GATED MULTIMODAL UNITS FOR INFORMATION FUSION

Arevalo, John
Dept. of Computing Systems and Industrial Engineering
Universidad Nacional de Colombia
Cra 30 No 45 03-Ciudad Universitaria
jearevaloo@unal.edu.co

Solorio, Thamar
Dept. of Computer Science
University of Houston
Houston, TX 77204-3010
solorio@cs.uh.edu

Montes-y-Gómez, Manuel
Instituto Nacional de Astrofísica, Óptica y Electrónica
Computer Science Department
Luis Enrique Erro No. 1, Sta. Ma. Tonantzintla
C.P. 72840 Puebla, Mexico
smmontesg@inaoep.mx

González, Fabio A.
Dept. of Computing Systems and Industrial Engineering
Universidad Nacional de Colombia
Cra 30 No 45 03-Ciudad Universitaria
fagonzalezo@unal.edu.co

ABSTRACT

This paper presents a novel model for multimodal learning based on gated neural networks. The Gated Multimodal Unit (GMU) model is intended to be used as an internal unit in a neural network architecture whose purpose is to find an intermediate representation based on a combination of data from different modalities. The GMU learns to decide how modalities influence the activation of the unit using multiplicative gates. It was evaluated on a multilabel scenario for genre classification of movies using the plot and the poster. The GMU improved the macro f-score performance of single-modality approaches and outperformed other fusion strategies, including mixture of experts models. Along with this work, the MM-IMDb dataset is released which, to the best of our knowledge, is the largest publicly available multimodal dataset for genre prediction on movies.

1 INTRODUCTION

Representation learning methods have received a lot of attention by researchers and practitioners because of its successful application to complex problems in areas such as computer vision, speech recognition and text processing (LeCun et al., 2015). Most of these efforts have concentrated on data involving one type of information (images, text, speech, etc.), despite data being naturally multimodal. Multimodality refers to the fact that the same real-world concept can be described by different views or data types. Collaborative encyclopedias (such as Wikipedia) describe a famous person through a mixture of text, images and, in some cases, audio. Users from social networks comment events like concerts or sport games with small phrases and multimedia attachments (images/videos/audios). Medical records are represented by a collection of images, sound, text and signals, among others. The increasing availability of multimodal databases from different sources has motivated the development of automatic analysis techniques to exploit the potential of these data as a source of knowledge in the form of patterns and structures that reveal complex relationships (Bhatt & Kankanhalli, 2011; Atrey et al., 2010). In recent years, multimodal tasks have acquired attention by the representation learning community. Strategies for visual question answering (Antol et al., 2015), or image captioning (Vinyals et al., 2015; Xu et al., 2015; Johnson et al., 2015) have developed interesting ways of combining different representation learning architectures.

Most of these models are focused on mapping from one modality to another or solving an auxiliary task to create a common representation with the information of all modalities. In this work, we design a novel module that combines multiple sources of information, which is optimized with respect to the end goal objective function. Our proposed module is based on the idea of gates for selecting which parts of the input are more likely to contribute for correctly generating the desired output. We

Figure 1: Predictions of genre labels depending on the input modality. Red and blue labels indicate false positives and true positives respectively.

use multiplicative gates that assign importance to various features simultaneously, creating a rich multimodal representation that does not require manual tuning, but instead it learns directly from the training data. Our gated model can be reused in different network architectures for solving different tasks, and can be optimized end-to-end with other modules in the architecture using standard gradient-based optimization algorithms.

As an application use case, we explore the task of identifying a movie genre based on its plot and its poster. Genre classification has several application areas like document categorization (Kanaris & Stamatatos, 2009), recommendation systems (Makita & Lenskiy, 2016a), and information retrieval systems, among others. Figure 1 depicts the challenging task of assigning genres to a particular movie based solely on the usage of one modality. Such predictions were done with *MaxoutMLP_w2v* and *VGG_transfer* approaches (See Section 3), both of them are models based on representation learning. It can be seen that even a human might be confused if both modalities are not available. The main hypothesis of this work is that a model using gating units, in contrast to a hand-coded multimodal fusion architecture, will be able to learn an input-dependent gate-activation pattern that determines how each modality contribute to the output of hidden units.

The rest of the paper is organized as follows: Section 2 presents a literature review and some considerations of the previous work. Section 3 describes the methods used as baseline as well as our representation-leaning-based model proposed. Section 4 presents the experimental evaluation setup along with the details of the MM-IMDb dataset. Section 5 shows and discusses the results for movie genre classification. Finally, Section 6 draws the conclusions and future work.

2 Related work

2.1 Multimodal fusion

Different reviews (Atrey et al., 2010; Bhatt & Kankanhalli, 2011; Li Deng, 2014; Deng, 2014) have summarized strategies that addressed multimodal analysis. Most of the collected works claimed the superiority of multimodal over unimodal approaches for automatic analysis tasks. A conventional multimodal analysis system receives as input two or more modalities that describe a particular concept. The most common multimodal sources are video, audio, images and text. In recent years there has been a consensus with respect to the use of representation learning models to characterize the information of this kind of sources (LeCun et al., 2015). However, the way that such extracted features are combined is still in exploration.

Multimodal combination seeks to generate a single representation that makes easier automatic analysis tasks when building classifiers or other predictors. A simple approach is to concatenate features to get a final representation (Kiela & Bottou, 2014; Pei et al., 2013; Suk & Shen, 2013). Although it is a straightforward strategy, it ignores inherent correlations between different modalities.

More complex fusion strategies include Restricted Boltzmann Machines (RBMs) and autoencoders. Ngiam et al. (2011) concatenated higher level representations and train two RBMs to reconstruct the original audio and video representations respectively. Additionally, they trained a model to recon-

struct both modalities given only one of them as input. In an interesting result, Ngiam et al. (2011) were able to mimic a perceptual phenomenon that demonstrates an interaction between hearing and vision in speech perception known as McGurk effect. A similar approach was proposed by Srivastava & Salakhutdinov (2012). They modified feature learning and reconstruction phases with Deep Boltzmann Machines. Authors claimed that such strategy is able to exploit large amounts of unlabeled data by improving the performance in retrieval and annotation tasks. Other similar strategies propose to fusion modalities using neural network architectures (Andrew et al., 2013; Feng et al., 2013; Kang et al., 2012; Kiros et al., 2014a; Lu et al., 2014; Mao et al., 2014; Tu et al., 2014; Wu et al., 2013) with two input layers separately and including a final supervised layer such as softmax regression classifier.

