

# Multimodal Classification Fusion in Real-World Scenarios

Ignazio Gallo, Alessandro Calefati, Shah Nawaz  
University of Insubria  
Varese, Italy  
Email: ignazio.gallo@uninsubria.it

**Abstract**—In this paper, we propose a multimodal setting in real-world scenarios based on weighting and meta-learning combination methods that integrate the output probabilities obtained from text and visual classifiers. While the classifier built on the concatenation of text and visual features may worsen the results, the model described in this paper can increase classification accuracy to over 6%. Typically, text or images are used in classification; however, ambiguity in either text or image may reduce the performance. This leads to combine text and image of an object or a concept in a multimodal approach to enhance the performance. In our approach, a text classifier is trained on Bag of Words and a visual classifier is trained on features extracted through a Deep Convolutional Neural Network. We created a new dataset of real-world texts and images called Ferramenta. Some of the images and related texts in this dataset contain ambiguities, which is an ideal situation to test a multimodal approach. Experimental results reported on Ferramenta and PASCAL VOC2007 datasets indicate that the combination methods described performs better in a multimodal setting.

**Index Terms**—multimodal classification; deep convolutional neural network;

## I. INTRODUCTION

With the rise of e-commerce websites, users are provided information often coming from different sources, for example text and image. For each item on sale, a user can select a product based on a text and an image that show characteristics, colors and other features of the product. However, sometimes, the image and the text of an advertisement are not consistent, which confuses the users that are interested in buying that product. We use different kind of data to perform a multimodal classification, a technique that leverages on features extracted from different modalities to enhance the classification performance. The proposed approach is summarized in Figure 2 and uses Convolutional Neural Network (CNN) [1] and other classifiers to achieve the above mentioned goal. This method can obtain high classification accuracy, especially on data characterized by noisy text (grammatically ill formed sentences, short text document, technical details, etc.). Experiments are conducted on advertisements, as shown in Figure 1, where the description contains a noisy text and an ambiguous image in some cases.

The image and the text of a document usually contain information describing the same object or concept. In ambiguous situations it is useful to extract the information content from the text and image. For example, the image and the text in



(a) Extol craft 108811 - Scissors / shears - Segmental diamond cut blade 115 x 22,2 mm Dry cooling Weight: 0,12



(b) Fumasi Shear Sheet metal Italy 220 - 8033116531634 - Model Italy Gambi rights Lame execution burnished



(c) Finether 3.2M Portable Aluminum Telescoping Ladder with Finger Protection Spacers for Home Loft Office, EN131 Certified, 330 Lb Capacity.



(d) Custom multifunction dynamic construction scaffolding (11'6 x 4' x 2'6 Base), simple for decoration -up 150 kg -weights only 16 kg

Fig. 1: In the top row, two examples of ambiguous textual descriptions that can be disambiguated through the analysis of the respective images. In the bottom row, two examples of ambiguous images that can be disambiguated through the analysis of the respective description.

Figures 1b and 1c describe a pair of shears and a ladder respectively, without ambiguities. But if we look at Figure 1a and we want to classify it only using the text, a classifier may incorrectly classify it as “shears”. Conversely, if we look at the example in Figure 1d and we want to classify that advertisement only analyzing the image, a classifier may incorrectly annotate it as a “ladder”. In this way, by combining text and image it is possible to disambiguate wrong classifications and improve the classification result. This leads to the use of a multimodal approach using textual and visual features on a variety of tasks including modeling semantic relatedness, compositionality and classification [2], [3], [4], [5], [6].

