# The Effects of Unimodal Representation Choices on Multimodal Learning

**Fernando Tadao Ito, Helena de Medeiros Caseli, Jander Moreira**
Universidade Federal de São Carlos
Rod. Washington Luís Km 235, Postal Box 676, CEP: 13565-905, São Carlos-SP
f.tadaoito@gmail.com, helenacaseli@dc.ufscar.br, jander@dc.ufscar.br

**Abstract**

Multimodal representations are distributed vectors that map multiple modes of information to a single mathematical space, where distances between instances delineate their similarity. In most cases, using a single unimodal representation technique is sufficient for each mode in the creation of multimodal spaces. In this paper, we investigate how different unimodal representations can be combined, and argue that the way they are combined can affect the performance, representation accuracy and classification metrics of other multimodal methods. In the experiments present in this paper, we used a dataset composed of images and text descriptions of products that have been extracted from an e-commerce site in Brazil. From this dataset, we tested our hypothesis in common classification problems to evaluate how multimodal representations can differ according to their component unimodal representation methods. For this domain, we selected eight methods of unimodal representation: LSI, LDA, Word2Vec, GloVe for text; SIFT, SURF, ORB and VGG19 for images. Multimodal representations were built by a multimodal deep autoencoder and a bidirectional deep neural network.

**Keywords:** distributed representations, multimodal representation, multimodal autoencoders, classification

## 1. Introduction

In the real world, multiple modes of information are gathered to create knowledge in a way humans can understand. From structured language, we can express abstract ideas in a standardized fashion; while visualization can help us to detect and observe objects and intention. However, in mathematical terms, each kind of data has different intrinsic statistical features that cannot be compared in a trivial way.

To analyze these features, we use distributed representation models to map real-world unimodal data to mathematical vectors with reduced dimensionality (Van Gelder, 1992). This idea is based on the distributional hypothesis (which states that "items with similar distributions have similar meanings") to discover non-trivial patterns and relationships between data instances (Harris, 1954).

The challenge of melding different data domains into a single conceptual space is the goal of **multimodal representations**. A multimodal representation model is a function that maps objects that have multiple representations to a single vectorial space, combining the semantic implications that are expressed by distance similarities in unimodal representations (Atrey et al., 2010). Using a multimodal model, we can correlate words in a textual description of an object to the visual features of an image representing it, learning its "concept". For example, a multimodal model could take as input the image of a sofa and return words that have similar meaning such as "chair" and "armchair", or take one of these words as input and output a related image. The hypothesis of a multimodal model is that it can find features that are invariant to information modality, intrinsic to the concept, and can be subsequently used in any machine learning task (e.g. e-commerce product classification).

In this paper, we explore unimodal and multimodal representations and measure their performance in a classification task using real-world data. Our hypothesis is that multimodal representations are directly influenced by the underlying unimodal representations used in its creation. Previous works on the area have already used multimodal techniques to increase accuracy in multimedia classification tasks (Ngiam et al., 2011). In our work, we want to expand these previous works by assessing new combinations of representations.

## 2. Related Works

Early multimodal methods focused on representation fusions, either by combining representations before classification (*feature level* fusion) or by combining the results of classifications performed in single-mode representations in another analysis (*decision level* fusion) (Atrey et al., 2010). Feature level fusion has been used in multimodal pedestrian tracking using multiple detection algorithms (Yang et al., 2005) and traffic surveillance with multiple video sources (Wang et al., 2003). Decision level fusion can be compared to ensemble learning (Dietterich, 2000), combining different algorithms and data sources to create a more accurate representation.

One of the first multimodal model-based representation approaches is the one described by Ngiam et al. (2011), applying Restricted Boltzmann Machines (RBM) (Salakhutdinov and Hinton, 2009) and autoencoders (Bourlard and Kamp, 1988) for video-audio joint representations. The unimodal representations are first used to pre-train the input layers of the multimodal autoencoder network via RBM unsupervised training. The initial weights are then used in the complete autoencoder network trained using single-modality data and multi-modality data. In that paper, the grayscale image is cropped in a specific region of interest (the mouth), rescaled into a $60 \times 80$ pixel matrix and then translated to a reduced dimensionality space (32 dimensions) by Principal Component Analysis (PCA) whitening (Friedman, 1987). The audio was represented by spectrogram signals with temporal derivatives, resulting in a 483 dimension vector further reduced (100 dimensions) by PCA whitening. Using digit images from the MNIST dataset (LeCun and Cortes, 2010) and noisy variants as a second mode, multimodal features performed better on an image classification task when compared to single-mode feature

representations.