An alternative approach involves an objective or loss function suited for the target task (Akata et al., 2014; Frome et al., 2013; Kiros et al., 2014b; Mao et al., 2014; Socher et al., 2013; 2014; Zheng et al., 2014). These strategies usually assume that there exists a common latent space where modalities can express the same semantic concept through a set of transformations of the raw data. The semantic embedding representations are such that two concepts are similar if and only if their semantic embeddings are close (Norouzi et al., 2014). In (Socher et al., 2013) a multimodal strategy to perform zero-shot classification was proposed. They trained a word-based neural network model (Huang et al., 2012) to represent textual information, whilst use unsupervised feature learning models proposed in (Coates & Ng, 2011) to get image representation. The fusion was done by learning an image linear mapping to project images into the semantic word space learned in the neural network model. Additionally a Bayesian framework was included to decide whether an image is of a seen or unseen class. Frome et al. (2013) learn the image representation using a CNN trained with the Imagenet dataset and a word-based neural language model (Mikolov et al., 2013b) to represent the textual modality. To perform the fusion they re-train CNN using text representation as targets. This work outperforms scalability with respect to (Socher et al., 2013) from 2 to 20,000 unknown classes in the zero-shot learning task. A modified strategy of Frome et al. (2013) was presented by Norouzi et al. (2014). Instead of re-train the CNN network, they built a convex combination with probabilities estimated by the classifier and semantic embedding vector of the unseen label. This simple strategy outperforms state-of-the-art results. Because the cost function involves both multimodal combination and supervision, these family of models are tied to the task of interest. Thus, if the domain or task conditions changes, adaptations are required.

The proposed model is closely related to the mixture of experts (MoE) approach (Jacobs et al., 1991). However, the common usage of MoE is focused on performing decision fusion, i.e. combining predictors to address a supervised learning problem (Yuksel et al., 2012). Our model is devised as a new component in the representation learning scheme, making it independent from the final task (e.g. classification, regression, unsupervised learning, etc) provided that the defined cost function be differentiable.

2.2 Movie genre classification

With respect to movie genre classification, several strategies also have been proposed. These strategies have used different modalities to characterize each movie, such as textual features, image features and multimedia features (audio and/or video). Huang et al. (2007) were one of the first teams exploring this task. They classified movie previews into 3 genres by extracting handcrafted features from the video and training a decision tree classifier. They evaluated the model using 44 films. Using only textual modality, Shah et al. (2013) performed single-label genre classification of movie scripts using clustering algorithms with 260 movies. Later, combining two modalities, Pais et al. (2012) classified movies between drama and non-drama using visual and textual features with 107 samples. Hong & Hwang (2015) explored different PLSA models to combine 3 modalities: audio, image and text to predict genre of movie previews. It was single label classification with 4 genres for 140 movies taken from IMDb.

Recently, Fu et al. (2015) used a set of handcrafted visual features for poster characterization and bag-of-words for synopsis. Then, they trained one SVM per each modality to combine their predictions. The dataset contained $2,400$ movies with one genre (out of 4) each.

The previous mentioned works present this problem in a single label setup. However, a more realistic scenario would be multilabel, since most of the movies belong to more than one genre, (e.g.

Matrix(2000) is a Sci-fi/Action movie). In this setup, Anand (2014) explores the efficiency of using keywords and users' tags to perform multilabeling using the movies from MovieLens 1M dataset which contains $1,700$ movies. Also Ivasic-Kos et al. (2014; 2015) performed multilabel classification using handcrafted features from posters, with $1,500$ samples for 6 genres. Makita & Lenskiy (2016a;b) use movie ratings matrix and genre correlation matrix to predict the genre. It used a smaller version of the Movielens dataset with 18 movie genres.

Most of the above works have used the publicly available MovieLens datasets. However, there is not a single experimental setup defined so that all methods can be systematically compared. Also, to the best of our knowledge, none of the previous works contain more than $10,000$ samples. With this work we will release a dataset created with the movies of the MovieLens 20M dataset. We include not only genre, poster and plot information used in this work, but also the poster of the movie as well as more than 50 characteristics taken from the IMDb website. We will also release the source code to automatically add more movies and genres.

3 METHODS

This paper presents a neural-network-based strategy for multilabel classification of multimodal data. The key component of the strategy is a novel type of hidden unit, the Gated Multimodal Unit (GMU), which learns to decide how modalities influence the activation of the unit using gates. The details of the GMU are presented in Subsection 3.1.

Statistical properties usually are not shared across modalities (Srivastava & Salakhutdinov, 2012). And thus, they require different representation strategies according to the nature of data. This work explored several strategies to address text and visual representation. For text information we evaluated word2vec models, n-grams models and RNN models. The details are discussed in Subsection 3.2. On the other hand, two different convolutional neural networks were evaluated for processing visual data and are presented in Subsection 3.3.

3.1 GATED MULTIMODAL UNIT FOR MULTIMODAL FUSION

Multimodal learning is closely related to data fusion. Data fusion looks for optimal ways of combining different information sources into an integrated representation that provides more information than the individual sources (Bhatt & Kankanhalli, 2011). This fusion can be performed at different levels, that can be categorized into two broad categories: feature fusion and decision fusion. Feature fusion, also called early fusion, looks for a subset of features from different modalities, or combinations of them, that better represent the information needed to solve a particular problem. On the other hand, decision fusion, or late fusion, combines decisions from different systems, e.g. classifiers, to produce consensus. This consensus may be reached by a simple average, a voting system or a more complex Bayesian framework.

In this work we present a model, based on gated neural networks, for data fusion that combines ideas from both feature and decision fusion. The model, called Gated Multimodal Unit (GMU), is inspired by the flow control in recurrent architectures like GRU or LSTM. A GMU is intended to be used as an internal unit in a neural network architecture whose purpose is to find an intermediate representation based on a combination of data from different modalities. Figure 2.a depicts the structure of a GMU. Each x_i corresponds to a feature vector associated with modality i. Each feature vector feeds a neuron with a $\tanh$ activation function, which is intended to encode an internal representation feature based on the particular modality. For each input modality, x_i, there is a gate neuron (represented by σ nodes in the diagram), which controls the contribution of the feature calculated from x_i to the overall output of the unit. When a new sample is fed to the network, a gate neuron associated to modality i receives as input the feature vectors from all the modalities and uses them to decide whether the modality i may contribute, or not, to the internal encoding of the particular input sample.

