Discovering Disentangled Representations with the F Statistic Loss

Karl Ridgeway
Department of Computer Science
University of Colorado, Boulder
karl.ridgeway@colorado.edu

Michael C. Mozer
Department of Computer Science
University of Colorado, Boulder
mozer@colorado.edu

Abstract

We propose and evaluate a novel loss function for discovering deep embeddings that make explicit the categorical and semantic structure of a domain. The loss function is based on the F statistic that describes the separation of two or more distributions. This loss has several key advantages over previous approaches, including: it does not require a margin or arbitrary parameters for determining when distributions are sufficiently well separated, it is expressed as a probability which facilitates its combination with other training objectives, and it seems particularly well suited to disentangling semantic features of a domain, leading to more interpretable and manipulable representations.

In typical classification tasks, the input features—whether images, speech, text, or other measurements—contain only implicit information about category labels, and the job of a classifier is to transform the input features into a representation that makes category labels explicit. The traditional representation has been a *localist* or one-hot encoding of categories, but an alternative approach has recently emerged in which the representation is a *distributed* encoding in a high dimensional space that captures category structure via metric properties of the space. The middle panel of Figure 1 shows a projection of instances from three categories to a two-dimensional space. The projection separates inputs by category and therefore facilitates classification of unlabeled instances via proximity to the category clusters. Such a *deep embedding* also allows new categories to be 'learned' with a few labeled examples that are projected to the embedding space. The literature is somewhat splintered between researchers focusing on deep embeddings which are evaluated via k-shot learning [1, 2, 3] and researchers focusing on k-shot learning who have found deep embeddings to be a useful method [4, 5].

Figure 1 illustrates a fundamental trade off in formulating an embedding. From left to right frames, the intra-class variability increases and the inter-class structure becomes more conspicuous. In the leftmost panel, the clusters are well separated but the classes are all equally far apart. In the rightmost panel, the clusters are highly overlapping and the blue and purple cluster centers are closer to one another than to the yellow. Separating clusters is desirable, but so is capturing inter-class similarity. If this similarity is suppressed, then instances of a novel class will not be mapped in a sensible manner—a manner sensitive to input features, semantic features, and their correspondence. The middle panel reflects a compromise between discarding variability between instances of the same class and preserving relationships among the classes. With this compromise, deep embeddings can

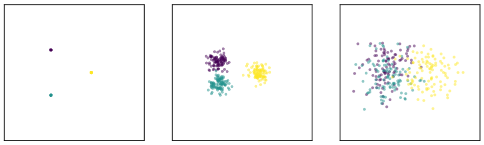

Figure 1: Alternative two-dimensional embeddings of instances of three categories. Points represent instances and color the category label. In the leftmost frame, the points are superimposed on one another.

be used to model hierarchical category structure and can facilitate partitioning the instances along multiple dimensions, e.g., disentangling content and style [6].

The trade off in Figure 1 points to a challenge in constructing deep embeddings. Some existing methods aim to perfectly separate categories in the training set [1], which may not be appropriate if there are labeling errors or noise in the data. Other methods require a margin or other parameter to determine how well separated the categories should be in order to prevent overfitting [7, 2, 3, 8]. We propose a new method that automatically balances the trade off using the currency of probability and statistical hypothesis testing. It also manages to align dimensions of the embedding space with categorical and semantic features, thereby facilitating the disentangling of representations.

1 Using the F statistic to separate classes

For expository purposes, consider two classes, $\boldsymbol{C} = \{1, 2\}$, having n_1 and n_2 instances, which are mapped to a one-dimensional embedding. The embedding coordinate of instance j of class i is denoted z_{ij}. The goal of any deep embedding procedure is to separate the coordinates of the two classes. In our approach, we will quantify the separation via the probability that the true class means in the underlying environment, μ_1 and μ_2, are different from one another. Our training goal can thus be formulated as minimizing $\Pr\left(\mu_1 = \mu_2 \mid \boldsymbol{s}(\boldsymbol{z}), n_1, n_2\right)$, where $\boldsymbol{s}(\boldsymbol{z})$ denotes summary statistics of the labeled embedding points. This posterior is intractable, so instead we operate on the likelihood $\Pr\left(\boldsymbol{s}(\boldsymbol{z}) \mid \mu_1 = \mu_2, n_1, n_2\right)$ as a proxy.