Expanding on the multimodal approach introduced above, Andrew et al. (2013) use two different autoencoders and align the projections generated by maximizing the canonical correlation between the features. This model assumes that the projection of a vectorial space into the other is sufficient to portrait the features shared by two different modalities. This model was not tested in a classification task, only the correlation of the generated features was taken into account. The databases used (MNIST and XRMB (Westbury, 1994)) did not undergo any preprocessing or representation methods. This model creates correlated projections of data that share information between modalities without needing both modes to be calculated.

Wang et al. (2016) present an overview of multimodal techniques, including the ones described by Ngiam et al. (2011) and Andrew et al. (2013), expand on their objective functions and propose new architectures fusing these objective functions. In particular, the Deep Canonical Correlated Autoencoder (DCCAE) uses both objectives from previous works simultaneously to increase correlation between obtained features and fidelity to original data. This architecture is then tested and compared with others in unsupervised classification tasks in noisy MNIST and XRMB datasets: accuracy was measured by how the representations were clustered together, and multimodal representations had the best results.

Vukotić et al. (2016) present a new approach to multimodal fusion using paired crossmodal neural networks (BiDNN). Two neural networks are created, each mapping one modality to another directly, with the weights of hidden layers of both networks tied. The result of this training is a central layer in both networks that maps any given modality to a shared representation space between the two given modalities. This architecture is compared to others using metrics obtained from analyzing the MediaEval 2014 dataset (Eskevich et al., 2014) using audio transcripts and video segments to link a video to a specific concept (anchor), obtaining superior results in this classification task. The transcripts were represented with Word2Vec embeddings, and the video was transcribed to human visual concepts.

## 3. Methodology

In this section, we describe the representations used for comparison in our experiments. Four methods of each modality were chosen. For texts, we used: (1) Latent Semantic Indexing (Landauer et al., 1998) and (2) Latent Dirichlet Analysis (Blei et al., 2003) for topical representation, and (3) Word2Vec (Mikolov et al., 2013) and (4) Global Vectors for word representation (Pennington et al., 2014). For images, we applied: (1) Scale-Invariant Feature Transform (Lowe, 2004), (2) Speeded-Up Robust Features (Bay et al., 2008), (3) Oriented FAST and Rotated BRIEF (Rublee et al., 2011) for Bag-of-Visual-Words (Yang et al., 2007) vector generation, and (4) neural features obtained from the VGG19 pre-trained network (Simonyan and Zisserman, 2014).

### 3.1. Textual Representation

Textual distributed representation is based on the aforementioned distributional hypothesis (Harris, 1954) which states that similar statistical distributions denotes semantic similarity between two items. In the textual domain, this means that words that appear in the same context have similar semantic meaning.

In this paper, we review methods of textual representation from two ways: one by **topic representation**, obtaining vectors that encode the pertinence of words in sentences if they appear together in a *corpus*, and one by **word embeddings**, creating word representations based on their occurrence context and combining them in a single document vector. Thus, the methods for textual representation investigated in this paper are:

- Latent Semantic Indexing (LSI) (Landauer et al., 1998): This method uses the term-document matrix that encodes the frequency of each term by document and its eigenvalues and eigenvectors to decompose this data and find a new representation. This is accomplished by Singular Value Decomposition, selecting only the highest eigenvalues and their correspondent eigenvectors to recompose the term-document matrix, leading to a reduced topic-document matrix.

- Latent Dirichlet Analysis (LDA) (Blei et al., 2003): Similar to LSI, this method codifies a topical distribution of words using a term-document matrix. But, instead of matrix operations to simplify the original data, it uses probability and parameter estimations to find word-topic and topic-document distributions.

- Word2Vec (W2V) (Mikolov et al., 2013): This neural model uses a shallow neural network to create distributed representations based on the context of each word. This model can be trained in two ways: training the model to find the context of a given word (Skipgram) or find the central word of a given context (CBoW). Either way, the model codifies a representation of a word in its hidden layer. The simplest way of codifying a document vector from its words is to add all present vectors and divide them by the number of words in it, as the semantic information is kept on this vector combination. This document vector performs poorly in large texts, as it loses semantic information shared between contexts the same way as the standard Bag-of-Words representation does. Another way of codification of document vectors that circumvents the aforementioned limitation is Doc2Vec (Le and Mikolov, 2014), adding a document identification token to each context and then calculating its embedding.

