

Multimodal Sentiment Analysis: A Multitask Learning Approach

Mathieu Pagé Fortin and Brahim Chaib-draa

Department of Computer Science and Software Engineering, Laval University, Québec, Canada
mathieu.page-fortin.1@ulaval.ca, brahim.chaib-draa@ift.ulaval.ca

Keywords: Multimodal, Sentiment Analysis, Emotion Recognition, Multitask Learning, Text and Image Modalities.

Abstract: Multimodal sentiment analysis has recently received an increasing interest. However, most methods have considered that text and image modalities are always available at test time. This assumption is often violated in real environments (e.g. social media) since users do not always publish a text with an image. In this paper we propose a method based on a multitask framework to combine multimodal information when it is available, while being able to handle the cases where a modality is missing. Our model contains one classifier for analyzing the text, another for analyzing the image, and another performing the prediction by fusing both modalities. In addition to offer a solution to the problem of a missing modality, our experiments show that this multitask framework improves generalization by acting as a regularization mechanism. We also demonstrate that the model can handle a missing modality at training time, thus being able to be trained with image-only and text-only examples.

1 INTRODUCTION

Using a computational method, *sentiment analysis* aims to recognize whether a given source of data conveys a positive or a negative sentiment. A more fine-grained version of this problem aims to predict subtle categories of emotions such as *anger*, *sadness*, *excitement*, etc (Duong et al., 2017). Sentiment analysis has important applications for the study of social media, where billions of messages are published everyday¹. They often contain opinions and are thus a very rich source of information.

Multimodal sentiment analysis has received an increasing level of attention recently (Soleymani et al., 2017). On social media, users are encouraged to post images with a text description, and these two modalities can offer complementary information to better guide the analysis and reduce the classification error. However, most recent methods still consider that both modalities are always available at test time. This assumption is often violated in real environments (e.g. social media) as users do not always publish a text with an image.

Late-fusion can be used as a solution to the problem of missing modality by training two monomodal models and by producing the multimodal prediction by weighting each score (Atrey et al., 2010). If a modality is absent, the other monomodal classifier

can still make a prediction. However, the performance of late-fusion generally suffers from its simplicity, as it considers each modality independently and cannot learn discriminative multimodal interactions (Glodek et al., 2011).

Another approach is to develop a method that is robust to a missing modality, for instance with a generative model (Ngiam et al., 2011) or by minimizing a metric over multimodal representations (Sohn et al., 2014; Duong et al., 2017). However, such method requires much more complex training strategies.

Additionally, these methods cannot easily be trained with monomodal data. Collecting and especially labelling multimodal datasets with sentiment or emotion are laborious tasks. These datasets are therefore very often either noisy or small. A fraction of the dataset can even be composed of image-only or text-only examples. For instance, You et al. (2016) published a dataset in which each example has been annotated by at least five people. However, from the total of 22K images, only 8K of them are paired with a text. With typical multimodal methods, only the image-text pairs are used for training and the image-only examples are wasted (Duong et al., 2017). This drawback has been neglected in previous work, whereas it is an important issue for small datasets.

In this paper we propose a new method to tackle the problem of missing modality at test time *and* training time. It leverages a multitask framework to combine multimodal information when it is available,

¹<https://www.socialpilot.co/blog/social-media-statistics>

while being able to handle the cases where a modality is missing. Our model contains one classifier for each task: one that only analyzes the text, another that only analyzes the image, and another that performs predictions based on the fusion of both modalities. This method overcomes the simplicity of late-fusion and the problem of a missing modality for two reasons. First, the multimodal classifier can use any fusion technique to learn complex multimodal interactions. Second, the monomodal classifiers enable the model to perform accurate predictions even in situations where image-text pairs are not always present. Similarly to Vielzeuf et al. (2018), our approach is multitask in the sense that it considers a multimodal and two monomodal classification problems at the same time.