Figure 2.b shows a simplified version of the GMU for two input modalities, x_v (visual modality) and x_t (textual modality), that will be used in the remaining of the paper. It should be noted that both models are not completely equivalent, since in the bimodal case the gates are tied. Such weight tying constraints the model, so that the units trade off between both modalities while they use less

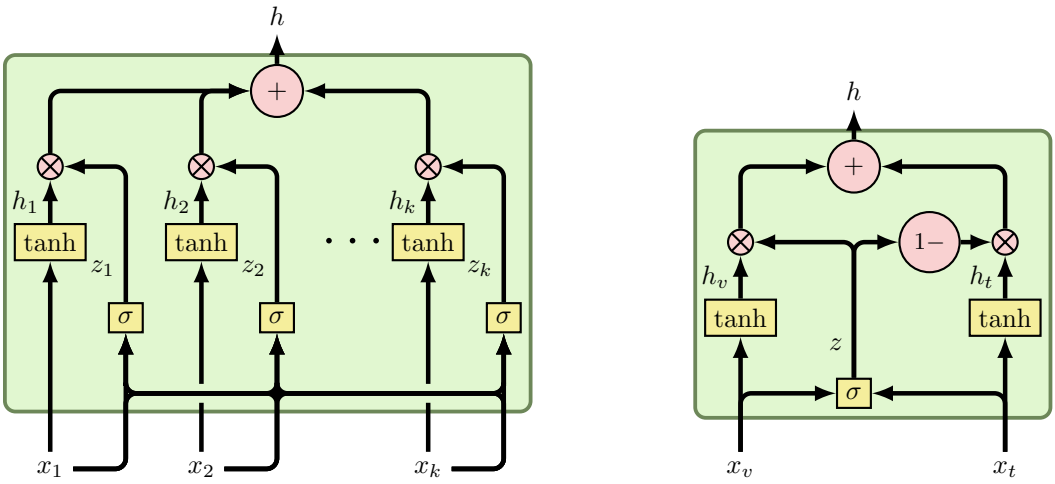


Figure 2: Illustration of gated units. a) The proposed model to use with more than two modalities. b) A simplification for the bimodal approach.

parameters than the multimodal case. The equations governing this GMU are as follows:

$$
\begin{aligned}
h_v &= \tanh\left(W_v \cdot x_v\right) \\
h_t &= \tanh\left(W_t \cdot x_t\right) \\
z &= \sigma\left(W_z \cdot [x_v, x_t]\right) \\
h &= z * h_v + (1 - z) * h_t \\
\Theta &= \{W_v, W_t, W_z\}
\end{aligned}
$$

with Θ the parameters to be learned and $[\cdot, \cdot]$ the concatenation operator. Since all are differentiable operations, this model can be easily coupled with other neural network architectures and trained with stochastic gradient descent.

3.2 Text representation

Text representation is a critical step when classification tasks are addressed using machine learning methods. Traditional approaches are based on counting frequencies of n-gram occurrences such as words or sequences of characters (e.g. bag-of-words models). The main drawback of such approaches is the difficulty to model relationships between words and their context. An alternative approach was initially proposed by Bengio et al. (2003), by building a language model based on a neural network architecture (NNLM). The NNLM was able to learn distributed representations of words that capture contextual information. Later, this model was simplified to deal with large corpora by removing hidden layers in the neural network architecture (word2vec) (Mikolov et al., 2013a). This is a fully unsupervised model that takes advantage of large sets of unlabeled documents. Herein, three text representations were evaluated:

n-gram Following the strategy proposed by Kanaris & Stamatatos (2009), we used the n-gram strategy for representing text. Despite their simplicity, n-gram models have shown to be a competitive baseline.

Word2Vec Word2vec is an unsupervised learning algorithm that finds a vector representation for each word based on its context (Mikolov et al., 2013a). It has been shown that this model is able to find semantic and syntactic relationships using arithmetic operations between the vectors. Based on this property, we represent a movie as the average of the vectors of words in the plot outline. The main motivation to aggregate word2vec vectors is the property of additive compositionality that this representation has exposed over different set of tasks such as word analogies. The usual way to aggregate is to sum vectors. We instead take the average to avoid large input values to the neural network.

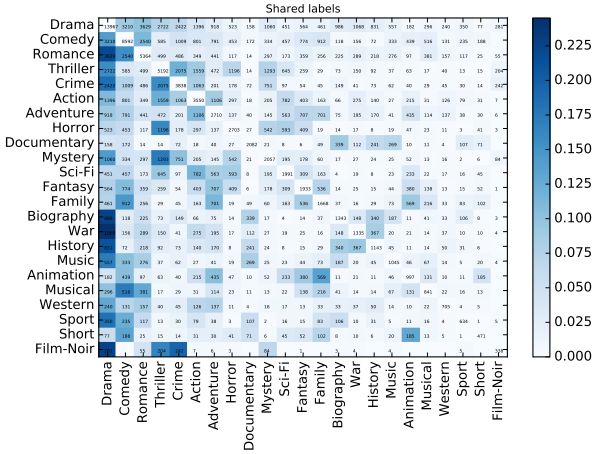


Figure 3: Co-ocurrence matrix of genre tags

Recurrent neural network Here we take the plot outline as a sequence of words and train a supervised recurrent neural network. We evaluated two variants. The first one (*RNN_w2v*) is a transfer learning model that takes as input the word vectors of word2vec as representations. The second one learns the word vectors from scratch (*RNN_end2end*).

3.3 VISUAL REPRESENTATION

In computer vision tasks, Convolutional neural networks have become the *de facto* standard. It has been shown that CNN models trained with a huge amount of data are able to learn common features shared across different domains. This characteristic is usually exploited by transfer learning approaches. For visual representation we explored 2 strategies: transfer learning and end-to-end training.

VGG Transfer In this approach, the VGG Network (Simonyan & Zisserman, 2014) trained with the ImageNet dataset is used as feature extractor by taking the last hidden activations as the visual representation.

End2End CNN Here, a CNN with 5 convolutional layers and an MLP (see Section 3.4) on top was trained from scratch.

3.4 CLASSIFICATION MODEL

Based on the defined representation, we explored two methods to map from feature vectors to genre classification. In particular we explored a simple Logistic regression and a neural network architecture. This is a multilayer perceptron (MLP) with two fully connected layers and maxout activation function. In particular, the maxout activation function $h_i : \mathbb{R}^n \rightarrow \mathbb{R}$ is a defined as:

$$h_i(\mathbf{s}) = \max_{j \in [1,k]} z_{i,j} \tag{1}$$

where $\mathbf{s} \in \mathbb{R}^n$ is the input vector, $z_{i,j} = \mathbf{s}^T \mathbf{W}_{...ij} + \mathbf{b}_{ij}$ is the output of the j-th linear transformation of the i-th hidden unit, and $\mathbf{W} \in \mathbb{R}^{d \times m \times k}$ and $\mathbf{b} \in \mathbb{R}^{m \times k}$ are learned parameters. It has been shown that maxout models with just 2 hidden units behave as universal approximators, while are less prone to saturate units (Goodfellow et al., 2013).

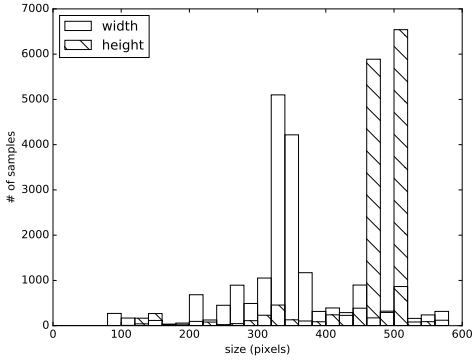


Figure 4: Size distribution of movie posters.