- Global Vectors for Word Representation (GloVe) (Pennington et al., 2014): This statistical model uses a term co-ocurrence matrix and ratios of co-ocurrence to find word vectors with distances between words relative to their co-ocurrence ratios. In the example given by the proposing paper, the ratio of $\mathrm{P}(solid|ice)/\mathrm{P}(solid|steam)$ is much higher than 1,

meaning that "solid" is more semantically related to "ice" than "steam". The objective function of this model reflects this ratio in the distances between these word vectors.

### 3.2. Visual Representation

A digital image is represented by a matrix of pixels (tuples of numbers with intensities of particular color channels). Albeit great for visualization, this matrix often has highly correlated neighboring pixels, creating redundant data. To extract meaningful mathematical information from an image, we must first find regions of interest that uniquely define it, and map these to a vector space where they can be compared.

This paper uses two different ways to generate features from an image: **hand-crafted descriptors**, which are predefined ways to find regions of interest and encode them into vectors; and **neural features**, automatically extracted and selected by an ImageNet pre-trained neural network. Thus, the methods for visual representation investigated in this paper are:

- Scale-Invariant Feature Transform (SIFT) (Lowe, 2004): This method extracts features that are invariant to scale, illumination and rotation. It is composed of four main steps: (1) keypoint extraction, via Differences of Gaussians in different scales; (2) keypoint localization, to refine and filter extracted keypoints; (3) orientation assignment, for each keypoint to achieve rotation invariance; (4) keypoint description, using the histogram of gradients in the neighborhood of the keypoint to encode a 128-position vector.

- Speeded-Up Robust Features (SURF) (Bay et al., 2008): As an extension of the SIFT method, SURF follows the same steps of SIFT applying different mathematical methods. For keypoint extraction and localization, the Differences of Gaussians are replaced by Box Filters and Hessian Matrix determinants; for orientation assignment and keypoint description, SURF uses Haar Wavelet responses around the keypoint. SURF achieves similar results to SIFT, generating smaller vectors (64-positions) with improved speed.

- Oriented FAST and Rotated BRIEF (ORB) (Rublee et al., 2011): This method is a fusion of two descriptors: FAST (Rosten and Drummond, 2006) and BRIEF (Calonder et al., 2010). ORB uses the FAST keypoint extraction method, achieves rotation invariance by analyzing the weighted centroid of intensities around each keypoint, and then combines this orientation with the BRIEF descriptor by prior rotation of the pixels in the described keypoint neighborhood.

- VGG19 classes (Simonyan and Zisserman, 2014): This method uses the classes detected from the VGG19 model pre-trained using the images from ImageNet, creating a probability vector of 1000 positions, probabilities of specific objects in a scene.

### 3.3. Multimodal Representation

To combine multiple data representations in a simple way, we can use feature concatenation prior to classification (Wang et al., 2003), create ensembles of single-mode classifiers (Radová and Psutka, 1997), align features of single-mode representations by similarity measures (Frome et al., 2013) or co-learn single-modality representations using another modality as basis (Information Resources Management Association, 2012, Chapter 28). But these representations do not encode a real fusion between two modalities in a **single mathematical vector space**, as they only add information to an existing space or combine inferences from separate spaces in a shallow way.

In this paper, we use both early-stage feature fusion with concatenation (Wang et al., 2003) and model-based feature fusion, creating a single vector space that maps both modalities to a single vectorial space. In this paper, we will use a simple multimodal autoencoder framework with single-mode pre-training, in a similar architecture to the one proposed in Ngiam et al. (2011):

1. Two deep autoencoders are pre-trained with single modalities until convergence;

2. The weights are then ported to a multimodal deep autoencoder with one (or more) shared hidden layers that will codify our multimodal representation;

3. Training will have instances that will try to reconstruct two modalities from one, reconstruct one modality from two, and reconstruct one modality based on the other.

The architecture described in Vukotić et al. (2016) (BiDNN) will also be used to compare results between the multimodal approaches. The simple autoencoder described above was also compared in Vukotić et al. (2016), and we replicate this experiment with our dataset.

## 4. Experiments

The experiments described in this paper were executed in a multimodal dataset composed of 6,400 textual descriptions of e-commerce electronic products paired with their respective images. Not all the textual descriptions have a related image. The original downloaded images have $(55 \times 55 \times 3)$ pixels each. Each product is assigned to a class automatically extracted by web crawlers.

