoning about distributions of representations in cases where there is possible uncertainty in matched data [10, 12, 13, 14] has proven to be effective in a variety of situations. For face recognition, predicting an embedding and an associated variance was found to improve matching accuracy between low quality and high quality images [12, 14]. Probabilistic embeddings have also been used in autoencoder based embedding learning [15] and image segmentation [16].

3. PROBLEM STATEMENT AND APPROACH

Given multiple views of a scene, for example an RGB and an IR image, we aim to jointly learn a feature representation of

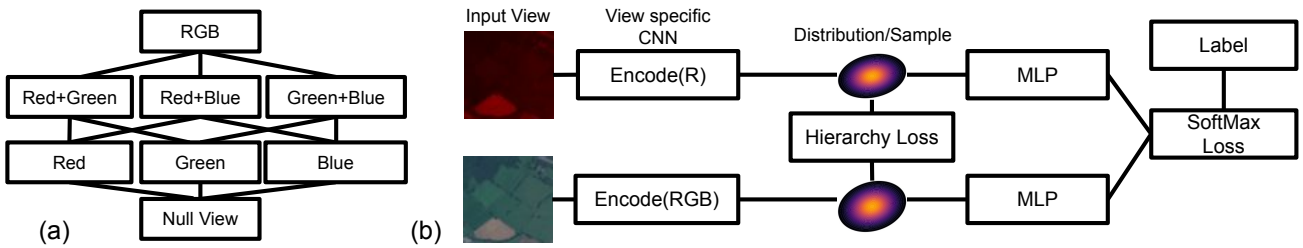

Fig. 1: (a) Nested hierarchy of views of RGB channels for EuroSat data. During training, we enforce the hierarchy by computing Bhattacharyya distance between distributions connected by an edge in the graph. (b) Training flow for hierarchical probabilistic embeddings using R and RGB channels. The input views are passed through parallel encoders which produce the parameters of a Gaussian distribution. Each encoder is a separate ResNet18. The weights on the MLP classifier are shared.

both views that captures the uncertainty inherent in the input view. We could also learn an embedding given both views as input. In this case, we expect the representation from the sub-views, RGB and IR, to be less certain than the representation from the joint-view, RGB + IR.

We learn a representation for the features f of an object as a probability distribution over a latent vector space in $\mathbb{R}^d$. We show that the uncertainty of the distribution reflects the amount of information presented by the input view. The hierarchy of views and sub-views forms a directed acyclic graph. See Figure 1 for an example of the directed acyclic graph of views for RGB channels.

3.1. Hierarchy of Probability Distributions

Views and sub-views obey a hierarchical ordering. The feature distributions corresponding to a view and its sub-views should also obey such a hierarchical ordering, where samples from the distribution corresponding to a super-view look like samples from a sub-view distribution.

We take all feature distributions to be multivariate Gaussian with diagonal covariance. Let v_0 be the *null*-view, with no input information. This is considered to be a sub-view of any other view. The probability distribution for v_0 is defined to be $P(f|v_0) = N(\mathbf{0}, \mathbf{1})$.

3.2. Multi-Modal Embedding Architecture

We treat each image channel and/or grouping of input channels as a separate view. The inputs are fed through a distinct ResNet18 CNN [17] corresponding to the input view. The output of the ResNets are then given to a shared probability module. The module consists of two fully connected layers which produce the parameters of the multivariate Gaussian. We then classify by passing a sample from the Gaussian distribution through a small MLP. For an overview of the training process for two views, see Figure 1.

3.3. Training Process

The loss function is split into: classification, hierarchy, and marginal statistics. It is given by

$$L = L_c + \lambda_h L_h + \lambda_m L_m$$

where the weights λ_h and λ_m are fixed before training.

We calculate the classification loss, L_c by drawing a sample from each feature distribution, and classifying using Soft-Max Loss against the label. The reparametrization trick [18] is used to sample from the distribution, allowing backpropagation.

The hierarchical loss, L_h, represents the loss of information between nested views of the same image. We take the Bhattacharyya distance between distributions that share a super-view/sub-view relationship. Although we expect the hierarchy of probability distributions to arise naturally during training, we find explicit enforcement of the hierarchy through direct comparison leads to a more stable training process and encourages learning of aligned distributions with a trend of decreasing uncertainty (see Section 4.2). The Bhattacharyya distance is fast to compute, and we have found it to be more effective than several natural alternatives, such as Earth Mover's Distance or KL Divergence.

