Investigating Company Logo Memorability with Convolutional Neural Embedding Models

Eoghan Keany egnkeany@gmail.com and
James McDermott james.mcdermott@nuigalway.ie

School of Computer Science, National University of Ireland, Galway

Abstract. The present study compared several state of the art neural embedding models for the correlation of their embeddings with human judgements, relating to both human memory and relevance ranking. These models included two embedding models, DeepRank and Ranknet; two classification models, ConvNet and VisNet; and a Variational Autoencoder. To assess each model's performance, two custom evaluation metrics were developed: a fine detail coefficient and a coarse detail coefficient. These measures revealed that the embeddings produced by the DeepRank model had the highest correlation with human judgement. This design combination of a tri-linear architecture, triplet loss function and semi-hard negative sampling did best at capturing the similarities between the images, achieving the highest overall result for both the fine detail and coarse detail coefficients. The embeddings produced by the DeepRank model were then used to investigate the memorability of each company logo. However, as image memorability cannot be characterised by low-level features alone our results suffered. In addition, the results show that deep features extracted from the embedding models show markedly better results on fine classification and retrieval tasks than their classification counterparts.

Keywords: Convolutional neural network · embedding · image dissimilarity · branding · logo

1 Introduction

Think clothes, TVs, computers, food, cars. From the moment we wake to the moment we sleep, we are constantly being bombarded with logos. Logos represent an interesting form of visual information, as they are specifically designed to be relatively simple, recognisable and memorable all in an attempt to improve brand recognition. But just how accurately can we remember these famous symbols?. In pursuit of an answer, the company Signs.com decided to carry out a study, "Branded in Memory", in which consumers were asked to draw several well-known company logos from memory[1]. Each individual image was given an accuracy score by a group of marketing experts as part of the study, creating a rich resource for computer vision and machine learning algorithms. A subset is shown in Fig. 1.

[1] https://www.signs.com/branded-in-memory

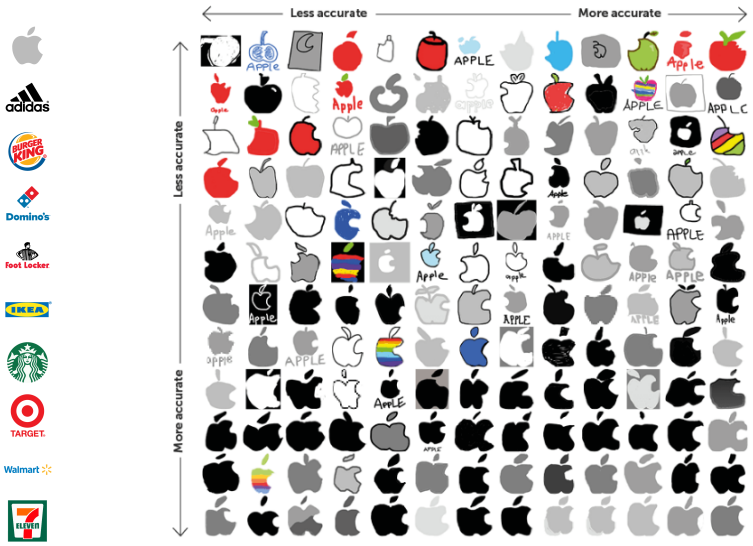


Fig. 1: A sample of the data set. Left: original logos. Right: logos for one company as drawn by subjects, arranged from left to right and top to bottom in increasing similarity to the true logo. From Signs.com.

Previous literature has shown that humans have a very strong visual memory. Each individual memory is stored and protected from interference, even when hundreds of images intervene between the first and second appearance of an image [14]. Research has also indicated an immense capacity for visual detail in long-term memory [3]. However the Signs.com data set apparently contradicts these statements as only a handful of the drawings are highly accurate. Despite a strong body of research showing that multiple exposures to stimuli can result in relatively accurate memory, other studies have demonstrated that exposure does not necessarily lead to enhanced memory but may contribute to more general, gist based memory. While psychologists have studied human capacity to remember visual stimuli little work has been conducted on the differences in stimuli that make them more or less memorable. While image memorability is partly subjective, some images are intrinsically more memorable than others, independent of context and subjects' biases [10]. This paper describes an attempt to automatically recreate some findings from the Signs.com study using several computer vision models.