The 10 classes in this dataset are:

- **Automotive** – items related to automobile sound systems, tires and related electronic gadgets;

- **House and Electronics** – products related to kitchen appliances and house maintenance;

- **Games** – video-game related products such as consoles, joysticks and games;

- **Hardware** – computer components such as processors, GPUs and coolers;

- **Computing** – stand-alone products that are related to computers such as routers, no-breaks and power cables;
- **Stationary** – office products such as organizing boxes and tacks;
- **Peripherals** – computer peripherals such as headphones, keyboards and mouses;
- **Used** – an amalgam of all the other products in used conditions and user-made descriptions;
- **Smartphones** – smartphones and its accessories;
- **Telephony** – landlines, cables and radios.

Single-modality and multi-modality experiments have been executed with five simple classification algorithms:

- Support Vector Machine with both stochastic gradient descent (SGD) and logistic regression learning (SVC);
- Binary Support Vector Machine (one-versus-all approach) with both stochastic gradient descent (SGD-B) and logistic regression learning (SVC-B);
- Multilayer Dense Neural Network (MLP) with 256, 128 and 64 neurons on each layer, respectively;

All representations except the VGG19-derived ones were generated in a 128 dimensional space. Word2Vec and GloVe word representations were calculated for all words in a document and averaged to create a 128-dimension vector. To generate Word2Vec and GloVe representations, we used the tokenized texts without rare (less than 2 occurrences on the corpus) or large (more than 50 letters) tokens. To generate LSI and LDA word representations, each text was first converted to Bag-of-Words and these count values were used to generate *term frequency–inverse document frequency* (TFIDF) (Wu and Salton, 1981) values. Each image had its SURF, SIFT and ORB descriptors extracted, and a Bag-of-Visual-Words was generated by a k-means clustering algorithm, with $k = 128$. Images were passed through the VGG19 network and a vector with 1000 positions was generated for each image. The scripts were written in Python, with CUDA acceleration (Nickolls et al., 2008) and Tensorflow (Abadi et al., 2015) in the neural network training phases. The classification experiment used 10-fold cross-validation.

### 4.1. Baselines

Firstly, we tested single-modality classification and early-stage feature fusion, concatenating the vectors to generate 256-dimension vectors (or 1128-dimension vectors in the VGG19 modality) for combined representations.

The combinations were made with two different textual representations (text×text), two different image representations (image×image) or one representation of each (image×text). Although simple, this approach improved classification accuracy and recall on some multimodal combinations, with the best results depicted in bold on Table 1. These results show that textual classification outperforms its image counterpart by a large amount in this domain.