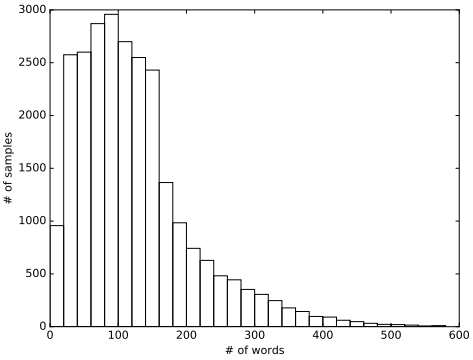


Figure 5: Length distribution of movie plots.

4 EXPERIMENTAL EVALUATION

4.1 MULTIMODAL IMDB DATASET

With this work we will make publicly available the Multimodal IMDb (**MM-IMDb**)[1] dataset. MM-IMDb dataset is built with the IMDb id's provided by the Movielens 20M dataset [2] that contains ratings of $27,000$ movies. Using the IMDbPY [3] library, movies which do not contain their poster image were filtered out. As the final result, the MM-IMDb dataset comprises $25,959$ movies along with their plot, poster, genres and other 50 additional metadata fields such as year, language, writer, director, aspect ratio, etc.

Notice that one movie may belong to more than one genre. Figure 3 shows the co-occurrence matrix, where the color bar indicates the representative co-occurrence per row, while Figure 4 and Figure 5 depict the distribution of the movie poster sizes and length of movie plots respectively. Each plot contains on average 92.5 words, while the longest one contains $1,431$ words and the average of genres per movie is 2.48. In this work, we defined the task of movie genre prediction based on its plot and image poster. Nevertheless, the additional metadata information encourages other interesting tasks such as rating prediction and content-based retrieval, among others.

4.2 EXPERIMENTAL SETUP

The MM-IMDb dataset has been split in three subsets. Train, development and test subsets contain 15552, 2608 and 7799 respectively. The distribution of samples is listed in Table 1. The sample was stratified so that training, dev and test sets comprises 60%, 10%, 30% samples of each genre respectively.

In the multilabel classification the performance evaluation can be more complex than traditional *multi-class* classification and the differences can be significant among several measures (Madjarov et al., 2012). Herein, four averages of the f-score (f_1) are reported: *samples* computes the f-score per sample and then averages the results, *micro* computes the f-score using all predictions at once, *macro* computes the f-score per genre and then averages the results. *weighted* is the same as *macro* with a weighted average based on the number of positive samples per genre. Concretely, we calculate them as follows (Madjarov et al., 2012):

$$f_1^{sample} = \frac{1}{N}\sum_{i=1}^{N}\frac{2\times|\hat{y}_i \cap y_i|}{|\hat{y}_i|+|y_i|} \quad f_1^{macro} = \frac{1}{Q}\sum_{j=1}^{Q}\frac{2\times \mathrm{p}_j \times \mathrm{r}_j}{\mathrm{p}_j+\mathrm{r}_j} \quad f_1^{weighted} = \frac{1}{Q^2}\sum_{j=1}^{Q} Q_j \frac{2\times \mathrm{p}_j \times \mathrm{r}_j}{\mathrm{p}_j+\mathrm{r}_j}$$

[1] http://lisi1.unal.edu.co/mmimdb/
[2] http://grouplens.org/datasets/movielens/
[3] http://imdbpy.sourceforge.net/

Table 1: Genre distribution per subset

Genre	Train	Dev	Test	Genre	Train	Dev	Test
Drama	8424	1401	4142	Family	978	172	518
Comedy	5108	873	2611	Biography	788	144	411
Romance	3226	548	1590	War	806	128	401
Thriller	3113	512	1567	History	680	118	345
Crime	2293	382	1163	Music	634	100	311
Action	2155	351	1044	Animation	586	105	306
Adventure	1611	278	821	Musical	503	85	253
Horror	1603	275	825	Western	423	72	210
Documentary	1234	219	629	Sport	379	64	191
Mystery	1231	209	617	Short	281	48	142
Sci-Fi	1212	193	586	Film-Noir	202	34	102
Fantasy	1162	186	585				

$$\mathrm{p}^{micro} = \frac{\sum_{j=1}^{Q} tp_j}{\sum_{j=1}^{Q} tp_j + \sum_{j=1}^{Q} fp_j} \quad \mathrm{r}^{micro} = \frac{\sum_{j=1}^{Q} tp_j}{\sum_{j=1}^{Q} tp_j + \sum_{j=1}^{Q} fn_j} \quad f_1^{micro} = \frac{2 \times \mathrm{p}^{micro} \times \mathrm{r}^{micro}}{\mathrm{p}^{micro} + \mathrm{r}^{micro}}$$

With N the number of examples; Q the number of labels; Q_j the number of true instances for the j-th label; p the precision, r the recall; $\hat{y}_i, y_i \in (0,1)^Q$ the prediction and ground truth binary tuples respectively; tp_j, fp_jandfn_j the number of true positives, false positives and false negatives for the j-th label respectively.

TEXTUAL REPRESENTATION

The pretrained Google Word2vec[4] embedding space was used. After intersecting the Google word2vec available words with the MM-IMDb plots, the final vocabulary contains 41,612 words. Other than lowercase, no text preprocessing was applied. Since it is our intention to measure how the network's depth affects the performance of the model, we also evaluate the architecture with a single fully connected layer. In order to compare the performance of this textual representation, we evaluate it using two publicly available datasets: *7genre* dataset that comprises 1,400 web pages with 7 disjoint genres and *ki-04* dataset that comprises 1,239 samples classified under 8 genres. We compare the model with the state of the art results (Kanaris & Stamatatos, 2009) which used character n-grams with structured information from the HTML tags to predict the genre of web pages.

VISUAL REPRESENTATION

Since the first approach was to use VGG as a feature extractor. This model is referred as *VGG_Transfer*. The second approach takes as input the raw images to a CNN. Since all the images do not have the same size, all images were scaled, and cropped when required, to 160×256 pixels keeping the aspect ratio. This CNN comprises 5 CNN layers of $5, 3, 3, 3, 3$ squared filters and 2×2 pool sizes. Each convolutional layer has 16 hidden units. The convolutional layers are connected with the *MaxoutMLP* on top.

MULTIMODAL REPRESENTATION

We evaluate 4 different ways to combine both modalities as baselines.

Average probability This can be seen as a late-fusion strategy. The probabilities obtained by the best model of each modality are averaged and thresholded.

concatenation Different works have found that a simple concatenation of representations of different modalities are good for combining the information (Suk & Shen, 2013; Pei et al., 2013; Kiela & Bottou, 2014). Herein, we concatenated both representations to train the *MaxoutMLP* architecture.