We also show that, compared to training different models on each task individually, some benefits arise from multitask learning. In addition to being more simple to develop (only one model is trained end-to-end), our experiments support that it also presents two other advantages. First, the results show that the multimodal classification can generalize better compared to when each classifier is trained individually. Second, it becomes easy to train a multimodal model with additional monomodal data. This enables the feature extractors and the monomodal classifiers to be trained with image-only or text-only examples, which can improve the performances. Therefore we differ from Vielzeuf et al. (2018) as we do not only consider the monomodal classifiers as a regularization mechanism. We also leverage their potential to deal with a missing modality.

2 RELATED WORK

Textual Sentiment analysis is a very active area of research in Natural Language Processing (Soleymani et al., 2017). Typically, the approaches in this area can be divided in two groups: *lexicon-based* and *machine learning-based* sentiment analysis. Recently, machine learning approaches gained more interest since they showed a real potential for automatically learning rich text representations from data, which are then used to classify the sentiment. Several techniques based on machine learning have been proposed, including classification by using *word embeddings* (Mikolov et al., 2013), the use of Recurrent Neural Networks (RNN) (Camacho-Collados and Pilehvar, 2017) or the use of Convolutional Neural Networks (CNN) (Kim, 2014).

Visual sentiment analysis is a more recent research avenue (Campos et al., 2015). The challenge of bridg-

ing the *affective gap* (Machajdik and Hanbury, 2010) makes this avenue more complex than the detection of concrete objects. Still, various approaches have been proposed by using state-of-the-art visual models. For instance, You et al. (2015) proposed a CNN inspired by AlexNet (Krizhevsky et al., 2012). Wang et al. (2016) used two parallel AlexNets to predict the adjective and noun labels of the image, and then used both representations to predict the sentiment. Similarly, You et al. (2017) proposed an attention mechanism based on the adjective label that weights the feature maps extracted by VGG-16 (Simonyan and Zisserman, 2014). Their weighted sum is then used for the sentiment classification.

Multimodal sentiment analysis combines innovations from both textual and visual sentiment analysis. Chen et al. (2017) used the CNN from (Kim, 2014) to extract a representation of texts, together with a network similar to AlexNet to analyze the images. The two representations are then fused and classified by a small neural network. Xu and Mao (2017) proposed to use the representations given by a VGG model pre-trained on ImageNet (Deng et al., 2009) and another VGG model pre-trained on a place dataset. These representations are then used to guide an attention mechanism on top of an LSTM that extracts a representation of the text.

One drawback of these previous multimodal methods is that, although they provide more accurate results than monomodal models, they are limited to multimodal data and cannot process image-only or text-only inputs at test time. In real situations and especially on social media, it is common to only have access to one of the two modalities. Few work has been done to solve this issue.

Sohn et al. (2014) proposed to minimize the *variation of information* between modality representations. By reducing such metric, the model learns to be more robust to a missing modality. Duong et al. (2017) developed a model that applies a similar idea to sentiment analysis. The model learns to minimize the distance between the representations of a text-image pair while maximizing the distance between the representations of an image and a randomly chosen text of another class.

Perhaps the most similar work to ours is the CentralNet (Vielzeuf et al., 2018). The architecture also contains two monomodal classifiers and a central multimodal classifier. However, this approach is mainly presented as a multimodal fusion technique where the authors leverage a multitask learning only to benefit from the implicit regularization that it introduces. In this paper, we also support that such multitask learning improves generalization, but we em-

phasize that the monomodal classifiers are important to further augment the training set with monomodal data and to deal with a missing modality at test time.

3 MULTIMODAL MULTITASK LEARNING

In this section, we describe the problem more formally, the model proposed to tackle it and the specific training procedure to handle a missing modality at train and test time.

3.1 Problem Formulation and Overview of the Proposed Framework

Given a dataset X composed of image-text pairs (I_p, T_p) , unpaired images I_u , and unpaired texts T_u , with their corresponding labels $y \in Y$, we aim to learn a model that can predict the label y^i of an instance x^i . We tackle the problem where the label is a sentiment or an emotion. We want the model to be able to handle the three following cases: the instance x^i at training and test time can be either a pair (I_p^i, T_p^i) , an image I_u^i , or a text T_u^i .