Many computer vision models involve embeddings, that is lower-dimensional spaces to which the original data space (e.g. the space of images) is mapped. Embeddings are useful if topological properties of the embedding such as distances are well-aligned with human perception. However, in many contexts it is difficult for humans to give assessments of image dissimilarity since it is entangled with semantic and cultural factors: is an image of a dalmatian similar to an image of a zebra, or not? In this context, the Signs.com dataset is particularly inter-

esting because it is in a highly constrained domain and it includes approximate dissimilarity labels. It allows us to investigate whether the embeddings created by common computer vision models have the desirable property of correlation between embedding distance and human perception of image dissimilarity.

Our study thus directly compares several state of the art neural models for the correlation of distances in their embeddings with human judgements. This also allows insight into logo memorability. The study demonstrates that embeddings created by explicit embedding models show markedly better results on fine classification and retrieval tasks than their classification counterparts.

2 Related work

Traditionally, feature extraction was accomplished by designing hand crafted features with the aid of an expert. However, in recent years many known typical image descriptors like SIFT, HOG and local binary patterns [13], [5], [1] have been replaced by state of the art image CNN's which learn the set of features directly from the observations themselves by undergoing supervised training. With regards to training a multi-class classification model, softmax cross-entropy loss is still the most popular choice [12]. Although this loss function has been successfully applied across numerous domains, this metric inherently cannot learn from the between-class relationships which can be very informative and will become necessary in our study as discussed in section 4 [9].

In order to capture between-class relationships, several embedding models were introduced. These models explicitly learn a mapping to a new feature space by varying the position of each sample point in this new space relative to another point. For example, triplet loss operates by minimising the distance between a sample image and a positive anchor whilst maximising the distance between the sample and a negative anchor.

$$l(p_i, p_i^+, p_i^-) = max\{0, alpha + D(f(p_i), f(p_i^+)) - D(f(p_i), f(p_i^-))\}$$

Despite embedding model's ability to capture the between-class relationships, they do have some inherent drawbacks. The models converge at a very slow and unsteady rate during training and they also require complicated sampling operations. In the literature it has been common to select from all possible pairs at random for contrastive loss [2]. On its own, random sampling of triplets may mostly yield "easy" examples that induce no loss [20]. Hard negative mining has been shown to contribute to faster convergence [17]. In contrast, hard negative mining paired with triplet loss can lead to a collapsed model (where every image has the same embedding). Thus in response semi-hard negative mining was created, first used in FaceNet. It is widely accepted as the standard sampling approach for triplet loss [15]. In this study a random approach was chosen for the contrastive loss implementation and a semi-hard method for the triplet loss function. As a result of this complexity, these Siamese architectures are much more difficult to optimise in comparison to their cross-entropy counterparts [20].

3 Models

In total six models were created and tested, including the raw pixel value representation as a baseline. Each of the 5 neural models were trained on the entire data set which contained 10 classes/brands and 1440 images in total. Of these 5 neural models, four were trained under supervised conditions with the exception of the VAE. Both DeepRank and RankNet are embedding models that explicitly learn the embeddings directly. Whereas, both VisNet and ConvNet are classification models that implicitly learn the embeddings, thus the penultimate layer was used as the embedded representation in these cases. The differences in model architectures and training regimes are described in the following sub sections.

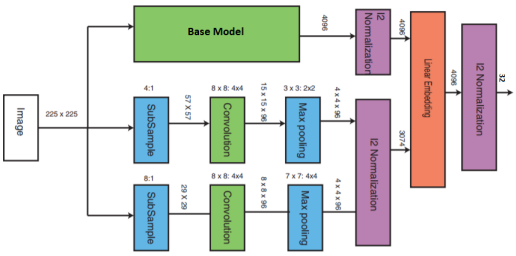
3.1 The Base Model

To mitigate experimental bias each of the neural models were constructed around a shared base deep residual convolutional neural network inspired by the ILSVRC 2015 winner Resnet [18]. The base model was constructed using skip gram connections to increase the training efficiency of the model as it contained just over 50 layers in total. The model takes a [50x50] RGB image as an input and outputs a 32 dimensional L2 normalised embedded representation in the form of a dense vector. The base model is comprised of 6 residual blocks in conjunction with other standard layers such as convolution, pooling, normalisation and activation layers. There are two types of residual blocks implemented in this model, the identity block and the residual block. The identity block contains the basic skip gram connection and is used when the input and output are of the same dimension, whereas the residual block is used when the dimensions differ. Identity mapping still takes place, however an adjustment to the resolution and channels of the alternative pathway occurs by means of a 1x1 convolution operation before recombination. In both blocks the main pathway is subjected to a sequence of 3x3 convolution, batch normalisation, max pooling with a 2x2 kernel and RELU activation layers. Each residual block is three layers deep meaning that the main path goes through nine successive layers following the sequence described above before being combined with the skip connection. The output feature map from the successive residual blocks are then reduced using average pooling and flattened into a 32 dimensional vector. Each model is built on this base model by applying different loss functions and additional architectures.