[4] https://code.google.com/archive/p/word2vec/

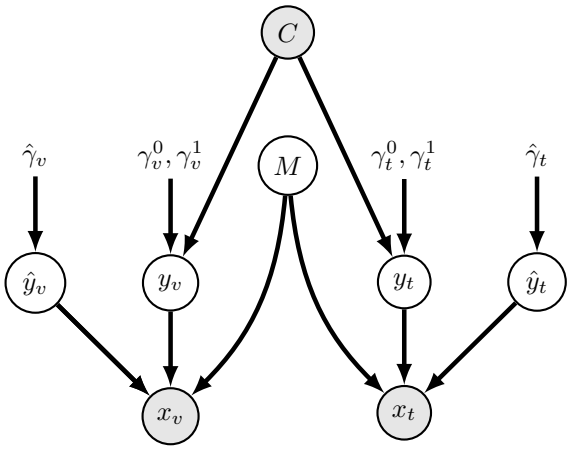


Figure 6: Generative model for the synthetic task. Grayed nodes represent visible variables, the other nodes represent hidden variables.

linear sum Following the way Vinyals et al. (2015) combine text and images representation into a single space, this model adds a linear transformation for each modality so that both outputs have the same size to be summed up and then followed by the *MaxoutMLP* architecture.

MoE The mixture of experts (MoE) (Jacobs et al., 1991) model was adapted for multilabel classification. two gating strategies were explored: *tied*, where a single gate multiplies all the logistics outputs, and *untied* where every logistic output has its own gate. Logistic regression and *MaxoutMLP* were evaluated as experts.

NEURAL NETWORK TRAINING

Neural network models were trained using using Batch Normalization scheme (Ioffe & Szegedy, 2015). This strategy applies a normalization step across samples that belong to the same batch, so that each hidden unit in the network receive a zero-mean and unit variance. Stochastic gradient descent with ADAM optimization (Kingma & Ba, 2014) was used to learn the weights of the neural network. Dropout and max-norm regularization were used to control overfitting. Hidden size ($\{64, 128, 256, 512\}$), learning rate ($\left[10^{-3}, 10^{-1}\right]$), dropout ($[0.3, 0.7]$), max-norm ($[5, 20]$) and initialization ranges ($\left[10^{-3}, 10^{-1}\right]$) parameters were explored by training 25 models with random (uniform) hyperparameter initializations and the best was chosen according to validation performance. It has been reported that this strategy is preferable over grid search when training deep models (Bergstra & Bengio, 2012). All the implementation was carried on with the Blocks framework (Van Merriënboer et al., 2015)[5].

During the training process, we noticed that batch normalization considerably helped in terms of training time and convergence, resulting in less sensitivity to hyperparameters such as initialization ranges or learning rate. Also, dropout and max-norm regularization strategies helped to increase the performance at test time.

5 RESULTS

5.1 EVALUATION OVER SYNTHETIC DATA

In order to evaluate if the model is able to identify which modality is contributing more information to classify a particular sample, we created a synthetic task based on a generative model, which is depicted in Figure 6. In this model we define the random binary variable C as the target and $x_v, x_t \in \mathbb{R}^2$ as the input features. M is a random binary variable that decides which modality will contain the relevant information that determines the class. The input features of each modality can

[5] https://github.com/johnarevalo/gmu-mmimdb

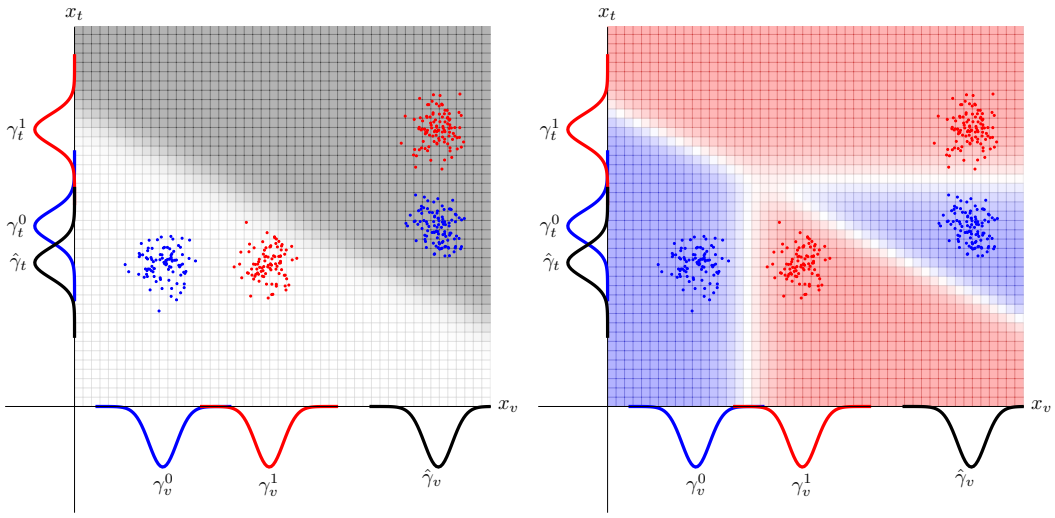


Figure 7: Activations of z (left) and prediction (right) for a synthetic experiment with $x_v, x_t \in \mathbb{R}^1$.

be generated by a random source, $\hat{y}_v$ and $\hat{y}_t$, or by an informed source, y_v and y_t. The generative model is specified as follows:

$$
\begin{aligned}
C &\sim \text{Bernoulli}(p_C) & x_v &= My_v + (1-M)\hat{y}_v \\
M &\sim \text{Bernoulli}(p_M) & y_t &\sim \mathcal{N}(\gamma_t^C) \\
y_v &\sim \mathcal{N}(\gamma_v^C) & \hat{y}_t &\sim \mathcal{N}(\hat{\gamma}_t) \\
\hat{y}_v &\sim \mathcal{N}(\hat{\gamma}_v) & x_t &= M\hat{y}_t + (1-M)y_t
\end{aligned}
$$

We trained a model with a single GMU and applied a sigmoid function over h, then the binary cross entropy was used as loss function. Using the generative model, 200 samples per class were generated for each experiment. 1000 synthetic experiments with different random seeds were run and the GMU outperformed a logistic regression classifier in 370 of them, while obtaining equal results in the remainder ones. Our goal in these simulations was to show that the model was able to learn a latent variable that determines which modality carries the useful information for the classification. An interesting result is that between M and the activations of the gate z there is a correlation of 1. This means the model was capable of learning such latent variable by only observing the x_v and x_t input features.

We also wanted to project back the z activations to the feature space in order to visualize regions depending on the modality. Figure 7 shows the activations in a synthetic experiment generated by the setup of Figure 6 for $x_v, x_t \in \mathbb{R}^1$. Each axis represents a modality, red and blue dots are the samples generated for the two classes and black Gaussian curves represent the $\hat{\gamma}_v$ and $\hat{\gamma}_t$ noises. The contour of the left figure (gray) represents the activation of z. Notice that in white regions ($z = 1$), the model gives more importance to the x_v modality while in gray regions ($z = 0$) the x_t modality is more relevant; i.e. the z gate is isolating the noise. The contour of the right figure (blue-red) represents the model prediction. It is noteworthy that the boundary defined by the gates still holds when the model solves the task. This also encourages the inclusion of non-linearities to the z gate so that it is able to discriminate more complex interactions between modalities.