We propose to solve this problem with a multi-task learning framework as shown in Figure 1. The whole model is composed of two features extractors: a Visual network Φ_v , a Textual network Φ_t ; and three auxiliary classifiers: a Visual classifier C_v , a Textual classifier C_t , and a Multimodal classifier C_m . The Visual network and the Textual network respectively extract a representation of the image and the text, while each classifier specializes itself either in the classification of the image representation, the text representation, or the multimodal representation. During training, when an instance is a text-image pair, the whole model is updated, whereas when an instance is an image or a text, only the corresponding features extractor and classifier are trained. Similarly, at test time the prediction is done with the corresponding classifier.

3.2 Modules

Visual Network. The objective of the Visual network Φ_v is to extract a representation of the input image. In our experiments, we use the DenseNet-121 (Huang et al., 2017) pretrained on ImageNet. The classification layer is replaced by a fully-connected layer of 300 neurons to produce the image representation $r_v \in R^{300}$.

Textual Network. Similarly to Chen et al. (2017), we use a slight variation of the simple CNN from

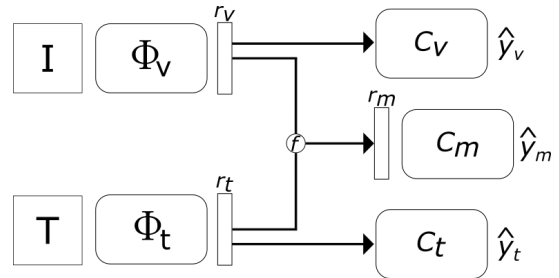


Figure 1: Overview of our multimodal multitask approach. The image I is analyzed by the CNN Φ_v to produce a representation r_v . Similarly, a representation r_t is extracted from the text T by Φ_t . Then, three tasks are considered: 1) the prediction of r_v by the classifier C_v , 2) the prediction of r_t by C_t , and 3) the prediction of the multimodal fusion $f(r_v, r_t)$ by C_m .

(Kim, 2014) that we present here. First, the input text is embedded using pre-trained representations of words in 300 dimensions (Mikolov et al., 2013). Then Dropout is applied with a probability of 0.25 to reduce overfitting. A convolutional layer of 100 filters extracts features with a window size of $h_i \times 300$. The nonlinear activation ReLU is used. Finally, global max, average and min pooling are used, such that:

$$p_i = [\max(c_i), \text{avg}(c_i), \min(c_i)] \in R^{100 \times 3}, \quad (1)$$

where c_i is the feature maps given by the convolutional layer i .

The process for one layer with filters of size $h_i \times 300$ has been described. For our Textual network, three parallel layers are used with $h_1 = 3, h_2 = 4, h_3 = 5$ to capture different n -gram patterns. The three resulting p_i are concatenated and a fully-connected layer with 300 neurons is used to obtain the text representation $r_t \in R^{300}$.

Visual Classifier. The Visual classifier aims to predict the sentiment expressed by the image. It learns to predict the class by taking an image representation r_v . In our experiments, this classifier is made of two fully-connected layers of 224 and 128 neurons, followed by a classification softmax layer.

Textual Classifier Similarly, the Textual classifier only depends on the text information to make its prediction. It uses the text representation r_t and is trained to recognize the class. Its architecture is similar to the Visual classifier.

Multimodal Classifier. Finally, this central classifier leverages both representations of the modalities r_v and r_t to make predictions. An initial step is to perform the fusion of representations with a function $f(r_v, r_t)$. To this end, several techniques

have been proposed, ranging from simple (e.g. concatenation, summation) to more complex ones (Hori et al., 2017; Zadeh et al., 2017; Vielzeuf et al., 2018). In our experiments, we use the concatenation for its simplicity and let sophisticated fusion techniques for future work. A classifier made of two fully-connected layers and a softmax layer is used to predict the class.

3.3 Training Procedure

We consider three tasks that are each performed by a different classifier. One task is to predict the sentiment \hat{y}_v from the visual information only:

$$\begin{aligned} r_v &= \Phi_v(I), \\ \hat{y}_v &= C_v(r_v), \end{aligned} \quad (2)$$

where Φ_v is the Visual network and C_v is the Visual classifier.