ConvNet The first model implemented in this study was ConvNet [18]. This network follows a classic structure that uses the categorical cross entropy (negative log likelihood) loss function in conjunction with the base model. This loss function simply measures the dissimilarity between the true and predicted probability distributions obtained from the final softmax activated layer. Once training was complete, the 32 dimensional dense vector located before the final softmax activated layer was extracted and used as an embedded representation.

VisNet Visnet expands on ConvNet by introducing a tri-linear parallel CNN structure by combining the base model with two smaller shallow networks [6]. This has the added benefit of using the base model to encode strong invariance that can capture image semantics, while the other two parts of the network take down-sampled representations that have less invariance and capture more of the input's visual appearance. Both of the smaller networks contain an [8x8] convolutional layer but differ in stride length and max pooling filter size. All three outputs from the base model and two shallower architectures are L2 normalised before being concatenated together to produce a 32 dimensional embedded representation. Similar to ConvNet, during training a 10 dimensional softmax layer was used with the categorical cross-entropy loss.

Fig. 2: Visnet Tri-linear Network Architecture [19]



DeepRank DeepRank is an embedding model where the network can be thought of as a function that simply maps an input to a point in Euclidean space. Unlike the original implementation [19], which evaluates the hinge loss of a triplet, this study utilised the triplet loss function to learn the appropriate embeddings [4]. The model takes an image triplet as an input. Each image is then fed independently to three identical deep neural networks with a shared architecture and parameters. This architecture follows the tri-linear CNN structure seen above to capture both the image semantics and visual appearance. The triplet loss function operates by minimising the distance between a sample image (the "anchor') and a same-class image whilst maximising the distance between the anchor and a different-class image.

$$l(p_i, p_i^+, p_i^-) = \max\{0, \alpha + D(f(p_i), f(p_i^+)) - D(f(p_i), f(p_i^-))\}$$

The margin α defines a minimum threshold between the positive and negative images. This encourages the positive samples to maintain a minimum distance between each other. However, this parameter has an optimal balance point, as if the margin is increased the number of hard negatives or good training samples falls. Thus, for this implementation an alpha value of 0.2 was heuristically

chosen [19]. An offline approach for choosing semi-hard triplets was also implemented to improve convergence time [15]. Triplets are chosen by choosing a random image as anchor and then negative and positive images based on the constraint:

$$|f(x_i^a) - f(x_i^p)|^2 + alpha < |f(x_i^a) - f(x_i^n)|^2$$

This ensures that the no "easy" samples are produced which give zero loss and hence do not improve weights.

RankNet As with DeepRank, RankNet uses a Siamese architecture and also seeks to learn an embedding directly. However, RankNet uses the contrastive loss function. As with any other distance-based loss function, it aims to produce an embedding that captures the semantic similarity between images. This function can be expressed mathematically as [7]:

$$L(\theta) = \frac{1-Y}{2} D(X_q, X_p)^2 + \frac{Y}{2}(max(0, m - D(X_q, X_n)^2))$$

When a similar image pair (label Y = 0) is fed to the network, the first part becomes 0 and the loss becomes equal to the positive pair distance between two similar images. Gradient descent will push them closer together. On the other hand, when two dissimilar images (label Y = 1) are fed to the network, the second part of the equation disappears and the remainder works as a hinge loss. This allows the function to directly optimise the distance between samples by encouraging all positive pair distances to approach 0, whilst keeping negative pair distances above a certain threshold m. However, one defect of contrastive loss is that a constant margin m has to be applied for all negative pairs. This causes visually diverse classes to be embedded in the same small space as visually similar ones. In contrast, triplet loss tries to keep all positive points closer than any negative points for each image. This allows the embedding space to be distorted and does not enforce a constant margin [16], [20].