5.2 Genre classification results

Before using our text representation in the multimodal task, we wanted to be sure such representation was good enough to address the genre classification task. Thus, we evaluated it on 2 public datasets. We found *MaxoutMLP_w2v* achieves the state of the art results on the *ki-04* dataset and increases the performance in the *7Genre* dataset from 0.841 to 0.854 (Kanaris & Stamatatos, 2009). Notice

that the baseline uses additional information from the HTML structure from the web page, while this representation uses only the text data.

Table 2: Summary of classification task on the MM-IMDb dataset

Modality	**Representation**	**F-Score**			
		weighted	samples	micro	macro
Multimodal	GMU	**0.617**	**0.630**	**0.630**	**0.541**
	Linear_sum	0.600	0.607	0.607	0.530
	Concatenate	0.597	0.605	0.606	0.521
	AVG_probs	0.604	0.616	0.615	0.491
	MoE_MaxoutMLP	0.592	0.593	0.601	0.516
	MoE_MaxoutMLP (tied)	0.579	0.579	0.587	0.489
	MoE_Logistic	0.541	0.557	0.565	0.456
	MoE_Logistic (tied)	0.483	0.507	0.518	0.358
Text	MaxoutMLP_w2v	0.588	0.592	0.595	0.488
	RNN_transfer	0.570	0.580	0.580	0.480
	MaxoutMLP_w2v_1_hidden	0.540	0.540	0.550	0.440
	Logistic_w2v	0.530	0.540	0.550	0.420
	MaxoutMLP_3grams	0.510	0.510	0.520	0.420
	Logistic_3grams	0.510	0.520	0.530	0.400
	RNN_end2end	0.490	0.490	0.490	0.370
Visual	VGG_Transfer	0.410	0.429	0.437	0.284
	CNN_end2end	0.370	0.350	0.340	0.210

Table 2 shows the results in the proposed dataset. For the textual modality, the best performance is obtained by the combination of word2vec representation with an MLP classifier. The behavior of all representation methods are consistent across the performance measures. Learning from scratch the RNN model performed the worst. We hypothesize this has to do with the lack of data to learn meaningful relations among words. It has been shown that millions of words are required to train a model such as word2vec that is able to exploit common regularities between word co-occurrences.

For the visual modality, the usage of pretrained models works better than training the model from scratch. It seems it is still a small dataset to learn all the complexities of the posters. Now, comparing the performance independently per genre, as in Table 3, it is interesting to notice that in *Animation* the visual modality outperforms the textual one.

In the multimodal scenario, by adding the GMU as building block to learn the fusion we obtained the best performance, improving independent modalities in the averaged measures and in 16 of out 23 genres and outperforming all other evaluated fusion strategies. The concatenation or the linear

Table 3: Macro F-Score reported per genre for single and multimodal approaches.

Genre	**Textual**	**Visual**	**GMU**	**Genre**	**Textual**	**Visual**	**GMU**
Drama	0.74	0.67	**0.77**	Fantasy	0.42	0.25	**0.46**
Comedy	0.65	0.59	**0.68**	Family	0.5	0.46	**0.58**
Romance	**0.53**	0.33	0.51	Biography	**0.4**	0.02	0.25
Thriller	0.57	0.39	**0.62**	War	0.57	0.19	**0.64**
Crime	**0.61**	0.25	0.59	History	**0.35**	0.06	0.29
Action	0.58	0.37	**0.6**	Animation	0.43	0.61	**0.68**
Adventure	**0.51**	0.32	**0.51**	Musical	0.14	0.18	**0.28**
Horror	0.65	0.41	**0.69**	Western	0.52	0.37	**0.65**
Documentary	0.67	0.18	**0.76**	Sport	0.64	0.11	**0.7**
Mystery	0.38	0.11	**0.39**	Short	0.2	0.24	**0.27**
Sci-Fi	0.63	0.3	**0.66**	Film-Noir	0.02	0.11	**0.37**
Music	**0.51**	0.01	0.48				

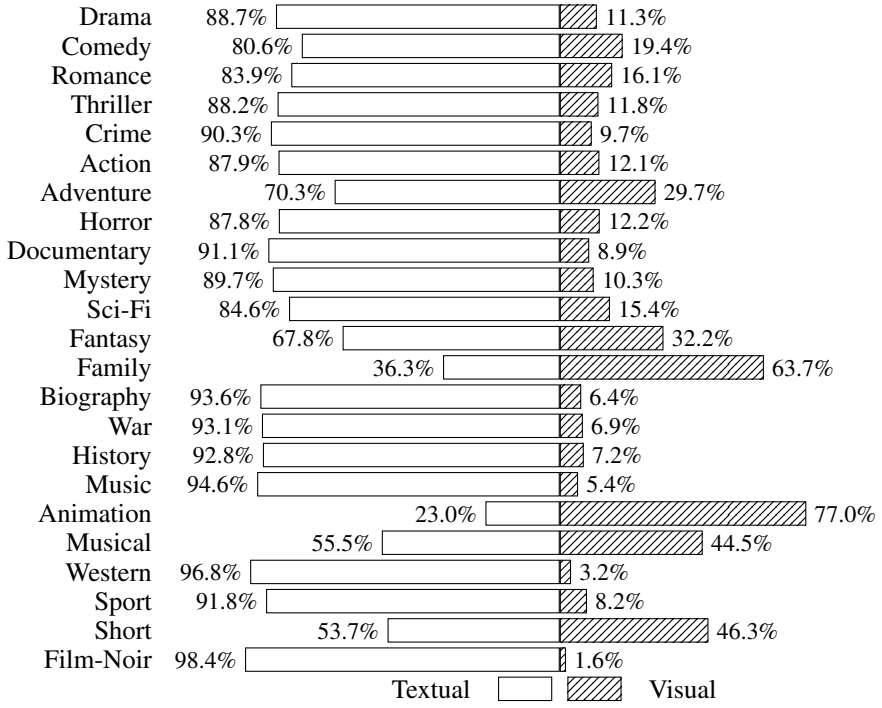


Figure 8: Percentage of gates activations ($z > 0.5$: Visual; $z <= 0.5$: textual) for test samples to which the model assigned them the label.

combination approaches were not enough to model the correlation between the modalities and MoE models did not perform better than simpler approaches. This is an expected behavior for MoE in a relatively small dataset because the data is fractionated over different experts, and thus it doesn't make an efficient use of the training samples.

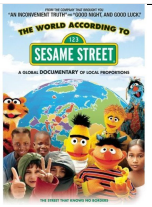
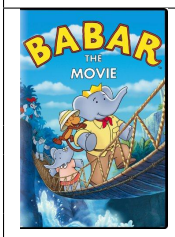
In order to evaluate which modality influences more the model when assigning a particular label, we averaged the activations of a subset of z gates of the test samples to which the model assigned them such label. We counted the number of samples that pays more attention to the textual modality ($z <= 0.5$) or to the visual modality ($z > 0.5$). The units were chosen taking into account the mutual information between the predictions and the z activations. The result of this analysis is depicted in Figure 8. As expected, the model is generally more influenced by the textual modality. But, in particular cases such as *Animation* or *Family* genres, the visual modality affects more the model. This is also consistent with results of Table 3 which reports better performances for visual modality.