Another task is to predict the sentiment \hat{y}_t from the text only:

$$\begin{aligned} r_t &= \Phi_t(T), \\ \hat{y}_t &= C_t(r_t), \end{aligned} \quad (3)$$

where Φ_t is the Textual network and C_t is the Textual classifier

Finally the main task is to predict the sentiment \hat{y}_m from the fusion of the image and text representations:

$$\begin{aligned} r_m &= f(r_v, r_t), \\ \hat{y}_m &= C_m(r_m), \end{aligned} \quad (4)$$

where $f(\cdot, \cdot)$ is the fusion function and C_m is the Multimodal classifier.

For each of the three tasks $j \in \{v, t, m\}$, its auxiliary loss $L_j(\hat{y}_j, y)$ is defined by the cross-entropy. The whole model is trained to minimize:

$$L = \alpha_v L_v(\hat{y}_v, y) + \alpha_t L_t(\hat{y}_t, y) + \alpha_m L_m(\hat{y}_m, y), \quad (5)$$

where α_v , α_t and α_m are hyperparameters to weight the loss of each task. In this work we set them all to 1, but in future work we plan to investigate how we can learn them from data.

Furthermore, we remind that we consider the problem where a training instance can be either an image-text pair, an image (only), or a text (only). When an image-text pair is available, the two features extractors Φ_v and Φ_t and the three classifiers C are trained according to the multitask loss defined in equation 5.

When the instance is an unpaired image, the loss function only includes the prediction of the Visual classifier:

$$L = \alpha_v L_v(\hat{y}_v, y). \quad (6)$$

Similarly, when the instance is a text the loss function only includes the prediction made by the Textual classifier:

$$L = \alpha_t L_t(\hat{y}_t, y) \quad (7)$$

4 EXPERIMENTS AND RESULTS

We conducted experiments to evaluate our proposed approach for multimodal sentiment analysis. We now describe the datasets that are used, the experiments and the baselines. Then we present and discuss the results.

4.1 Datasets

Flickr Emotion. (You et al., 2016). This dataset contains images that have been annotated by at least 5 workers from Amazon Mechanical Turks. For each image, they were asked to attribute a label between eight emotions: *amusement*, *anger*, *awe*, *contentment*, *excitement*, *disgust*, *fear*, and *sadness*. We used the given URLs to download, with the Flickr API², the texts associated with the images. We only used the examples where the majority of workers agreed for a particular label. Since some examples can be ambiguous, we kept the top 20% of examples that received all the votes for the same emotion and we randomly divide it equally to form the validation and the test sets. The remaining 80% is used for training. Table 1 shows the statistics for each dataset and each class.

VSO. (Borth et al., 2013). Visual Sentiment Ontology (VSO) is widely used for sentiment analysis experiments due to its large number of examples. However, the way VSO was collected makes it very noisy. It has been built by querying Flickr with adjectives and nouns, which are then used to label the data as expressing positive or negative sentiment. Similarly to Chen et al. (2017), we downloaded the images and used the Flickr API to collect the texts associated with the images. We removed the examples with less than 5 words and more than 150 words, resulting in 301,042 pairs of images and texts. We randomly splitted the dataset into 80% training, 10% validation and 10% test.

4.2 Experimental Setting

We conduct two experiments and compare our results with 6 variants as baselines.

²www.flickr.com

Table 1: Statistics of datasets.

VSO		Flickr Emotion		
Sentiment	Quantity	Emotion	Quantity (I-T pairs)	Quantity (I-only)
positive	187,402	amusement	1,485	3,270
negative	113,640	anger	422	762
Total	301,042	awe	1,165	1,789
		contentment	1,872	3,312
		disgust	676	929
		excitement	1,266	1,482
		fear	367	613
		sadness	910	1,755
		Total	8,163	13,912

4.2.1 Experiments

We consider two variants of experiments to evaluate our model. We first compare the performance of baselines and our model trained with multitask learning. Then, we experiment the possibility of leveraging monomodal examples to improve multimodal classification.