| Modalities | Representation | SGD | SVC | MLP | SGD-B | SVC-B |
|---|---|---|---|---|---|---|
| Text | GloVe | 0.71 | 0.74 | 0.80 | 0.70 | 0.74 |
| | W2V | 0.70 | 0.74 | 0.80 | 0.71 | 0.73 |
| | LDA | 0.62 | 0.68 | 0.75 | 0.65 | 0.68 |
| | LSI | 0.84 | 0.85 | **0.87** | 0.83 | **0.87** |
| Image | SIFT | 0.39 | 0.38 | 0.40 | 0.38 | 0.40 |
| | SURF | 0.36 | 0.38 | 0.36 | 0.36 | 0.38 |
| | ORB | 0.29 | 0.32 | 0.30 | 0.29 | 0.29 |
| | VGG19 | 0.38 | 0.44 | **0.57** | 0.40 | 0.45 |
| Image×Image | SIFT+SURF | 0.45 | 0.42 | 0.43 | 0.41 | 0.41 |
| | SIFT+ORB | 0.41 | 0.40 | 0.41 | 0.41 | 0.37 |
| | SIFT+VGG19 | 0.47 | 0.50 | **0.53** | 0.45 | 0.49 |
| | SURF+ORB | 0.38 | 0.38 | 0.39 | 0.38 | 0.36 |
| | SURF+VGG19 | 0.46 | 0.49 | 0.52 | 0.48 | 0.48 |
| | ORB+VGG19 | 0.45 | 0.45 | 0.49 | 0.46 | 0.46 |
| Text×Text | GloVe+W2V | 0.76 | 0.79 | 0.83 | 0.76 | 0.81 |
| | GloVe+LDA | 0.76 | 0.78 | 0.81 | 0.76 | 0.79 |
| | GloVe+LSI | 0.83 | 0.85 | 0.84 | 0.82 | 0.86 |
| | W2V+LDA | 0.75 | 0.78 | 0.82 | 0.75 | 0.78 |
| | W2V+LSI | 0.86 | **0.88** | 0.87 | 0.85 | 0.86 |
| | LDA+LSI | 0.84 | 0.86 | 0.84 | 0.85 | 0.87 |
| Image×Text | SIFT+W2V | 0.72 | 0.72 | 0.74 | 0.71 | 0.75 |
| | SIFT+GloVe | 0.74 | 0.71 | 0.80 | 0.71 | 0.72 |
| | SIFT+LDA | 0.66 | 0.68 | 0.69 | 0.66 | 0.68 |
| | SIFT+LSI | 0.84 | 0.84 | 0.84 | 0.83 | **0.86** |
| | SURF+W2V | 0.70 | 0.73 | 0.75 | 0.71 | 0.72 |
| | SURF+GloVe | 0.68 | 0.71 | 0.80 | 0.72 | 0.71 |
| | SURF+LDA | 0.65 | 0.66 | 0.67 | 0.65 | 0.65 |
| | SURF+LSI | 0.84 | 0.84 | 0.83 | 0.83 | 0.85 |
| | ORB+W2V | 0.68 | 0.71 | 0.73 | 0.71 | 0.69 |
| | ORB+GloVe | 0.70 | 0.72 | 0.76 | 0.71 | 0.69 |
| | ORB+LDA | 0.61 | 0.66 | 0.65 | 0.61 | 0.66 |
| | ORB+LSI | 0.83 | 0.84 | 0.81 | 0.82 | 0.85 |
| | VGG19+W2V | 0.73 | 0.74 | 0.78 | 0.72 | 0.74 |
| | VGG19+GloVe | 0.70 | 0.74 | 0.81 | 0.69 | 0.71 |
| | VGG19+LDA | 0.68 | 0.72 | 0.75 | 0.70 | 0.72 |
| | VGG19+LSI | 0.80 | 0.83 | 0.83 | 0.81 | 0.84 |

Table 1: Unimodal and early-stage multimodal fusion experiments. Metric used is F-Score. Best scores for each modality are highlighted in bold.

| Modalities | Representation | MMAE | BIDNN |
|---|---|---|---|
| Image×Image | SIFT+SURF | 0.38 | 0.37 |
| | SIFT+ORB | 0.37 | 0.36 |
| | SIFT+VGG19 | **0.39** | 0.44 |
| | SURF+ORB | 0.35 | 0.34 |
| | SURF+VGG19 | **0.39** | **0.46** |
| | ORB+VGG19 | 0.32 | 0.43 |
| Text×Text | GloVe+W2V | 0.73 | 0.79 |
| | GloVe+LDA | 0.60 | 0.76 |
| | GloVe+LSI | 0.50 | 0.82 |
| | W2V+LDA | **0.74** | 0.82 |
| | W2V+LSI | **0.74** | **0.84** |
| | LDA+LSI | 0.61 | 0.83 |
| Image×Text | SIFT+W2V | 0.66 | 0.75 |
| | SIFT+GloVe | 0.52 | 0.61 |
| | SIFT+LDA | 0.60 | 0.66 |
| | SIFT+LSI | 0.32 | 0.74 |
| | SURF+W2V | 0.68 | 0.72 |
| | SURF+GloVe | 0.49 | 0.61 |
| | SURF+LDA | 0.58 | 0.63 |
| | SURF+LSI | 0.38 | 0.77 |
| | ORB+W2V | **0.70** | 0.72 |
| | ORB+GloVe | 0.44 | 0.54 |
| | ORB+LDA | 0.61 | 0.62 |
| | ORB+LSI | 0.39 | 0.73 |
| | VGG19+W2V | 0.67 | 0.73 |
| | VGG19+GloVe | 0.51 | 0.54 |
| | VGG19+LDA | 0.65 | 0.67 |
| | VGG19+LSI | 0.37 | **0.79** |

Table 2: Multimodal fusion experiments. Metric used is F-Score on a Multilayer Perceptron model. Best scores between modalities and algorithms are highlighted in bold.