Variational Autoencoder The variational auto encoder is the only unsupervised model used. Similar to the standard autoencoder, the VAE encodes the input image to a reduced latent space [11]. The network has an encoder-decoder architecture where the encoder produces a latent space representation and the decoder reconstructs the original image from a sampled point in the latent space. The VAE constrains the encoder network to create latent vectors that follow a Gaussian distribution. The network accomplishes this by producing both a mean and standard deviation for each latent variable. To extract the image embedding the mean latent vector from the encoder was chosen as opposed to the sampled latent vector. Both values were tested with the mean latent vector giving marginally better results.

Raw Pixels Finally, we take the raw pixel space (50x50x3) as a baseline.

4 Experiments and Results

This section compares the embeddings created by the models. Three evaluation metrics were used.

4.1 Measuring Model Success and Logo Memorability

We firstly discuss the measures of success used to assess the performance of the embeddings and the memorability of each company's logo. Firstly, the quality of the embeddings produced by the different architectures were evaluated using two measures of success, the coarse detail coefficient and the fine detail coefficient, described below. These metrics provided a means to identify and select an appropriate model that best suited our needs. This model was then used to investigate which company logo is the most memorable, using the Memorability Coefficient. All three of these metrics rely on a distance metric to quantify the separation between two images in the latent space. As the embeddings are L2 batch normalised this allows the squared Euclidean distance between these normalised vectors to be proportional to their cosine similarity.

Coarse Detail Coefficient An accurate embedding should minimise within-class distances whilst maximising between-class distances (where each company is considered a class). We can measure this using the nearest neighbour classification accuracy. As we do not require a model capable of making predictions on new data, we do use any evaluation on unseen data.

Fine Detail Coefficient Between-class clustering alone can only capture the coarse details of an image. To measure each model's success at producing an embedding in which distances correlate to human judgement, a further measure is proposed. The data set contains 10 x 144 hand-drawn imitations of ten different logos, each labelled with a measure of similarity to the original logo provided by marketing experts in the original study. A model-derived ranking was then created by sorting the Euclidean distances between the embedded vectors of the actual logo design and each hand-drawn imitation logo. This new ranking vector can then be compared to the original ranking sequence using Kendall's Tau Correlation.

Memorability Coefficient To estimate the visual memorability of a company's logo design, the total Euclidean distance between the actual brand logo and every hand-drawn imitation was calculated. This measure is predicated on the logic that each hand-drawn image is an attempt to recreate the true logo. The accuracy of each image can therefore be estimated by calculating its distance to the actual logo design in the embedding. It is an assumption for this that distance in the embedding is correlated with perceptual dissimilarity, an assumption partly supported by results of the coarse and fine detail coefficients.

Furthermore, calculating the total distance to every hand drawn image gives an indication on the total accuracy of the drawings. As these drawings are recreated from memory a smaller distance indicates a more memorable and more re-creatable design. However, this measurement takes no account of differing exposure of each subject to the brands.

4.2 Results: Quality of Embeddings

Embeddings are visualised in Figure 3. The clusters formed by RankNet and Visnet are visually much tighter, but the between-class positioning of the DeepRank embedding leads to superior overall performance when measured numerically. Each model was individually trained ten times on the entire data set and the average of each measure of success was calculated.