Experiment 1. Previous work have shown that multitask learning can improve generalization by introducing regularization mechanisms (Ruder, 2017; Vielzeuf et al., 2018). In our framework, multitask learning regularizes the network by introducing a representation bias and an inductive bias as described by Ruder (2017). By training monomodal and multimodal classifiers, the image and text representations are learned with the objective of being discriminative. On the one hand this improves the quality of the monomodal representations before their fusion, and on the other hand it has less tendency to overfit since each monomodal representation is shared by two classifiers (i.e. the monomodal and the multimodal classifiers).

In this first setup, we evaluate the effects of regularization introduced by multitask learning and the ability of our model to handle a missing modality. We also reuse the text-based and image-based classifiers to perform late-fusion as a baseline. We remove the unpaired images and texts, and the examples in which the text contains less than five words. The architectures, hyperparameters and train/val/test splits are the same for each experimentation. Each experimentation is repeated at least three times (to reduce the variance) and the average is reported. For the multitask model, early stopping is performed on the basis of the main task: multimodal classification.

Experiment 2. One benefit of the multitask approach in the context of multimodality is that we can leverage monomodal training data. We experiment how the generalization of each task is affected by this

additional training data. We first evaluate the performance of our model when enriching the training set of Flickr Emotion dataset with unpaired images. Then, to further demonstrate the advantage of leveraging unpaired data, we experiment this training procedure on VSO with different amounts of paired data (0.1%, 1%, 5%, 25%, 100%), while the rest is used as monomodal examples.

4.2.2 Baselines

We compare our model trained with multitask learning to the following variants:

1. *Image-based classifier* (SI): Visual network Φ_v with Visual classifier C_v .
2. *Text-based classifier* (ST): Textual network Φ_t with Textual classifier C_t .
3. *Single-task multimodal classifier* (SM): Visual network Φ_v and Textual network Φ_t with Multimodal classifier C_m .
4. *Late-fusion*: Trained *Image-based classifier* and *Text-based classifier* are reused and their predictions are averaged.
5. *Multimodal-text only* (SM_T): *Single-task multimodal classifier* is reused but the images are absent at test time.
6. *Multimodal-image only* (SM_I): *Single-task multimodal classifier* is reused but the texts are absent at test time.

4.3 Results and Discussion

4.3.1 Experiment 1: Regularization and Missing Modality

We evaluated the benefits of our multitask framework on the generalization of each three tasks: text, image,

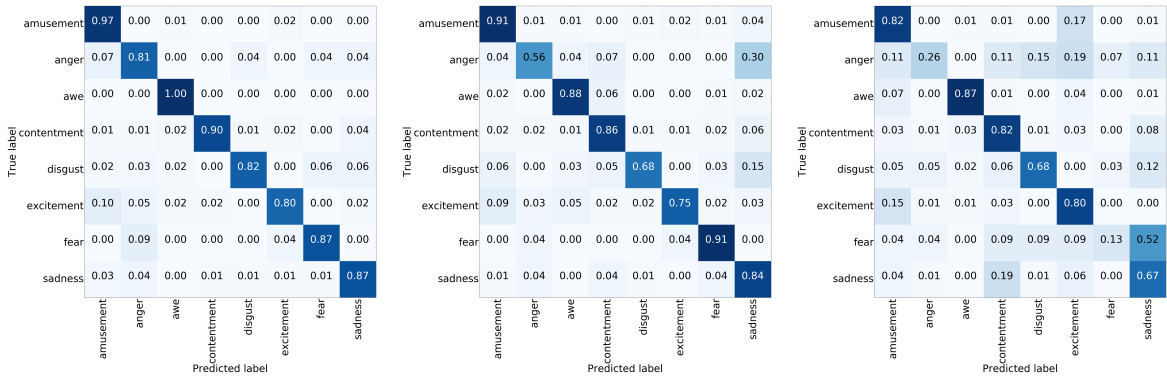


Figure 2: Confusion matrices for (from left to right) the multimodal, text and image classifiers (C_m , C_t , C_v) trained with multitask learning on Flickr Emotion.

and multimodal classification. The accuracy and F1-score on Flickr Emotion and VSO datasets are shown in Table 2 and Table 3. The macro-averaged F1 (F1-Macro) is the F1-score averaged over all classes.