In this work, we present two late fusion [7] mechanisms,

weighting and meta combination methods that combine the output of two individual classifiers trained on visual and text features respectively to classify an advertisement. The performance of the two above mentioned mechanisms outperforms an early fusion [7] classifier trained on text and visual features concatenation. Visual features are obtained by using CNN extracted features whereas text features are obtained using Doc2Vec (D2V) from [8] and BoW [7]. We also created a dataset called “*Ferramenta*”, which contains ambiguous images and noisy text descriptions of commercial offers. Existing datasets in literature (such as [9]) are mainly characterized by a couple of labels or keywords associated to an image representing a concept. Our dataset provides images and descriptions representing adverts, that are usually available on an e-commerce website. For the purposes of academic research, we will publish our dataset and we believe that it can be used for a variety of useful tasks. Table IV shows some of the examples with our fusion method on Ferramenta dataset. Examples (a) - (c) show that the method correctly classifies an advert even if one of the models or both make wrong classifications.

## II. RELATED WORK

In literature it is uncommon to find a dataset with both text and image such as the one presented in this work, which is created through the combined use of text and image in an advertisement. The majority of the datasets available in literature are related to datasets of images that are then associated to labels to force multimodality, such as PASCAL VOC2007 [9] extended with Flickr tags<sup>1</sup>.

The fusion of different modalities generally occurs at two levels: at the level of *features* or *early fusion* and at the level of *decisions* or *late fusion* as described in [7]. Some examples of early fusion such as [10], [6], directly concatenate text and image features to produce a single multimodal vector (see graphical representation in Figure 3), obtaining promising performance in other contexts other than classification. Thus, we can show that in certain contexts an early fusion approach results in a classification performance that is better than text or image classifiers: however, it never outperforms better classifiers. We can now explore and investigate the late fusion strategy.

## III. THE PROPOSED MODEL

The purpose of supervised learning is to categorize patterns into a set of classes. The main idea behind the ensemble methodology is to weigh several individual classifiers and combine them to obtain a classifier that outperforms individual classifiers, also called late fusion [7] in a multimodal approach. Empirically, ensembles tend to yield better results when there is a significant diversity among models [11]. Many ensemble methods, therefore, seek to promote diversity among the models they combine. In the ensemble fusion model, texts and images are first processed separately to provide decision-level

<sup>1</sup><http://lear.inrialpes.fr/people/guillaumin/data.php>

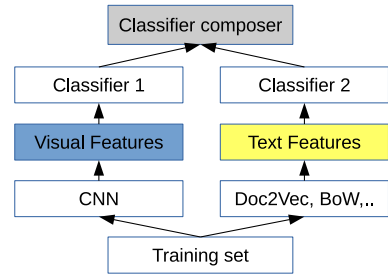


Fig. 2: The late fusion model. Text and images are used independently to extract the features in a supervised manner or by specific operators. In each of these two types of features, a classifier is trained to output class probabilities. The latter are fused together by special algorithms.

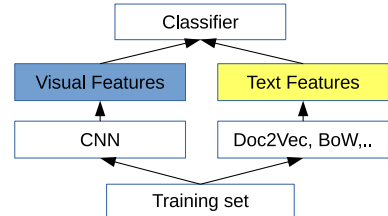


Fig. 3: A classical early fusion [7] multimodal approach where texts and images are used independently to extract the features in supervised manner or by special operators. These two types of features are concatenated together in order to train a single output classifier.

results, as described in [12], [7]. Results are then combined using two different approaches: weighting methods and meta-learning methods [11]. Weighting methods are useful if the base-classifiers perform the same task and have comparable success. Meta-learning methods are best suited for cases in which certain classifiers consistently correctly classify, or consistently mis-classify, certain instances.

In our model, as in Figure 2, text processing and image processing are carried out on text and images separately and a fusion algorithm is used to combine the results. The details of the model and different ensemble approaches are explained below.

### A. Weighting methods

When combining classifiers with weights, a classifier’s classification has a strength that is proportional to an assigned weight. This weight can be fixed or dynamically determined for the specific instance to be classified. Suppose we are using probabilistic classifiers, where  $P(y = c|x)$  denotes the probability of class  $c$  given an instance  $x$ . The idea of the *Distribution Summation* (DS) combining method [11] is to sum up the conditional probability vector obtained from each classifier. The selected class is chosen according to the highest value in the total vector. Formally, it can be written

$$class(x) = \arg \max_{c_i \in dom(y)} (P_t + P_v) \quad (1)$$

where  $P_t$  and  $P_v$  are the probabilities  $P_t(y = c_i|x)$  and  $P_v(y = c_i|x)$  of the text classifier and visual classifier respectively.