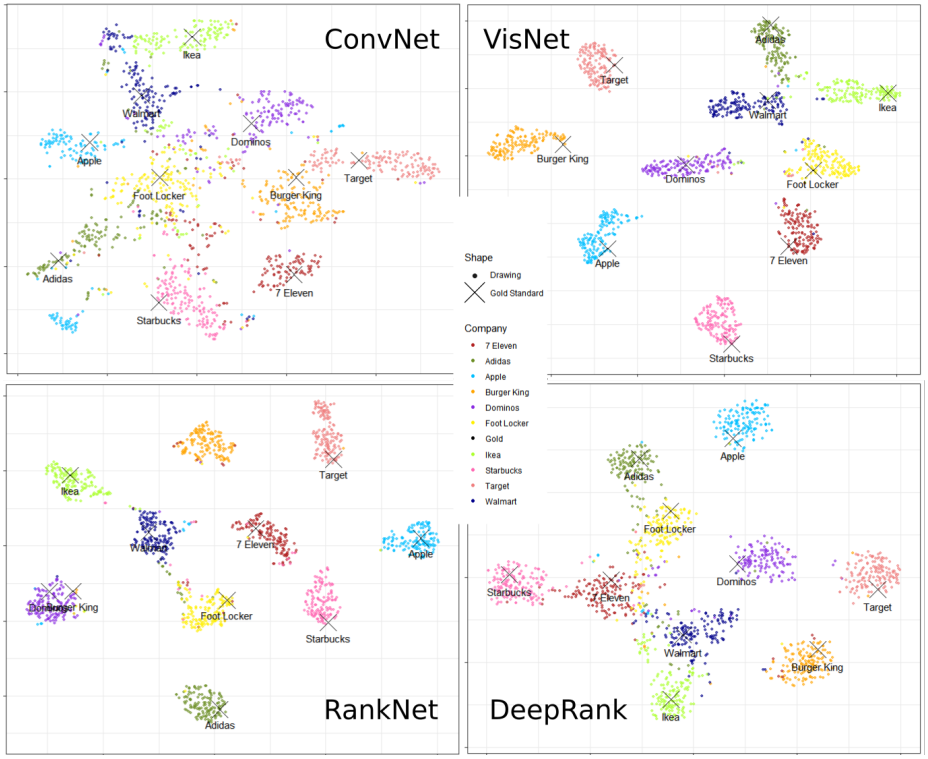


Fig. 3: Embeddings visualised with t-SNE. '$\times$' designates each actual logo image.

Numerical results are shown in Table 1. As expected the Raw Pixel representation performs poorly. However, it does retain some information as measured by the fine detail coefficient. This may be mostly due to correct colour-matching.

Table 1: Performance of each representation. DeepRank and RankNet consistently outperform other models, with DeepRank achieving the highest results in the Fine Detail Coefficient, while RankNet scored the highest in the Coarse Detail Coefficient.

Model Name	Fine Detail Coefficient	Coarse Detail Coefficient
Raw Pixels	0.14	0.5
VAE	0.15	0.56
ConvNet	0.2	0.85
Visnet	0.23	0.91
RankNet	0.28	0.94
DeepRank	0.35	0.92

The VAE attained the lowest model score. This poor result could be due to its unsupervised nature which causes the dominant information in the latent layer to be influenced solely on what contributes the most towards the loss function applied in the reconstruction, meaning that each dimension in the latent space can become entangled. Supervised methods can instead encourage the embedding to favour information about a specific feature of interest (cluster identity, etc.) between each logo. We believe this effect is exacerbated by the within-class disorder of various different logo iterations coupled with the poor quality drawings causing confusion and entanglement in the embedding space. To test this hypothesis it would be interesting to compare a Beta VAE which can mitigate the effect of entanglement in the latent space. It is also interesting to see that the performance of each model was correlated with the complexity and resources needed to train them, with the exception of the VAE. The results revealed that DeepRank was the best overall model. Achieving the highest combined score with an average fine detail score of 0.35 and an average coarse detail value of 0.92. In general the low scores in the fine detail could be affected by the challenging nature of the data set as learning fine image similarities is a challenging task in itself, as it needs to capture both the between-class and within-class image differences. In contrast, the coarse detail coefficient displayed a more optimistic result. Each model except for the VAE displayed a high propensity to produce useful embeddings that could capture the between-class differences. The RankNet model produced the most dense clustering and obtained the highest average coarse score of 0.94 making it the best model for a classification task. However, both Visnet and DeepRank achieved similar average performances with scores of 0.91 and 0.92 respectively.

The addition of multi-linear layers in the model architecture were hypothesised to have less within class variance and simply capture more of the input's visual appearance or low level features. This statement was supported by the increased performance in VisNet's ability to capture the coarse details in the images, achieving a coarse detail score of 0.91 compared to ConvNet's 0.85. Within this experiment there was also evidence to support the separation in performance

between the embedding models and their classification counterparts. With the embedding models performing better in both aspects of the evaluation including both fine and coarse-grained classification. However, this separation may not be reliable as it depends upon pair selection in the training process. In this experiment, a sampling process that could create a random semi-hard negative for every pair of anchor and positive [15] was implemented. However both the embedding models could benefit from an even more accurate sampling approach utilised in [19]. This method uses a pairwise relevance score for within class images, where the probability of an image being chosen as a query image is proportional to its relevance score. Applying this sampling technique would encode the fine details of the within-class images into the embeddings, where similar images within the same class will be embedded closer to one another. Nevertheless, this improvement would degrade the equality and efficacy of testing between the two model groups. These findings are not alone as previous literature has shown that the performance of the classification based features are heavily dependent upon the size of training set. When the size of the data set is small or the number of classes is very large, the embedding models will outperform classification models [8].

4.3 Results: Memorability of Logos

Table 2: The ranked memorability of each company logo where the actual rank refers to the annotated data set and the predictions were made with the DeepRank embeddings.