First, our results confirm the benefits of multimodal classification. The multimodal classifiers outperform the text-based and the image-based classifiers on both datasets by a large margin. The baselines SM_I and SM_T show the performances that the *single-task multimodal classifier* obtains on Flickr Emotion dataset with image-only and text-only examples, respectively. The results clearly show that this classifier is not robust to a missing modality, since the accuracy drops to 47.92% and 79.90%. The monomodal classifiers (C_v and C_t) of our model obtain an accuracy of 70.34% and 83.99% with text-only and image-only examples, respectively. These results show that our model is more robust to a missing modality.

Interestingly, the results shown in Table 2 support that training our model with multitask learning improves generalization for the task of multimodal classification. On Flickr Emotion dataset, the model trained with multitask learning improves the accuracy by 1.24%. We performed a t-test that supported a significant increase (p-value = 0.016). The accuracy scores obtained by C_v and C_t are similar to those obtained by the single-task models SI and ST. However, the macro-averaged F1-scores are higher by 2.59% and 1.13% on Flickr Emotion dataset for the Visual and Textual classifiers trained with multitask learning. These improvements are consistent with the experiments performed by Vielzeuf et al. (2018), in which they used a similar multitask learning as a regularization mechanism.

Late-fusion obtains significantly lower performances on Flickr Emotion dataset due to the lack of interactions between multimodal features. Surprisingly, late-fusion and SM obtain similar performances to our model on VSO (see Table 3). A t-test sug-

Table 2: Results of experiment 1 for the proposed multi-task model and baselines on Flickr Emotion dataset. (See Section 4.2.2 for baselines).

Method	Classifier	Accuracy	F1-Macro
Baselines	SI	70.59	0.5808
	ST	83.70	0.7982
	L-Fus	89.09	0.8564
	SM	89.93	0.8659
	SM_I	47.92	0.4064
	SM_T	79.90	0.7486
Ours	C_v	70.34	0.6067
	C_t	83.99	0.8095
	C_m	91.17	0.8803

Table 3: Results of experiment 1 for the proposed multitask model and baselines on VSO dataset.

Method	Classifier	Accuracy	F1-score
Baselines	SI	69.91	0.7767
	ST	84.15	0.8779
	L-fus	85.73	0.8913
	SM	85.79	0.8868
Ours	C_v	69.73	0.7648
	C_t	83.79	0.8775
	C_m	86.35	0.8894

gests that the superior accuracy of C_m is not significant (p-value = 0.09). We speculate that since VSO is noisy, the upper-bound that any algorithm can approach on this dataset is relatively low. On the other hand, since the dataset is very large, it is possible that simple models – such as late-fusion or even a text-based classifier – are already able to obtain performances that approach it. This would explain why the multimodal classifier only improves the accuracy by 2% over a text-based classifier, and why a multimodal fusion technique does not outperform by a significant margin a simple late-fusion.

We also report in Figure 2 the confusion matri-

Table 4: Results of the experiment 2 for the proposed multitask model trained with more image-only examples on Flickr Emotion dataset.

Training	Classifier	Accuracy	F1-Macro
Multitask + images	C_v	75.47	0.6507
	C_t	83.92	0.8063
	C_m	91.59	0.8810

ces for the three classifiers of our model trained on Flickr Emotion dataset. They offer some insights on the monomodal and multimodal classifications. For instance, we observe that the Visual classifier hardly recognizes the negative emotions, especially fear and anger (0.13 and 0.26 respectively). This problem is mostly solved by multimodality (0.87 and 0.81 for fear and anger).

4.3.2 Experiment 2: Monomodal Training Data

As discussed in Section 3.3, our multitask framework enables the model to be trained with monomodal data. This is particularly useful for Flickr Emotion dataset which is limited to 8,163 image-text pairs, but actually contains 13,912 image-only examples (see Table 1).

We trained our model with these additional images and we report the results in Table 4. When compared to the results obtained in the first experiment, the accuracy of the Visual classifier now significantly improves by 5%. This also slightly increases the performance of the multimodal classifier by 0.42%. This observation can be explained by the fact that the Visual network Φ_v is trained with more data and thus the quality of the visual representation is improved.