We wanted to qualitative explore test examples in which performance was improved by a relative large margin. Table 4 illustrates cases where the model takes advantage of the most accurate modality, and in some cases removes false positives. It is noteworthy that some of these examples can be confusing for a human if one modality is missing, or additional context information is not given.

6 Conclusions

This work presented a strategy to learn fusion transformations from multimodal sources. Similarly to the way recurrent models control the information flow, the proposed model is based on multiplicative gates. The Gated Multimodal Unit (GMU) receives two or more input sources and learns to determine how much each input modality affects the unit activation. In synthetic experiments the GMU was able to learn hidden latent variables, and in a real scenario it outperformed the single-modality approaches. An interesting property of GMU is that, being a differentiable operation, it

Table 4: Examples of predictions in test set. Red and blue genres are false positives and true positives respectively.

The World According to Sesame Street		
	a documentary which examines the creation and co - production of the popular children ' s television program in three developing countries: bangladesh , kosovo and south africa .	
	Ground Truth	Documentary
	Textual	Documentary, History
	Visual	Comedy, Adventure, Family, Animation
	GMU	Documentary
Babar: the movie		
	in his spectacular film debut , young babar , king of the elephants , must save his homeland from certain destruction by rataxes and his band of invading rhinos .	
	Ground Truth	Adventure, Fantasy, Family, Animation, Musical
	Textual	Adventure, Documentary, War, Music
	Visual	Comedy, Adventure, Family, Animation
	GMU	Adventure, Family, Animation
Letters from Iwo Jima		
	the island of iwo jima stands between the american military force and the home islands of japan . (...) when the american invasion begins , both kuribayashi and saigo find strength , honor , courage , and horrors beyond imagination .	
	Ground Truth	Drama, War, History
	Textual	Drama, Action, War, History
	Visual	Thriller, Action, Adventure, Sci-Fi
	GMU	Drama, War, History
The Last Elvis		
	a tragic accident causes an elvis impersonator to reassess his priorities.	
	Ground Truth	Drama
	Textual	Comedy, Documentary, Family, Biography, Music
	Visual	Drama, Romance
	GMU	Drama

is easily coupled in any other neural network architecture and trained with standard gradient-based optimization algorithms. With this work we will also release a new dataset that contains around $27,000$ movie plots, images and other metadata. To the best of our knowledge, this is the biggest dataset used to perform movie genre classification based on multimodal information and the first one to be publicly available. In our future work we expect to explore deep architectures of GMU layers as well as integration with attention mechanism over the input modalities. Also, It will be interesting to explore in more depth the interpretability of the learned features.

ACKNOWLEDGMENTS

Arevalo thanks Colciencias for its support through a doctoral grant in call 617/2013. The authors also thank for K40 Tesla GPU donated by NVIDIA and which was used for some representation learning experiments.

REFERENCES

Zeynep Akata, Honglak Lee, and Bernt Schiele. Zero-Shot Learning with Structured Embeddings. *CoRR*, abs/1409.8, 2014. URL http://arxiv.org/abs/1409.8403.

Deepa Anand. Evaluating folksonomy information sources for genre prediction. In *Advance Computing Conference (IACC), 2014 IEEE International*, pp. 887–892, feb 2014. doi: 10.1109/IAdCC.2014.6779440.

Galen Andrew, Raman Arora, Jeff A Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *ICML (3)*, pp. 1247–1255, 2013.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *International Conference on Computer Vision (ICCV)*, 2015.

Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):345–379, April 2010. ISSN 0942-4962. doi: 10.1007/s00530-010-0182-0.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155, 2003.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012.

Chidansh Bhatt and Mohan Kankanhalli. Multimedia data mining: state of the art and challenges. *Multimedia Tools and Applications*, 51(1):35–76, 2011. ISSN 1380-7501. doi: 10.1007/s11042-010-0645-5.

Adam Coates and Andrew Y Ng. The importance of encoding versus training with sparse coding and vector quantization. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 921–928, 2011.

Li Deng. A tutorial survey of architectures, algorithms, and applications for deep learning. *APSIPA Transactions on Signal and Information Processing*, 3, 2014. ISSN 2048-7703. doi: 10.1017/atsip.2013.9. URL http://journals.cambridge.org/article{_}S2048770313000097.

Fangxiang Feng, Ruifan Li, and Xiaojie Wang. Constructing hierarchical image-tags bimodal representations for word tags alternative choice. *arXiv preprint arXiv:1307.1275*, 2013.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc\textquotesingle Aurelio Ranzato, and Tomas Mikolov. DeViSE: A Deep Visual-Semantic Embedding Model. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger (eds.), *Advances in Neural Information Processing Systems 26*, pp. 2121–2129. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5204-devise-a-deep-visual-semantic-embedding-model.pdf.

Zhikang Fu, Bing Li, Jun Li, and Shuhua Wei. Fast Film Genres Classification Combining Poster and Synopsis. In Xiaofei He, Xinbo Gao, Yanning Zhang, Zhi-Hua Zhou, Zhi-Yong Liu, Baochuan Fu, Fuyuan Hu, and Zhancheng Zhang (eds.), *Lecture Notes in Computer Science*, volume 9242 of *Lecture Notes in Computer Science*, pp. 72–81. Springer International Publishing, Cham, 2015. doi: 10.1007/978-3-319-23989-7_8. URL http://link.springer.com/10.1007/978-3-319-23862-3http://link.springer.com/10.1007/978-3-319-23989-7{_}8.

Ian Goodfellow, David Warde-farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Maxout networks. In Sanjoy Dasgupta and David Mcallester (eds.), *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pp. 1319–1327. JMLR Workshop and Conference Proceedings, May 2013.

Hao-Zhi Hong and Jen-Ing G Hwang. Multimodal PLSA for Movie Genre Classification. In Friedhelm Schwenker, Fabio Roli, and Josef Kittler (eds.), *Multiple Classifier Systems: 12th International Workshop, MCS 2015, G{ü}nzburg, Germany, June 29 - July 1, 2015, Proceedings*, pp. 159–167. Springer International Publishing, Cham, 2015. ISBN 978-3-319-20248-8. doi: 10.1007/978-3-319-20248-8_14. URL http://dx.doi.org/10.1007/978-3-319-20248-8{_}14.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pp. 873–882. Association for Computational Linguistics, 2012.

Hui-Yu Huang, Weir-Sheng Shih, and Wen-Hsing Hsu. A Film Classifier Based on Low-level Visual Features. In *2007 IEEE 9th Workshop on Multimedia Signal Processing*, volume 3, pp. 465–468. IEEE, 2007. ISBN 978-1-4244-1273-0. doi: 10.1109/MMSP.2007.4412917. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4412917.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of The 32nd International Conference on Machine Learning*, pp. 448–456, 2015.