Brand Name	Actual Ranking	Predicted Ranking
Ikea	1	1
Target	2	2
Apple	3	3
7 Eleven	4	8
Burger King	5	4
Dominos	6	10
Adidas	7	6
Walmart	8	5
Foot Locker	9	9
Starbucks	10	7

In the final set of results, we compare logo designs for memorability as opposed to comparing neural models. Table 2 shows the results of the DeepRank Memorability coefficient for each company logo and compares it to expert judgements made as part of the Signs.com study. Simple and effective designs such as the Ikea, Apple and Target logos were consistently ranked among the more memorable designs by DeepRank (and other models), in accordance with the expert judgements. However, the models penalised companies who had multiple logo design iterations throughout their history such as Starbucks, Dominos and Adidas. Also designs based around text such as 7-eleven and Walmart were predicted by models to be more memorable. Both of these conflicted with the experts' opinion.

We also measure the correlation between the experts' and models' rankings of memorability. Despite memorability being an intrinsic property of an image, it cannot be characterised by common low-level image features. Indicatively, this makes it a difficult task for computer vision and the results here demonstrate this: the correlation was low, with a value of 0.11. This poor result can be partly attributed to the evolution and re-branding strategies of the companies themselves, and the use of text which is not processed as such by the models. Another source of error within the memorability coefficient could be explained by the poor image quality of the actual brand logos. In contrast to the training images these images had to be expanded from a dimension of [32x32] to [50x50] in order to run them through the networks. Also, it could be argued that some of the annotated scoring is unrealistic especially in its depiction of intricate logos such as Foot Locker and Starbucks, which are hindered by the drawing program used and the subject's drawing ability.

5 Conclusion

In this paper a novel method was presented to automatically recreate the findings from the study "Branded in Memory by Signs.com". The embeddings produced by the DeepRank model had the highest correlation with human judgement. Its combination of a tri-linear architecture, triplet loss function and semi-hard negative sampling allowed the model to capture the similarities between the logos. It achieved the highest result for both the fine detail and memorability coefficients with values of 0.35 and 0.25 respectively. Despite this success it would be naive to expect to replicate the complexities of human memory with a single model and the results for each individual model reflect this, with an average score of 0.11 for the memorability coefficient. The results also suggest that, overall, the embedding models performed better than their classification counterparts. However, it is vital not to over-interpret these results. Instead this study should be seen as motivation to conduct a further comparison, by highlighting the gaps in our algorithmic understanding of logo dissimilarity and memorability.

References

1. Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (12):2037–2041, 2006.
2. Sean Bell and Kavita Bala. Learning visual similarity for product design with convolutional neural networks. *ACM Transactions on Graphics*, 34(4):98, 2015.
3. Timothy F. Brady, Talia Konkle, George A. Alvarez, and Aude Oliva. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, 105(38):14325–14329, 2008.
4. Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. Large scale online learning of image similarity through ranking. *JMLR*, 11:1109–1135, Mar 2010.

5. Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *International Conference on Computer Vision Pattern Recognition*, pages 886–893, 2005.
6. Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1915–1929, 2012.
7. Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1735–1742. IEEE, 2006.
8. Shota Horiguchi, Daiki Ikami, and Kiyoharu Aizawa. Significance of softmax-based features in comparison to distance metric learning-based features. *arxiv:1712.10151*, 2019.
9. Le Hou, Chen-Ping Yu, and Dimitris Samaras. Squared earth mover's distance-based loss for training deep neural networks. *arXiv:1611.05916*, 2016.
10. Phillip Isola, Devi Parikh, Antonio Torralba, and Aude Oliva. Understanding the intrinsic memorability of images. In *Advances in Neural Information Processing Systems*, pages 2429–2437, 2011.
11. Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *arXiv:1312.6114*.
12. Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
13. David G Lowe et al. Object recognition from local scale-invariant features. In *iccv*, volume 99, pages 1150–1157, 1999.
14. Raymond S. Nickerson. Short-term memory for complex meaningful visual configurations: A demonstration of capacity. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, 19(2):155, 1965.
15. 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.
16. Vishvakarma A. Sharma R. Retrieving similar e-commerce images using deep learning. *arXiv:1901.03546*, 2019.
17. Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, Pascal Fua, and Francesc Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 118–126, 2015.
18. Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
19. Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1393, 2014.
20. Chao-Yuan Wu, R. Manmatha, Alexander J. Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2840–2848, 2017.