The weights  $\alpha_t$  and  $\alpha_v$  of each classifier can be set proportional to its accuracy performance on the training set or validation set, obtaining the following *Performance Weighting* (PW) formula

$$\text{class}(x) = \arg \max_{c_i \in \text{dom}(y)} (\alpha_t P_t + \alpha_v P_v) \quad (2)$$

where each  $\alpha_k$  denotes the weight of the classifier, such that  $\alpha_k \geq 0$  and  $\sum \alpha_k = 1$ .

According to the *Logarithmic Opinion Pool* (LOP) defined in [11], the selection of the preferred class can be also performed in this way:

$$\text{class}(x) = \arg \max_{c_i \in \text{dom}(y)} e^{\alpha_t \log P_t + \alpha_v \log P_v} \quad (3)$$

In this paper we used the equation 3 as weighting method, but many other methods can be used, as described in [11].

### B. Meta-combination methods

Meta-learning means learning from the classifiers produced by the inducers and from the classifications of these classifiers on training data.

In this work we tested the *Stacking* (S) meta-combination method. Stacking is a technique for achieving the highest generalization accuracy [11]. First, two algorithms are trained on images and text descriptions using the available data, then a combiner algorithm is trained to make a final prediction using all the predictions of the other algorithms as additional inputs. If an arbitrary combiner algorithm is used, then stacking can theoretically represent any of the ensemble techniques described above. Stacking performance can be improved by using output probabilities for each class label from the base-level classifiers. Each training instance  $i$  of a stacking meta-combiner, consists of a first set of  $n$  probabilities  $P_{1,t}(y = c_1|x) \dots P_{n,t}(y = c_n|x)$ , computed by the model used for the text classification, and concatenated to a second set of  $n$  probabilities from the visual model  $P_{1,v}(y = c_1|x) \dots P_{n,v}(y = c_n|x)$ . All these input probabilities are associated to the same set of binary outcome variables  $y_1 \dots y_n$ .

In this work, we experimented with two Stacking combiner algorithms: a simple Logistic regression (S-L) model with a ridge estimator and a Multilayer Perceptron Classifier (S-MLP) that uses back-propagation to classify instances [11]. These two algorithms are available in the Weka open-source library [13].

## IV. EXPERIMENTS AND DISCUSSIONS

We evaluated our text and visual late fusion classification method on two different datasets. The purpose of these experiments is twofold:

- to identify the best configuration of our method for a real-world scenario;



Fig. 4: The *Ferramenta* dataset. Each image is representative of one of the 52 classes in the dataset.

TABLE I: Comparison of the overall accuracy computed on the Ferramenta test set. The first row contains the results of the various models trained on the features extracted from the CNN. The next four rows report the accuracy of the models trained on two different configurations of BoW and D2V. The last two rows show the results of the model shown in Figure 3.

Features	#attr.	SVM(%)	RF(%)	DT(%)
CNN	4096	<b>90.31</b>	<b>88.01</b>	<b>79.27</b>
BoW-500	500	80.70	91.00	89.38
BoW-1000	1000	76.05	<b>91.73</b>	<b>90.58</b>
D2V win2	100	83.88	86.39	63.68
D2V win10	100	<b>88.08</b>	87.38	66.17
CNN + BoW	5096	89.51	<b>88.12</b>	<b>89.60</b>
CNN + D2V	4196	<b>89.95</b>	87.16	79.53

- to show that the multimodal method here presented can be more accurate in classification than two individual classifiers trained on noisy text and images respectively;

In all the experiments conducted in this work, we used a CNN proposed by [1] known as AlexNet. Instead of using the model as a classifier as usual, we use it as a feature extractor; in fact, we feed an image and obtained the 4096-dimensional vector of the last fully connected layer. For the training and test phases, we resized images to  $256 \times 256$  to fit the input of the CNN.