We borrow a particular statistic from analysis of variance (ANOVA) hypothesis testing for equality of means. The statistic is a ratio of between-class variability to within-class variability:

$$s = \tilde{n} \frac{\sum_i n_i (\bar{z}_i - \bar{\bar{z}})^2}{\sum_{i,j} (z_{ij} - \bar{z}_i)^2}$$

where $\bar{z}_i = \langle z_{ij} \rangle$ and $\bar{\bar{z}} = \langle \bar{z}_i \rangle$ are expectations and $\tilde{n} = n_1 + n_2 - 2$. Under the null hypothesis $\mu_1 = \mu_2$ and an additional normality assumption, $z_{ij} \sim \mathcal{N}(\mu, \sigma^2)$, our statistic s is a draw from a Fisher-Snedecor (or F) distribution with degrees of freedom 1 and $\tilde{n}$, $S \sim F_{1,\tilde{n}}$. Large s indicate that embeddings from the two different classes are well separated relative to two embeddings from the same class, which is unlikely under $F_{1,\tilde{n}}$. Thus, the CDF of the F distribution offers a measure of the separation between classes:

$$\Pr\left(S < s \middle| \, \mu_1 = \mu_2, \tilde{n}\right) = I\left(\frac{s}{s + \tilde{n}}, \frac{1}{2}, \frac{\tilde{n}}{2}\right) \tag{1}$$

where I is the regularized incomplete beta function, which is differentiable and thus can be incorporated into an objective function for gradient-based training.

Several comments on this approach. First, although it assumes the two classes have equal variance, the likelihood in Equation 1 is fairly robust against inequality of the variances as long as $n_1 \approx n_2$. Second, the F statistic can be computed for an arbitrary number of classes; the generalization of the likelihood in Equation 1 is conditioned on *all* class instances being drawn from the same distribution. Because this likelihood is a very weak indicator of class separation, we restrict our use of the F statistic to class pairs. Third, this approach is based entirely on *statistics* of the training set, whereas every other deep-embedding method of which we are aware uses training criteria that are based on individual instances. For example, the triplet loss [7] attempts to ensure that for specific triplets $\{z_{11}, z_{12}, z_{21}\}$, z_{11} is closer to z_{12} than to z_{21}. Objectives based on specific instances will be more susceptible to noise in the data set and may be more prone to overfitting.

1.1 From one to many dimensions

Our example in the previous section assumed one-dimensional embeddings. We have explored two extensions of the approach to many-dimensional embeddings. First, if we assume that the Euclidean distances between embedded points are gamma distributed—which turns out to be a good empirical approximation at any stage of training—then we can represent the numerator and denominator in the F statistic as sums of gamma random variables, and a variant of the unidimensional separation measure (Equation 1) can be used to assess separation based on Euclidean distances. Second, we can apply the unidimensional separation measure for multiple dimensions of the many-dimensional embedding space. We focus on the latter approach in this article.

For a given class pair (α, β), we can compute $\Pr\left(S < s \middle| \mu_{1k} = \mu_{2k}\right)$ for each dimension k of the embedding space We select a set, $\boldsymbol{D}_{\alpha,\beta}$, of the d dimensions with largest $\Pr\left(S < s \middle| \mu_{1k} = \mu_{2k}\right)$ i.e., the dimensions that are best separated already. Although it is important to separate classes, they needn't be separated on *all* dimensions because the pair may have semantic similarity or equivalence along some dimensions. The pair is separated if they can be distinguished reliably on a subset of dimensions.

For a training set or a mini-batch with multiple instances of a set of classes $\boldsymbol{C}$, our embedding objective is to maximize the joint probability of separation for all class pairs (α, β) on all relevant dimensions, $\boldsymbol{D}_{\alpha,\beta}$. Framed as a loss, we minimize the log probability:

$$\mathcal{L}_F = -\sum_{\{\alpha,\beta\}\in\boldsymbol{C}} \sum_{k\in\boldsymbol{D}_{\alpha,\beta}} \ln \Pr\left(S < s \middle| \mu_{1k} = \mu_{2k}\right)$$