The marginal statistics loss, L_m, also serves as a regularization term for the distributions. The marginal statistics for a view v_i with respect to the null-view v_0, are given by $N(\mathbf{0}, \mathbf{1}) = p(f|v_0) = \int_{v_i} p(f|v_i)p(v_i)dv_i$. That is, the average distribution corresponding to the view v_i across the dataset will be standard normal. In practice, we enforce the marginal statistics at the minibatch level. Given a minibatch of size N, corresponding to views $\{v_k^{(i)}\}_{i=1}^N$, we first find the Gaussian that minimizes KL-divergence to the mini-batch mixture $\frac{1}{N}\sum_{i=1}^N P(f|v_k^{(i)})$, where $P(f|v_k^{(i)}) = N(a_i, A_i)$. This is given by $N(b, B)$ with:

$$b = \sum_{i=1}^{N} \frac{1}{N} a_i, \quad B = \sum_{i=1}^{N} \frac{1}{N} (A_i + a_i a_i^\top - b b^\top).$$

We then enforce the marginal statistics during training with

$$L_m = D_{KL}(N(b, B), N(\mathbf{0}, \mathbf{1})),$$

where D_{KL} is the KL-divergence.

4. EVALUATION

We evaluate our approach on the EuroSat dataset, which consists of Sentinel-2 imagery over 34 European countries. Patches from these images were extracted and classified into one of 10 classes: industrial buildings, residential buildings, annual crop, permanent crop, river, sea & lake, herbaceous vegetation, highway, pasture, and forest. The patches measure 64×64 pixels. There are 27,000 images in the EuroSat dataset. We use an 80/20 split of training to testing data. We examine our methods on both RGB and 13-channel multispectral images. For RGB imagery, we take advantage of a complete set of views and sub-views, by using each combination of R, G, and B channels. For multispectral data, we use the selected grouping of channels discussed in Section 4.3.

4.1. Implementation Details

We first train the multi-modal embedding network on the directed acyclic graph of RGB views, as depicted in Figure 1. We set the weights in the loss function to $\lambda_h = .05$ and $\lambda_m = 1$. For ease of training and visualization, we embed features as Gaussian distributions over a 2-dimensional latent space.

For comparison, we train the same architecture without enforcing the hierarchy. In this case, we take only the Bhattacharyya distance from each distribution to the standard normal distribution corresponding to the null view. Note that this is similar to the training method used for probabilistic face embeddings described in [14] with Bhattacharyya distance replacing KL-divergence.

4.2. Results

We measure the accuracy and average variance for the multi-view embeddings in Table 1. When trained with hierarchy enforcement, the accuracy increases and average variance decreases as more views are given to the network. We compare this to another model with the same architecture trained without enforcing the hierarchy. In this case, the accuracy follows the expected progression as views increase, but the average variance does not. This suggests that without enforcing the hierarchy during training, the learned distributions coming from different views are not directly comparable to each other. The network is learning the embeddings independently, and so uncertainty in input view is not reflected by the variance of the feature distributions.

We also include a comparison of log-likelihoods to measure the nesting of distributions. We take the distribution from each sub-view and calculate the average log probability at 100 samples of the distribution from the full RGB view. The results are averaged throughout the test set. The higher average log-likelihood when hierarchy is enforced demonstrates that samples from super-view distributions are compatible with samples from sub-view distributions, whereas this is not the case without hierarchy enforcement.

Another measure of the view-based uncertainty is shown in Figure 2. For each test image, we take 50,000 samples from the feature distribution and plot the number of unique predicted classifications resulting from these samples. The majority of the time, the distribution from an RGB image predicts a single classification, whereas the distribution coming from a single channel view more often has two or more possible classifications.

Channels	Accuracy	Avg. Variance	Log-likelihood
Hierarchy Enforced			
R	0.8515	0.0239	-3.8314
G	0.8378	0.0230	-5.8161
B	0.7641	0.0232	-3.9687
RG	0.8954	0.0217	-1.4938
RB	0.8878	0.0215	-2.000
GB	0.9044	0.0211	-1.749
RGB	0.9270	0.0200	-
Hierarchy Not Enforced			
R	0.8559	0.0172	-12.1669
G	0.8546	0.0213	-11.7454
B	0.7267	0.0174	-16.8876
RG	0.8920	0.0183	-6.2442
RB	0.8874	0.0205	-5.4862
GB	0.9000	0.0172	-4.3706
RGB	0.9243	0.0203	-

Table 1: Accuracy, average uncertainty, and average log-likelihood on RGB channels from EuroSat, using a 2-dimensional embedding. For log-likelihood we compute the average log-likelihood of 100 samples of the full RGB view.

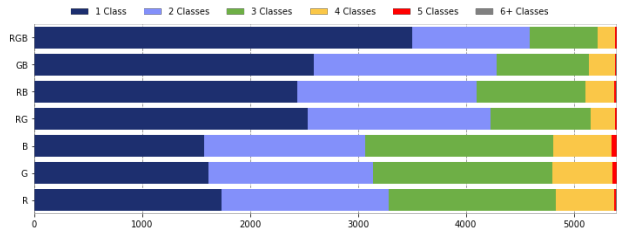

Fig. 2: For each test image, we extract a feature distribution and take 50,000 samples before running each sample through the classifier. The number of unique predicted classes among the samples is presented.