In our experiments we measured the performance of the models mainly by using the overall accuracy (Acc), but in some experiments that required a more in-depth analysis we used also the precision (P), recall (R) and F-measure (F1) criterion. In other experiments we used the area under the ROC and PRC curves to better understand the behavior of the model.

### A. Datasets

In our experiments we use a real-world dataset that we called *Ferramenta* and the multimodal version of the PASCAL VOC2007 [9] dataset. By using text features only, the classification works well in scenarios when the text is represented by a set of labels describing the image, such as in the multimodal version of the PASCAL VOC2007. When

TABLE II: Multimodal fusion accuracy results with the Ferramenta test set. In the top half of the table the results of the proposed late fusion model with the best classifiers for each feature. The bottom half shows the results of the best visual model combined with the best textual model, based on the D2V features. The two rows *FConc* (Feature Concatenation) show the results of the early fusion model shown in Figure 3, whereas the other rows labeled as *Fusion* show the results of the model shown in Figure 2. *Prob* means that the final combiner receives in input the probabilities of the two underlying models.

	features	algorithm	P(%)	R(%)	F1(%)	ROC(%)	PRC(%)	Acc(%)
Text	bow1000	RF	91.80	91.70	91.70	99.10	94.60	91.73
Visual	CNN4096	SVM	90.60	90.30	90.30	99.40	92.80	90.31
FConc.	CNN+BoW	RF	89.90	88.10	88.00	99.20	93.50	88.14
Fusion	prob	DS	93.80	93.70	93.70	<b>99.80</b>	<b>97.50</b>	93.74
Fusion	prob	PW	93.80	93.70	93.70	<b>99.80</b>	<b>97.50</b>	93.74
Fusion	prob	LOP	<b>94.60</b>	<b>94.40</b>	<b>94.40</b>	99.50	97.30	<b>94.42</b>
Fusion	prob	S-L	90.40	89.50	89.50	94.40	81.40	89.53
Fusion	prob	S-MLP	93.40	93.20	93.20	99.30	96.00	93.21
Text	D2V	SVM	88.20	87.40	87.10	98.80	92.00	87.38
Visual	CNN4096	SVM	90.60	90.30	90.30	99.40	92.80	90.31
FConc.	CNN+D2V	SVM	90.40	89.50	89.50	94.40	81.40	89.53
Fusion	prob	DS	92.50	92.40	92.30	<b>99.70</b>	<b>96.70</b>	92.37
Fusion	prob	PW	92.50	92.40	92.30	<b>99.70</b>	<b>96.70</b>	92.38
Fusion	prob	LOP	<b>93.20</b>	<b>92.90</b>	<b>92.90</b>	99.40	96.50	<b>92.94</b>
Fusion	prob	S-L	91.80	91.60	91.40	98.40	88.70	91.57
Fusion	prob	S-MLP	92.00	92.00	91.90	99.50	95.30	92.03

TABLE III: Performance values of the experiment on PASCAL VOC2007 dataset. The experiment on the single modal approach was done using classifiers that performed better on the Ferramenta dataset, as shown in Table I. RandomForest was used for the text and SVM was used for the visual features. As expected, it emerges that best results are obtained using only textual features because the text describes images very well, also for images that contain objects representing the class in a small size or not in a central position. The approach that combines textual and visual features does not perform well on this dataset.