Finally, we further evaluated the ability of the model to be trained with few multimodal data on VSO dataset. More specifically, a given fraction of the training set is kept as multimodal data, while the remaining portion is divided in two: the texts are removed from the first half and the images from the second. The validation and test sets are 100% multimodal. As a comparison, the same model is trained only with the multimodal split.

Figure 3 shows the accuracy for different proportions of multimodal data. As expected, a model only trained with few multimodal data generalizes very poorly with an accuracy of 62.56%. In that case, the advantage of performing multimodal classification is only visible when the model can be trained with a large amount of image-text pairs. On the contrary, when our model can also be trained with monomodal data, the multimodal classifier always performs better than a text-based classifier even with very few training image-text pairs. The superiority of multimodal clas-

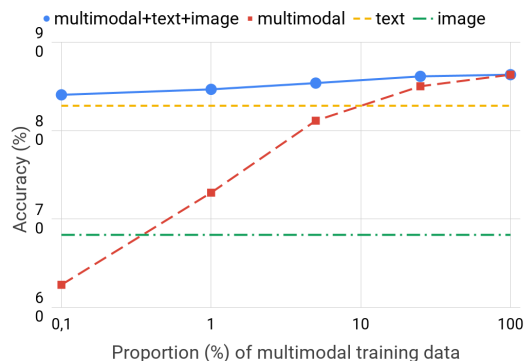


Figure 3: Classification accuracy on VSO dataset for different proportions of multimodal training data. The results obtained by the monomodal text-based and image-based classifiers are also shown as baselines.

sification improves with the amount of multimodal data, but even when 0.1% ($n = 230$) of the training set is multimodal, it performs 1.24% better than the text-based classifier with an accuracy of 84.07%.


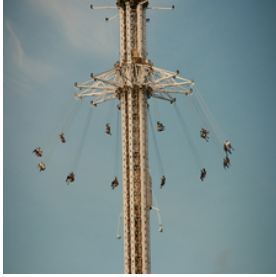
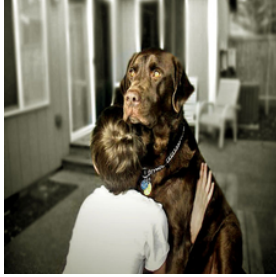
4.3.3 Discussion

The previous results highlighted three benefits of our multitask approach. First, it offers a simple solution to the problem of missing modality at test time. As discussed above, this is an important feature for social media analysis. Second, we observed empirically that training the model with multitask learning improves the performances of the multimodal classifier. This regularization mechanism is an additional tool that can easily be used in any multimodal models. Third, the multitask approach enables to leverage monomodal training data. We showed that our model, when trained with monomodal data, already outperforms the other baselines with as few as 230 multimodal training examples. Since datasets are often hard to collect, and even more for multimodal ones, the ability to learn with monomodal data can be a very useful feature.

Our results also confirmed that sentiment analysis and emotion recognition are harder tasks with images than with texts. This is due to the problem of *affective gap* (Machajdik and Hanbury, 2010), which means that there is a large conceptual distance between low-level information contained in the input and the abstract concept of human sentiment that we aim to predict. This gap is larger in images than in texts, since words are more expressive and at a higher level of abstraction than pixels. The combination of text and image information certainly helps to bridge the *affective gap*.

We investigated the use of multitask mostly as a solution to the problem of missing modality. How-

Table 5: Examples of predictions for the three classifiers on Flickr Emotion test set.

Image			
Text	"since we knew what to expect or how comfortable we would feel we brought our rack and a single rope"	"not my idea of fun, each to their own"	"love between a boy and his best friend"
Ground Truth	awe	amusement	contentment
C_m prediction	awe	amusement	contentment
C_t prediction	contentment	fear	contentment
C_v prediction	excitement	amusement	sadness

ever, there can also be value in the monomodal classifications. Multimodal classification relies on the assumption that both modalities convey the same message, which is not always true. Table 5 shows examples of image-text pairs and the predictions made by the three classifiers of our model. We can observe that the text and the image can express emotions that are ambiguous or different. For instance, the image of a carousal with the text "not my idea of fun, each to their own" can at the same time express amusement and fear. In those cases, the monomodal and multimodal classifiers can together give a more complete description of the message posted by a user, which is not measured by common quantitative metrics such as accuracy and F1-score.