### 4.2. Our approach

After finding our baselines, we tested our proposed approach by using multimodal features generated by a Deep Multimodal Autoencoder (MMAE) (Ngiam et al., 2011)

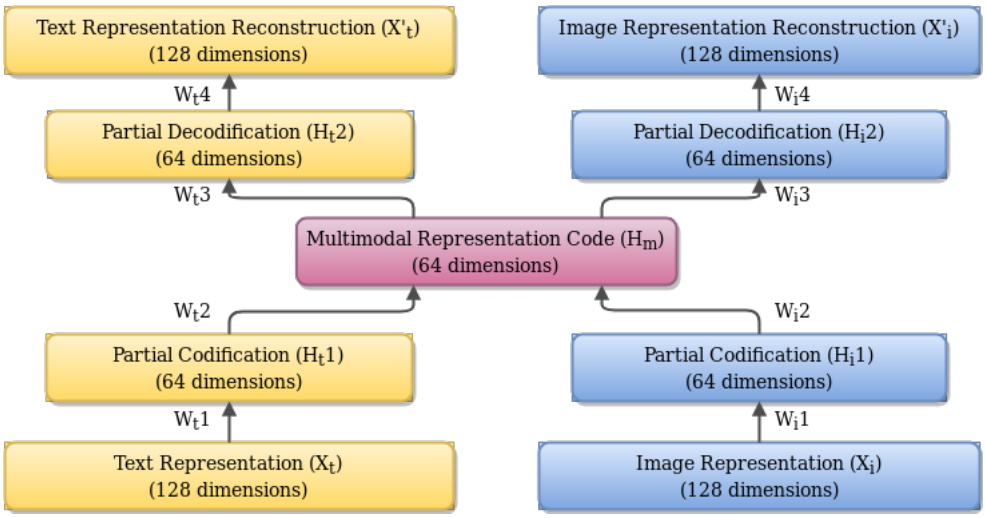


Figure 1: Deep Multimodal (MMAE) architecture used in the experiments.

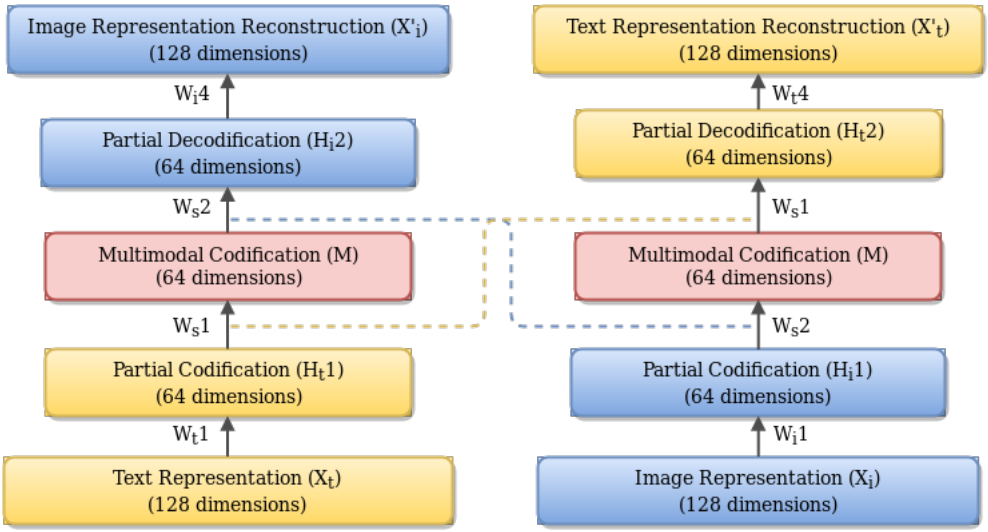


Figure 2: Bidirectional Deep Autoencoder (BIDNN), with shared weights.

and a Bidirectional Neural Network (BIDNN) (Vukotić et al., 2016) with tied weights, as illustrated in Figures 1 and 2.

These features have 64 dimensions, fusing text and image representations. Both models were trained in batches of 256 instances for 1000 epochs, optimized using ADADELTA (Zeiler, 2012) with binary cross-entropy loss. The results of these tests are depicted in Table 2. Only the Multilayer Perceptron classification model results are shown in this table, as the other models achieved poor F-Scores (below 0.4).

The results show that the Bidirectional Neural Network is consistently better in this task, corroborating the experiments in (Vukotić et al., 2016). But, when compared to the previous experiment, lower F-Scores were obtained in the neural multimodal features: the best results obtained with these features were 0.74 (Multimodal Autoencoder) and 0.84 (Bidirectional Neural Network), lower than the results obtained by early-fusion of textual features (0.87 using W2V+LSI) and comparable to the results of unimodal features (GloVe, W2V and LSI had results higher or equal than 0.80).