	airplane			bicycle			bird			boat			bottle		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Text	<b>95.30</b>	<b>70.10</b>	<b>80.80</b>	<b>80.60</b>	<b>46.90</b>	<b>59.30</b>	<b>94.50</b>	<b>55.00</b>	<b>69.50</b>	<b>80.70</b>	<b>39.00</b>	<b>52.50</b>	<b>44.70</b>	16.00	<b>23.60</b>
Visual	32.60	54.90	40.90	25.70	23.80	24.70	25.80	8.20	12.40	25.20	31.40	28.00	8.10	<b>34.90</b>	13.10
T+V	51.20	64.70	57.10	44.30	32.60	37.60	64.50	25.20	36.20	57.30	32.00	41.00	10.50	33.00	15.90
	bus			car			cat			chair			cow		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Text	<b>89.60</b>	<b>39.70</b>	<b>55.00</b>	<b>80.60</b>	41.50	<b>54.80</b>	<b>92.90</b>	<b>60.90</b>	<b>73.50</b>	<b>7.90</b>	<b>91.60</b>	<b>14.60</b>	<b>78.50</b>	<b>40.20</b>	<b>53.10</b>
Visual	9.20	13.20	10.80	36.50	47.40	41.20	20.40	33.50	25.40	6.80	69.30	12.50	16.70	7.10	9.90
T+V	12.90	12.60	12.80	48.50	<b>49.50</b>	49.00	42.40	44.70	43.50	7.10	77.20	13.10	38.50	11.80	18.10
	diningtable			dog			horse			motorbike			person		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Text	27.30	3.20	5.70	<b>88.10</b>	<b>53.80</b>	<b>66.90</b>	<b>91.30</b>	<b>69.00</b>	<b>78.60</b>	<b>86.70</b>	<b>52.70</b>	<b>65.50</b>	26.20	9.70	14.20
Visual	33.30	<b>5.80</b>	<b>9.90</b>	14.80	23.70	18.20	46.50	41.20	43.70	30.40	34.20	32.20	<b>46.30</b>	<b>32.40</b>	<b>38.10</b>
T+V	<b>38.50</b>	5.30	9.30	35.10	36.60	35.80	68.90	64.60	66.70	44.50	40.10	42.20	39.30	26.10	31.40
	pottedplant			sheep			sofa			train			tvmonitor		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Text	<b>58.60</b>	22.80	<b>32.80</b>	<b>80.40</b>	<b>42.30</b>	<b>55.40</b>	<b>50.00</b>	10.80	17.70	<b>88.50</b>	<b>74.10</b>	<b>80.70</b>	<b>67.00</b>	25.80	<b>37.20</b>
Visual	9.90	<b>24.10</b>	14.00	20.80	5.20	8.30	15.50	<b>24.70</b>	19.10	24.40	55.60	33.90	13.80	<b>26.60</b>	18.20
T+V	13.70	23.70	17.40	50.00	14.40	22.40	20.90	22.90	<b>21.80</b>	38.90	64.90	48.60	18.80	24.90	21.40



(a) **Tags:** brasil, japan, tour. **Class:** car, person.



(b) **Tags:** club, racing, water, yacht. **Class:** boat.

Fig. 5: Some examples of images extracted from PASCAL VOC2007 dataset supplied with tags from Flickr dataset.

merging text and image classifiers, the final classification can significantly improve; except in case of noisy text, such as in our proposed multimodal dataset. In our experiment on PASCAL VOC2007, for example, the proposed multimodal approach produces worse classification results. We chose the PASCAL VOC2007 dataset and proposed Ferramenta dataset to highlight this situation.

Ferramenta dataset consists of 88.010 images split in 66.141 images for train and 21.869 images for test, belonging to 52 classes (paint brush, hinge, tape, safe, cart, etc.). Text descriptions in Ferramenta dataset contain 22045 different words for the train set and 20083 for the test set, all randomly