Hierarchy Enforced			
Channels	Accuracy	Avg. Variance	Log-likelihood
Aerosol (1)	0.5753	0.0331	-6.000
Vapor/Cirrus (2)	0.7948	0.0380	-1.7700
RGB (3)	0.9241	0.0439	-1.580
Red Edge (4)	0.9066	0.0412	-0.4456
NIR (1)	0.8430	0.0386	-1.1418
SWIR (2)	0.9107	0.0414	0.0766
Atmosphere (3)	0.8839	0.0287	-0.4380
RGB + Edge (7)	0.9312	0.0270	-0.4892
IR (3)	0.9222	0.0266	-.1499
Full (13)	0.9628	0.0260	-
Hierarchy Not Enforced			
Aerosol (1)	0.5958	0.0134	-77.0801
Vapor/Cirrus (2)	0.7984	0.0193	-51.4137
RGB (3)	0.8402	0.0208	-20.4798
Red Edge 1-4 (4)	0.9139	0.0224	-9.9356
NIR (1)	0.8413	0.0188	-39.7989
SWIR (2)	0.8923	0.0199	-25.3889
Atmosphere (3)	0.8371	0.0214	-24.0449
RGB + Edge (7)	0.9369	0.0255	-8.1741
IR (3)	0.8735	0.0191	-25.41
Full (13)	0.9561	0.0236	-

Table 2: Accuracy, average uncertainty, and average log-likelihood on the 13-channel multispectral EuroSat dataset, using a 2-dimensional embedding. Number of image channels is indicated in parentheses.

4.3. Multispectral Data

For our final evaluation, we use all 13 spectral bands of the EuroSat dataset. Due to the exponential number of possible sub-views given the quantity of channels, we do not use all possible views, but instead break the data into the collection of aligned, nested views seen in Table 2. Here the Atmosphere view contains the Aerosol, Vapor, and Cirrus channels, IR contains both NIR and SWIR channels, and RGB+Edge contains the three color channels as well as four Red Edge channels. We compare trends in accuracy and average variance, as well log-likelihoods to the full 13-channel view in Table 2. Again, we see a natural trend of variances and log-likelihoods only when enforcing the hierarchy during training.

5. CONCLUSION

We introduced an approach for learning a hierarchical, probabilistic feature embedding where we expect varying quantities of information at inference time. Our approach makes it possible to achieve uncertainty estimations for feature distributions coming from sources with variable bands of information.

6. REFERENCES

[1] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," 2017. 1

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020. 1

[3] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim, "A metric learning reality check," *ECCV*, 2020. 1

[4] Y. Kong, Z. Ding, J. Li, and Y. Fu, "Deeply learned view-invariant features for cross-view action recognition," *IEEE Transactions on Image Processing*, 2017. 1

[5] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays, "Composing text and image for image retrieval-an empirical odyssey," in *CVPR*, 2019. 1

[6] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li, "Where am I looking at? Joint Location and Orientation Estimation by Cross-View Matching," *arXiv preprint arXiv:2005.03860*, 2020. 1

[7] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li, "Spatial-aware feature aggregation for image based cross-view geo-localization," in *NIPS*, 2019. 1

[8] Krishna Regmi and Mubarak Shah, "Bridging the domain gap for ground-to-aerial image matching," in *ICCV*, 2019. 1

[9] Sixing Hu, Mengdan Feng, Rang MH Nguyen, and Gim Hee Lee, "Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization," in *CVPR*, 2018. 1

[10] Sanjeev Muralikrishnan, Vladimir G Kim, Matthew Fisher, and Siddhartha Chaudhuri, "Shape unicode: A unified shape representation," in *CVPR*, 2019. 1

[11] Yonglong Tian, Dilip Krishnan, and Phillip Isola, "Contrastive multiview coding," *ArXiv*, 2019. 1

[12] Tianyuan Yu, Da Li, Yongxin Yang, Timothy M Hospedales, and Tao Xiang, "Robust person re-identification by modelling feature uncertainty," in *ICCV*, 2019. 1

[13] Yichun Shi and Anil K Jain, "Probabilistic face embeddings," in *ICCV*, 2019. 1

[14] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei, "Data uncertainty learning in face recognition," in *CVPR*, 2020. 1, 3

[15] Xudong Lin, Yueqi Duan, Qiyuan Dong, Jiwen Lu, and Jie Zhou, "Deep variational metric learning," in *ECCV*, 2018. 1

[16] Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger, "A probabilistic u-net for segmentation of ambiguous images," *NIPS*, 2018. 1

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016. 2

[18] Diederik Kingma and Max Welling, "Auto-encoding variational bayes," in *ICLR*, 2014. 2