## 5. Discussion

As can be noticed by the values in Table 1, LSI seems to be the best text representation model with 0.87 F-Score with MLP and SVC-B. It is also the best one to combine with image representation models since its performance with SIFT (SIFT+LSI), SURF (SURF+LSI) and ORB (ORB+LSI) was really close to the best values achieved when only text representations were used. We believe that this LSI's best performance is related to the fact that most descriptions of e-commerce products are mainly composed of highly specialized keyword groups (*eg.*, "smartphone" and "mobile" in the "smartphone" class, and "DPI" and "anti-ghosting" in "peripherals"), thus topical representations can adequately summarize concepts in this domain. And, as these products rarely share the same visual concepts of oth-

ers in their class (*eg.*,"smartphones" had images of smartphones, smartphone cables and accessories), visual descriptors are not efficient in this task. From these early-stage multimodal results we can conclude that text (at least, LSI) can help image in the classification scenario tested in our experiments.

Regarding our proposed neural multimodal approach, we could not significantly improve classification accuracy using multimodal features obtained via neural means (MMAE and BIDNN). In order to visualize the similarity between representations, we generated comparison images for selected products [1]. The four most similar products (top-4) to a target image of a printer are depicted in Figures 3 (using only neural image features), 4 (using only textual features) and 5 (using multimodal neural features), from a special receipt printer. In Figure 3, image representations could approximate other printers and rectangular objects to its immediate vectorial vicinity. In Figure 4, using textual representations, the two closest product vectors had no relation whatsoever to printing: one represents a repair toolkit for a CPU cooler and the other is a motion sensor. In Figure 5, using multimodal features, the four closest products are all printers, and the first two are receipt printers just as the target product is.

Analyzing product similarity, we can see that some products can be grouped more accurately using multimodal features (as seen in Figure 5). The information gain from multimodal features was not relevant for classification in our corpus since there is too much noise in certain categories (*eg.*, in Figure 4 in which a printer, a cable and a sensor were grouped together in the same category). However, we think that multimodal features can be used in other tasks involving similarity such as e-commerce search queries: improving comparison speed and quality using these features to navigate monolithic databases can increase site performance.

As there are few linguistic resources for the Portuguese language in this area, we intend to make this set a starting point for other works. The dataset can be obtained on request, and the code will be released in a public repository after the publication of this paper.

## 6. Acknowledgements

This research is part of the MMeaning project, supported by São Paulo Research Foundation (FAPESP), grant #2016/13002-0, and was also partly funded by the Brazilian Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

## 7. Bibliographical References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Andrew, G., Arora, R., Bilmes, J., and Livescu, K. (2013). Deep Canonical Correlation Analysis. Proceedings of the 30th International Conference on Machine Learning (ICML-13), 28:1247–1255.

Atrey, P. K., Hossain, M. A., El Saddik, A., and Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: A survey. Multimedia Systems, 16(6):345–379.

Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (2008). Speeded-Up Robust Features (SURF). Computer Vision and Image Understanding, 110(3):346–359.

Blei, D. M., Edu, B. B., Ng, A. Y., Edu, A. S., Jordan, M. I., and Edu, J. B. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3:993–1022.

Bourlard, H. and Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. Biological Cybernetics, 59(4):291–294, Sep.

Calonder, M., Lepetit, V., Strecha, C., and Fua, P. (2010). BRIEF: Binary robust independent elementary features. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 6314 LNCS, pages 778–792.

Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. Multiple Classifier Systems, 1857:1–15.

Eskevich, M., Aly, R., Racca, D. N., Ordelman, R., Chen, S., and Jones, G. J. (2014). The search and hyperlinking task at MediaEval 2014. In CEUR Workshop Proceedings, volume 1263.

Friedman, J. H. (1987). Exploratory projection pursuit. Journal of the American Statistical Association, 82(397):249–266.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., et al. (2013). Devise: A deep visual-semantic embedding model. In Advances in neural information processing systems, pages 2121–2129.

Harris, Z. S. (1954). Distributional structure. Word, 10(2-3):146–162.

Information Resources Management Association. (2012). Data mining: Concepts, methodologies, tools, and applications.