selected. Ferramenta dataset was collected from different sellers available in a price comparison website. The ground truth was created using a query based software that clusters commercial offers based on a text matching system. After each query, three co-located human annotators, as described in [14], analyzed the intra-class image similarity and exploiting the text to resolve ambiguity. We are aware that the first version of Ferramenta dataset contains few false positive offers; however, we will remove noise in the updated version by following the above method. Figure 4 shows some images from the Ferramenta dataset, one for each class of the dataset. Two examples of images and texts are shown in Figures 1c and 1d. Text in this dataset is in Italian language and we preferred not to translate the text into English, as we believe the translation process could alter the nature of the dataset. We have translated only the text in the examples in this article into English (Figure 1 and Table IV) to allow readers to understand the content of the dataset. Each image of this dataset has  $100 \times 100$  pixels. Unlike the datasets in the literature where modalities (text and image) are obtained from different sources like in the work [6] or images are labeled by different users over the Internet, i.e. ESP Game dataset [15], Ferramenta dataset provides a text and a representative image of commercial advertisements. We believe that this dataset can be used in future multimodal research work and on a variety of interesting tasks merging computer vision and natural language processing. Each image in the dataset represents an advert and comes with a unique identifier which is used to get the corresponding description.

The PASCAL VOC2007 has 20 different object categories (boat, bicycle, horse, etc.) with 9,963 images. For the PASCAL VOC2007 set we used the standard train and test split. We used a publicly available dataset obtained from Flickr tags for all the PASCAL VOC2007 images that can be downloaded from the Lear web site<sup>2</sup>. Each image of the dataset may have multiple objects from multiple classes in the same image, for this reason we used the dataset to classify each class against all the others. The Figure 5 shows two examples of image and text pairs typical of this dataset. The text associated with each image is a set of tags describing the image content.

### B. Model Configuration

Prior to the evaluation of our multimodal approach, we compared classification models on the individual visual and textual features to identify the best model for each type of feature and find the best configuration of features for each model. We conducted experiments with different configurations applied to BoW and D2V algorithms; however, the results here reported are obtained with default parameters provided by the Weka library. The best results of this comparison are summarized in the Table I. The analysis of this table shows that the best model built on visual features is a Support Vector Machine (SVM), which is even better than the same CNN used as a classifier, with which we get an accuracy of 88.64%. While the best model for the text features is the Random Forest

(RF) model applied to the 1000 BoW most significant features (the feature selection was done using the InfoGain algorithm, selecting the first 1000 best features). From the last two rows of Table I we can observe that the accuracy of the model shown in Figure 3 never exceeds the best accuracy rates of the same model applied to only one of the two features (text and image).

From the results obtained in Table I we chose to work with two different configurations of features. The first configuration is the best result for the visual features and the textual features: we created a multimodal model using a SVM for the visual features and a Random Forest for the BoW with the first 1000 most significant features. The second configuration uses the D2V textual features, in order to compare these with the BoW type. Using these configurations we compared the different fusion approaches described in sections III-A and III-B, and reported in Table II. By analyzing this second table, the best fusion approach appears to be the LOP.

In Table IV we reported some very interesting examples of how the best configuration of the proposed model works. The examples (a) - (c) display three different situations in which one or both of the models built on features were wrong, while the final combiner found the correct class by analyzing the probability values. Conversely, the examples (e) - (f) highlight two situations in which the error of one of the two classifiers leads to a wrong final classification. It never occurred that correct classifications of text and images lead to wrong final classification.

### C. Comparison







In this experiment, we used a standard multimodal datasets available for, which however has text information that was very different from those of the Ferramenta dataset.

For the images of this experiment, we used the PASCAL VOC2007 to train the CNN. The text from the dataset was represented as a BoW vector of size 804, that is the full size of the dictionary. Using the dictionary provided with the dataset, we were able to rebuild the original text for each image and the description of each image, in many cases, was composed by only two or three words. We discarded images that did not have any text description. The vectors from the described steps were applied to the LOP fusion approach that received the input probabilities from a Random Forest and a SVM. This is the best configuration from the previous experiment. As shown in Table III, best results are highlighted, the multimodal approach does not improve the classification performance. In most cases, best results are obtained using only text. This happens because in this dataset, many images belonging to a specific class, contain the main object in a small size or not in a central position.