This F-*statistic loss* has four desirable properties. First, the gradient rapidly drops to zero once classes become reliably separated on at least d dimensions, leading to a natural stopping criterion; the degree of separation obtained is related to the number of samples per class. Second, in contrast to other losses, the F-statistic loss is not invariant to rotations in the embedding space; this focus on separating along specific dimensions tends to yield disentangled features when the class structure is factorial or compositional. Third, embeddings obtained are relatively insensitive to the one free parameter, d. Fourth, because the loss is expressed in the currency of probability it can readily be combined with additional losses expressed similarly (e.g., a reconstruction loss framed as a likelihood). The following sections demonstrate the advantages of the F-statistic loss for identity classification and for disentangling attributes related to identity.

2 Identity Classification

In this section, we demonstrate the advantages of the F-statistic loss over state-of-the-art methods on identity classification. The task involves matching a person from a wide-angle, full-body photograph, taken at various angles and poses. We evaluate two data sets—CUHK03 [9] and Market-1501 [10]—following the methodology of [1]. Five-fold cross validation is performed for CUHK03, and a single train/test split used for Market-1501.

Training Details. Following [1], we use the Deep Metric Learning [3] architecture with a 500-dimensional embedding. All nets were trained using the ADAM [11] optimizer, with a learning rate of 10^{-4}. A validation set was withheld from the training set, and used for early stopping. To construct a mini-batch for training, we randomly select 12 identities, with up to 10 samples of each identity, as in [1]. In addition to the F-statistic loss, we evaluated histogram [1], triplet [2], and binomial deviance [12] losses. For the triplet loss, we use all triplets in the minibatch. For the histogram loss and binomial deviance losses, we use all pairs. For the F-loss, we use all class pairs. The triplet loss is trained and evaluated using L_2 distances. The F-statistic loss is evaluated using L_2 distances. As in [1], embeddings obtained discovered by the histogram and binomial-deviance losses are constrained to lie on the unit hypersphere; cosine distance is used for training and evaluation. For the F-statistic loss, we determined the best value of d, the number of dimensions to separate, using the validation set of the first split. Performance is relatively insensitive to d for $2 < d < 100$.

Results. Figure 2 reports Recall@k accuracy, the performance metric used in earlier work. For each query image in the test set, we compute the distance of its embedding vector to the embedding vectors in the remainder of the test set. A query returns a 1 if one of the k nearest neighbors in the embedding space is of the same class as the query image, 0 otherwise. Recall@k is the percentage

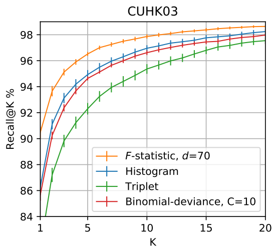

Figure 2: Recall@k for the CUHK03 dataset and four deep-embedding losses. For CUHK03, each line indicates the mean Recall@k over cross-validation splits, and vertical bars indicate ±1 standard error of the mean.

of queries that return a 1. The F-statistic loss leads to reliably better accuracy, especially for low k. On CUHK03, the F-statistic loss obtains a Recall@1 accuracy of $90.5\% \pm 0.4\%$, compared to next best, histogram loss, $86.3\% \pm 0.6\%$. On Market-1501, the single train-test split yields comparable performance for all losses for small k, e.g., Recall@1 accuracy is 66.51% for the histogram loss, 65.87% for the triplet loss, 65.75% for the F-statistic loss, and 65.45% for binomial deviance loss.

3 Disentangling Identity Attributes

Next, we show that the F-statistic loss obtains disentangled embeddings—embeddings whose dimensions are aligned with the categorical and semantic features of the input data. We explore disentangling with a data set of video game sprites—60×60 pixel color images of game characters viewed from various angles and in a variety of poses [13]. The identity of the game characters is composed of 7 attributes—body, arms, hair, gender, armor, greaves, and weapon—each with 2–5 distinct values, leading to 672 total unique identities which can be instantiated in various viewing angles and poses.

We used the encoder architecture of Reed et al. [13] as well as their embedding dimensionality of 22. We evaluated with five-fold cross validation, splitting by identity and including all variations in viewing angle and pose. A portion of the training set was reserved to determine when to stop training based on Recall@1 performance. For these experiments, we compare the F-statistic loss to the triplet loss; other losses using L_p norm distances should yield similar results.