Landauer, T. K., Folt, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. Discourse processes, 25(2):259–284.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In Eric P. Xing et al., editors, Proceedings of the 31st International Conference on Machine Learning, volume 32 of Proceedings of Machine Learning Research, pages 1188–1196, Bejing, China, 22–24 Jun. PMLR.

LeCun, Y. and Cortes, C. (2010). MNIST handwritten digit database. AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist.

Lowe, D. G. (2004). Distinctive image features from scale-

[1]The representations generated in the previous experiments were used to measure the vector distance between these selected products and all others in their vector space.


Figure 3: Top-4 similar products using VGG19-generated features.


Figure 4: Top-4 similar products using LSI-generated features.

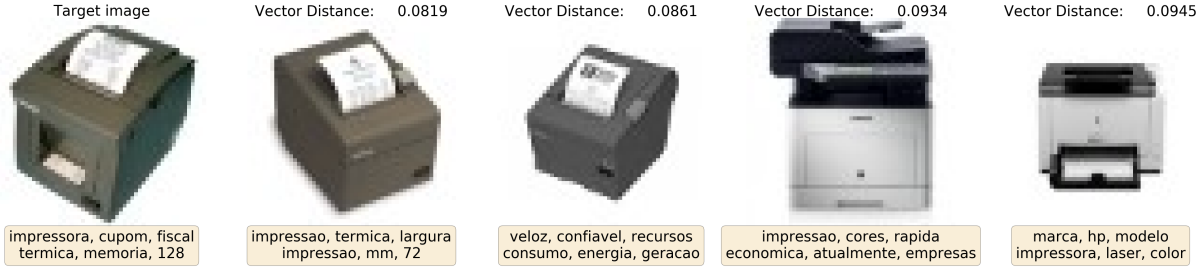


Figure 5: Top-4 similar products using Bidirectional Neural Network features.

invariant keypoints. International Journal of Computer Vision, 60(2):91–110.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, pages 1–12.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). Multimodal Deep Learning. Proceedings of The 28th International Conference on Machine Learning (ICML), pages 689–696.

Nickolls, J., Buck, I., Garland, M., and Skadron, K. (2008). Scalable parallel programming with CUDA. AMC Queue, 6(April):40–53.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543.

Radová, V. and Psutka, J. (1997). An approach to speaker identification using multiple classifiers. In Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on, volume 2, pages 1135–1138. IEEE.

Rosten, E. and Drummond, T. (2006). Machine learning for high-speed corner detection. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 3951 LNCS, pages 430–443.

Rublee, E., Rabaud, V., Konolige, K., and Bradski, G. (2011). ORB: An efficient alternative to SIFT or SURF. In Proceedings of the IEEE International Conference on Computer Vision, pages 2564–2571.

Salakhutdinov, R. and Hinton, G. (2009). Deep belief networks. Scholarpedia, pages 4–5.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556.

Van Gelder, T. (1992). Defining 'distributed representation'. Connection Science, 4(3/4):175.

Vukotić, V., Raymond, C., and Gravier, G. (2016). Bidirectional joint representation learning with symmetrical deep neural networks for multimodal and cross-modal applications. In Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, ICMR '16, pages 343–346, New York, NY, USA. ACM.

Wang, J., Kankanhalli, M. S., Yan, W., and Jain, R. (2003). Experiential sampling for video surveillance. In First ACM SIGMM international workshop on Video surveillance, pages 77–86. ACM.

Wang, W., Arora, R., Livescu, K., and Bilmes, J. (2016). On Deep Multi-View Representation Learning: Objectives and Optimization. Arxiv Preprint, XX(XX):XX.

Westbury, J. (1994). X-ray Microbeam Speech Production Database User's Handbook: Version 1.0 (June 1994). Waisman Center on Mental Retardation & Human Development.

Wu, H. and Salton, G. (1981). A comparison of search term weighting: term relevance vs. inverse document frequency. In ACM SIGIR Forum, volume 16, pages 30–39.

Yang, M.-T., Wang, S.-C., and Lin, Y.-Y. (2005). A multimodal fusion system for people detection and tracking. International journal of imaging systems and technology, 15(2):131–142.

Yang, J., Jiang, Y.-G., Hauptmann, A. G., and Ngo, C.-W. (2007). Evaluating bag-of-visual-words representations in scene classification. In Proceedings of the international workshop on Workshop on multimedia information retrieval - MIR '07, page 197.

Zeiler, M. D. (2012). ADADELTA: an adaptive learning rate method. CoRR, abs/1212.5701.