Comparing this result with the one obtained from the Ferramenta dataset we can conclude that the proposed model works well only when it is used to merge two types of features computed on noisy and ambiguous data, whereas when the text is clear and unambiguous, a classifier trained on textual features performs better.

<sup>2</sup><http://lear.inrialpes.fr/people/guillaumin/data.php>

TABLE IV: From (a) to (f), six interesting classification examples of instances (Description and Image) belonging to the test set of the Ferramenta dataset. The Visual and Text columns represent respectively the output produced by the classifier 1 and 2 showed in Figure 2. The output of the final composer is shown in the column Fusion.

	Text	Visual	Fusion	
(a)				
actual	screwdriver	screwdriver	screwdriver	
predicted	screwdriver	<b>glue</b>	screwdriver	
Description	silverline 918547 set 6 screwdrivers, 6 pcs			
(b)				
actual	cart	cart	cart	
predicted	<b>scissor</b>	cart	cart	
Description	cart box trolleys with solid tires			
(c)				
actual	screwdriver	screwdriver	screwdriver	
predicted	<b>socket wrench</b>	<b>cart</b>	screwdriver	
Description	ks tools 159.1203 screwdriver ergotorque plus key 5.5 mm., ks tools 159.1203 screwdriver key ergotorqueplus 5.5 mm.			
(d)				
actual	safe	safe	safe	
predicted	safe	safe	safe	
Description	88352 staco safe measure s, 88352 staco safe measure s			
(e)				
actual	chain	chain	chain	
predicted	chain	<b>circular saw blade</b>	<b>circular saw blade</b>	
Description	yale p1040sc deadbolt door locks high security chrome trim, yale locks p1040sc deadbolt door high security chrome trim			
(f)				
actual	nail	nail	nail	
predicted	<b>screw</b>	nail	<b>screw</b>	
Description	Hardware bulk pack of 20 nails for masonry 3 x 70 mm, bulk pack of 20 Hardware nails for masonry 3 x 70 mm			

## V. CONCLUSIONS

In this work, we presented a late fusion multimodal setting that combines text and visual features based on weighting and meta-learning combination methods. We also presented a multimodal dataset that can be used in future multimodal settings on a variety of real-world applications merging natural language processing and computer vision. Our results indicate that the proposed multimodal setting outperforms classifiers based on an early fusion approach. On the basis of the results obtained on Ferramenta and PASCAL VOC2007 dataset, our multimodal setting is recommended for applications where ambiguous text can exploit image to resolve ambiguities and, vice versa, to enhance performance.

## REFERENCES

- [1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.
- [2] M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 902–909.
- [3] Y. Feng and M. Lapata, "Visual information in semantic representation," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, pp. 91–99.
- [4] M. J. Huiskes, B. Thomee, and M. S. Lew, "New trends and ideas in visual concept detection: the mir flickr retrieval evaluation initiative," in *Proceedings of the international conference on Multimedia information retrieval*. ACM, 2010, pp. 527–536.
- [5] C. W. Leong and R. Mihalcea, "Going beyond text: A hybrid image-text approach for measuring word relatedness." in *IJCNLP*, 2011, pp. 1403–1407.
- [6] D. Kiela and L. Bottou, "Learning image embeddings using convolutional neural networks for improved multi-modal semantics," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, October 2014, pp. 36–45.
- [7] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, vol. 16, no. 6, pp. 345–379, 2010.
- [8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.
- [9] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.
- [10] E. Bruni, G. B. Tran, and M. Baroni, "Distributional semantics from text and images," in *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, ser. GEMS '11, 2011, pp. 22–32.
- [11] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, no. 1, pp. 1–39, 2010.
- [12] Y. Peng, X. Zhou, D. Z. Wang, I. Patwa, D. Gong, and C. V. Fang, "Multimodal ensemble fusion for disambiguation and retrieval," *IEEE MultiMedia*, vol. 23, no. 2, pp. 42–52, 2016.
- [13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009.
- [14] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015.
- [15] L. Von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2004, pp. 319–326.