Marina Ivasic-Kos, Miran Pobar, and Luka Mikec. Movie posters classification into genres based on low-level features. In *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, volume i, pp. 1198–1203. IEEE, may 2014. ISBN 978-953-233-077-9. doi: 10.1109/MIPRO.2014.6859750. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6859750.

Marina Ivasic-Kos, Miran Pobar, and Ivo Ipsic. Automatic Movie Posters Classification into Genres. In Madevska Ana Bogdanova and Dejan Gjorgjevikj (eds.), *ICT Innovations 2014: World of Data*, pp. 319–328. Springer International Publishing, Cham, 2015. ISBN 978-3-319-09879-1. doi: 10.1007/978-3-319-09879-1_32. URL http://dx.doi.org/10.1007/978-3-319-09879-1{_}32.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. *arXiv preprint arXiv:1511.07571*, 2015.

Ioannis Kanaris and Efstathios Stamatatos. Learning to recognize webpage genres. *Information Processing and Management*, 45(5):499–512, 2009. ISSN 03064573. doi: 10.1016/j.ipm.2009.05.003. URL http://dx.doi.org/10.1016/j.ipm.2009.05.003.

Yoonseop Kang, Saehoon Kim, and Seungjin Choi. Deep learning to hash with multiple representations. In *2012 IEEE 12th International Conference on Data Mining*, pp. 930–935. IEEE, 2012.

Douwe Kiela and Léon Bottou. Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-14)*, 2014.

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

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Multimodal neural language models. In *ICML*, volume 14, pp. 595–603, 2014a.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *arXiv preprint arXiv:1411.2539*, 2014b.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. doi: 10.1038/nature14539. URL http://dx.doi.org/10.1038/nature14539.

Dong Yu Li Deng. *Deep Learning: Methods and Applications*. NOW Publishers, May 2014. URL https://www.microsoft.com/en-us/research/publication/deep-learning-methods-and-applications/.

Xinyan Lu, Fei Wu, Xi Li, Yin Zhang, Weiming Lu, Donghui Wang, and Yueting Zhuang. Learning multimodal neural network with ranking examples. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 985–988. ACM, 2014.

Gjorgji Madjarov, Dragi Kocev, Dejan Gjorgjevikj, and Sašo Džeroski. An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, 45(9):3084–3104, 2012. ISSN 0031-3203. doi: http://dx.doi.org/10.1016/j.patcog.2012.03.004. URL http://www.sciencedirect.com/science/article/pii/S0031320312001203.

Eric Makita and Artem Lenskiy. A movie genre prediction based on Multivariate Bernoulli model and genre correlations. (May), mar 2016a. URL http://arxiv.org/abs/1604.08608.

Eric Makita and Artem Lenskiy. A multinomial probabilistic model for movie genre predictions. 2016b. URL http://arxiv.org/abs/1603.07849.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pp. 3111–3119, 2013b.

J Ngiam, A Khosla, and M Kim. Multimodal Deep Learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 689—-696, 2011. URL http://ai.stanford.edu/{~}ang/papers/icml11-MultimodalDeepLearning.pdf.

Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeff Dean. Zero-Shot Learning by Convex Combination of Semantic Embeddings. *CoRR*, abs/1312.5, dec 2014. URL http://arxiv.org/abs/1312.5650.

Gregory Pais, Patrick Lambert, Daniel Beauchene, Francoise Deloule, and Bogdan Ionescu. Animated movie genre detection using symbolic fusion of text and image descriptors. In *2012 10th International Workshop on Content-Based Multimedia Indexing (CBMI)*, number 1, pp. 1–6. IEEE, jun 2012. ISBN 978-1-4673-2369-7. doi: 10.1109/CBMI.2012.6269813. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6269813.

Deli Pei, Huaping Liu, Yulong Liu, and Fuchun Sun. Unsupervised multimodal feature learning for semantic image segmentation. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6. IEEE, aug 2013. ISBN 978-1-4673-6129-3. doi: 10.1109/IJCNN.2013.6706748. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.

Dharak Shah, Saheb Motiani, and Vishrut Patel. Movie Classification Using k-Means and Hierarchical Clustering. Technical report, 2013.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-Shot Learning Through Cross-Modal Transfer. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger (eds.), *Advances in Neural Information Processing Systems 26*, pp. 935–943. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5027-zero-shot-learning-through-cross-modal-transfer.pdf.

Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded Compositional Semantics for Finding and Describing Images with Sentences. *Transactions of the Association for Computational Linguistics (TACL)*, 2(April):207–218, 2014. URL http://nlp.stanford.edu/{~}socherr/SocherLeManningNg{_}nipsDeepWorkshop2013.pdf.

Nitish Srivastava and Ruslan Salakhutdinov. Multimodal Learning with Deep Boltzmann Machines. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 25*, pp. 2222–2230. Curran Associates, Inc., 2012. URL http://papers.nips.cc/paper/4683-multimodal-learning-with-deep-boltzmann-machines.pdf.

Heung Il Suk and Dinggang Shen. Deep learning-based feature representation for AD/MCI classification. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8150 LNCS, pp. 583–590, 2013. ISBN 9783642407628. doi: 10.1007/978-3-642-40763-5_72.

Jian Tu, Zuxuan Wu, Qi Dai, Yu-Gang Jiang, and Xiangyang Xue. Challenge Huawei challenge: Fusing multimodal features with deep neural networks for Mobile Video Annotation. In *Multimedia and Expo Workshops (ICMEW), 2014 IEEE International Conference on*, pp. 1–6, 2014. doi: 10.1109/ICMEW.2014.6890609.

Bart Van Merriënboer, Dzmitry Bahdanau, Vincent Dumoulin, Dmitriy Serdyuk, David Warde-Farley, Jan Chorowski, and Yoshua Bengio. Blocks and fuel: Frameworks for deep learning. *arXiv preprint arXiv:1506.00619*, 2015.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156–3164, 2015.

Pengcheng Wu, Steven C.H. Hoi, Hao Xia, Peilin Zhao, Dayong Wang, and Chunyan Miao. Online multimodal deep similarity learning with application to image retrieval. In *Proceedings of the 21st ACM international conference on Multimedia - MM '13*, MM '13, pp. 153–162, New York, New York, USA, 2013. ACM Press. ISBN 9781450324045. doi: 10.1145/2502081.2502112. URL http://doi.acm.org/10.1145/2502081.2502112{%}5Cnhttp://dl.acm.org/citation.cfm?doid=2502081.2502112.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2(3):5, 2015.

Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, 2012.

Yin Zheng, YJ Zhang, and Hugo Larochelle. Topic Modeling of Multimodal Data: an Autoregressive Approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. ISBN 2011000211. URL http://www.dmi.usherb.ca/{~}larocheh/publications/ZhengY2014.pdf.