5 CONCLUSION

In this paper, we proposed a multitask approach for multimodal sentiment analysis and emotion recognition. Our approach adds two auxiliary image-based and text-based classifiers to the traditional multimodal framework, which enables to handle a missing modality at test time and training time. Our experiments show that, not only it offers a viable and simple solution to a missing modality, but also that multitask learning acts as a regularization mechanism that can improve generalization.

To the best of our knowledge, we are also the first to explore the use of monomodal training data to improve multimodal classification. We showed that when a sufficient amount of monomodal data is available, very few multimodal data is necessary to ob-

tain excellent results. These observations suggest that by simply extending the multimodal framework with auxiliary monomodal classifiers, the training set can easily be augmented with additional monomodal data to improve its performances.

REFERENCES

- 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.
- Borth, D., Ji, R., Chen, T., Breuel, T., and Chang, S.-F. (2013). Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 223–232. ACM.
- Camacho-Collados, J. and Pilehvar, M. T. (2017). On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis. *arXiv preprint arXiv:1707.01780*.
- Campos, V., Salvador, A., Giro-i Nieto, X., and Jou, B. (2015). Diving deep into sentiment: Understanding fine-tuned cnns for visual sentiment prediction. In *Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia*, pages 57–62. ACM.
- Chen, X., Wang, Y., and Liu, Q. (2017). Visual and textual sentiment analysis using deep fusion convolutional neural networks. In *Image Processing (ICIP), 2017 IEEE International Conference on*, pages 1557–1561. IEEE.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.
- Duong, C. T., Lebet, R., and Aberer, K. (2017). Multimodal classification for analysing social media. *arXiv preprint arXiv:1708.02099*.

- Glodek, M., Tschechne, S., Layher, G., Schels, M., Brosch, T., Scherer, S., Kächele, M., Schmidt, M., Neumann, H., Palm, G., et al. (2011). Multiple classifier systems for the classification of audio-visual emotional states. In *Affective Computing and Intelligent Interaction*, pages 359–368. Springer.
- Hori, C., Hori, T., Lee, T.-Y., Zhang, Z., Harsham, B., Hershey, J. R., Marks, T. K., and Sumi, K. (2017). Attention-based multimodal fusion for video description. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 4203–4212. IEEE.
- Huang, G., Liu, Z., Weinberger, K. Q., and van der Maaten, L. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, volume 1, page 3.
- Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.
- Machajdik, J. and Hanbury, A. (2010). Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 83–92. ACM.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
- Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696.
- Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.
- Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Sohn, K., Shang, W., and Lee, H. (2014). Improved multimodal deep learning with variation of information. In *Advances in Neural Information Processing Systems*, pages 2141–2149.
- Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S.-F., and Pantic, M. (2017). A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14.
- Vielzeuf, V., Lechervy, A., Pateux, S., and Jurie, F. (2018). Centralnet: a multilayer approach for multimodal fusion. *arXiv preprint arXiv:1808.07275*.
- Wang, J., Fu, J., Xu, Y., and Mei, T. (2016). Beyond object recognition: Visual sentiment analysis with deep coupled adjective and noun neural networks. In *IJCAI*, pages 3484–3490.
- Xu, N. and Mao, W. (2017). Multisentinet: A deep semantic network for multimodal sentiment analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2399–2402. ACM.
- You, Q., Jin, H., and Luo, J. (2017). Visual sentiment analysis by attending on local image regions. In *AAAI*, pages 231–237.
- You, Q., Luo, J., Jin, H., and Yang, J. (2015). Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *AAAI*, pages 381–388.
- You, Q., Luo, J., Jin, H., and Yang, J. (2016). Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *AAAI*, pages 308–314.
- Zadeh, A., Chen, M., Poria, S., Cambria, E., and Morency, L.-P. (2017). Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.