The sprite dataset is factorial: every combination of attribute-values is present. An ideal disentangled representation will also be factorial, wherein all pairs of dimensions are statistically independent. However, due to the fact that the embedding dimensionality may allow for redundancy, simply measuring mutual information will not reveal disentangled structure: if one embedding is more compact than another, it will allow for more redundancy and consequently higher mutual information. As an alternative to mutual information, we measure how well each embedding dimension predicts each identity-attribute value (e.g., hair=blond, weapon=spear); in a disentangled representation, single dimensions should be highly predictive of these values. For each value, we assess how well each embedding dimension discriminates the given value from other values of the attribute, and record the AUC of the most predictive embedding dimension. There are in total 17 (nonredundant) attribute values, and five cross-validation splits, so we record 85 AUCs for each training loss. AUC is based on the entire dataset to ensure adequate coverage over all attribute values. Figure 3 shows the distribution of AUCs for embeddings based on the triplet, histogram, and F-statistic losses. The embeddings trained using the F-statistic loss are more likely to include dimensions that are aligned with the generative attributes of the domain (i.e., AUC close to 1). This property is robust for moderate values of d.

4 Discussion and Future Work

The F-statistic loss is a novel approach to learning deep embeddings that uses only summary statistics to judge embedding quality, in contrast to approaches that examine the relationships among the individual embedding points. Our approach beats state-of-the-art performance on the "person re-identification" task. Our approach also yields better disentangling of factors that compose identity, leading to more interpretable representations.

We are presently investigating the use of this loss for disentangling content and style (or, identity and non-identity) by incorporating an additional reconstruction loss to ensure that the combined content+style representation preserves information in the input [14, 15]. We further expect to improve

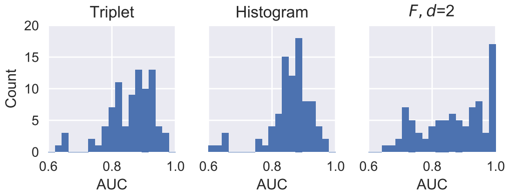

Figure 3: Predicting semantic attributes from single embedding dimensions. AUC distributions for triplet, histogram, and $d = 2$ for the F-statistic loss.

the disentangling of content and style by inverting the F-statistic loss for the style component of the embedding to reduce class separation. Finally, we are evaluating the content-style decompositions obtained with the F-statistic loss to those obtained by other losses, in an effort to demonstrate that the F-statistic decompositions are superior for image synthesis and for generating augmented data sets.

References

[1] Evgeniya Ustinova and Victor Lempitsky. Learning deep embeddings with histogram loss. *Advances in Neural Information Processing Systems*, pages 4170–4178, 2016.

[2] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[3] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Deep metric learning for person re-identification. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 34–39. IEEE, 2014.

[4] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In U. V. Luxburg, I. Guyon, S. Bengio, H. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages xxx–xxx. Curran Associates, Inc., 2017.

[5] E. Triantafillou, R. Zemel, and R. Urtasan. Few-shot learning through an information retrieval lens. In U. V. Luxburg, I. Guyon, S. Bengio, H. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages xxx–xxx. Curran Associates, Inc., 2017.

[6] J B Tenenbaum and W T Freeman. Separating style and content with bilinear models. *Neural Computation*, 12(6):1247–1283, 2000.

[7] S Chopra, R Hadsell, and LeCun Y. Learning a similiarty metric discriminatively, with application to face verification. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 349–356, 2005.

[8] Hyun Oh Song, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding query retrieval. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4004–4012, 2016.

[9] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014.

[10] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1116–1124, 2015.

[11] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[12] Dong Yi, Zhen Lei, and Stan Z. Li. Deep metric learning for practical person re-identification. *ICPR*, 11(4):1–11, 2014.

[13] Scott E. Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. *Advances in Neural Information Processing Systems*, pages 1252–1260, 2015.

[14] Salah Rifai, Yoshua Bengio, Aaron Courville, Pascal Vincent, and Mehdi Mirza. Disentangling factors of variation for facial expression recognition. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7577 LNCS(PART 6):808–822, 2012.

[15] Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *Advances in Neural Information Processing Systems*, pages 2539–2547, 2015.