Multimodal Multitask Emotion Recognition using Images, Texts and Tags

Mathieu Pagé Fortin
Laval University
Québec, Canada
mathieu.page-fortin.1@ulaval.ca

Brahim Chaib-draa
Laval University
Québec, Canada
brahim.chaib-draa@ift.ulaval.ca

ABSTRACT

Recently, multimodal emotion recognition received an increasing interest due to its potential to improve performance by leveraging complementary sources of information. In this work, we explore the use of images, texts and tags for emotion recognition. However, using several modalities can also come with an additional challenge that is often ignored, namely the problem of "missing modality". Social media users do not always publish content containing an image, text and tags, and consequently one or two modalities are often missing at test time. Similarly, the labeled training data that contain all modalities can be limited.

Taking this in consideration, we propose a multimodal model that leverages a multitask framework to enable the use of training data composed of an arbitrary number of modality, while it can also perform predictions with missing modalities. We show that our approach is robust to one or two missing modalities at test time. Also, with this framework it becomes easy to fine-tune some parts of our model with unimodal and bimodal training data, which can further improve overall performance. Finally, our experiments support that this multitask learning also acts as a regularization mechanism that improves generalization.

CCS CONCEPTS

• **Computing methodologies → Machine learning**; *Natural language processing*; *Computer vision*; *Multi-task learning*.

KEYWORDS

Multimodal; Multitask; Emotion recognition; Image emotion recognition; Text emotion recognition; Tag emotion recognition

ACM Reference Format:
Mathieu Pagé Fortin and Brahim Chaib-draa. 2019. Multimodal Multitask Emotion Recognition using Images, Texts and Tags. In *Crossmodal Learning and Application (WCRML'19), June 10, 2019, Ottawa, ON, Canada.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3326459.3329165

1 INTRODUCTION

The increasing use of social media, combined with the development of powerful data analysis technologies, recently opened interesting opportunities for the study of people's opinions and behaviour [22]. Billions of messages are published everyday[1], which forms a very rich source of information that can be profitable for several applications. In this paper, we study the task of recognizing the emotion (e.g. *excitement*, *anger*, *fear*, *etc.*) and the sentiment (*positive* or *negative*) expressed in these messages. The extraction of such high-level information has many applications such as personalized recommendations, opinion mining or health monitoring, among others [28, 34].

Early attempts at emotion recognition only used one modality as input such as texts, images or audio [20]. However the results can be hindered by the problem of "affective gap" [11], which means that there is a large conceptual distance between the low-level input (e.g. pixels) and the high-level notion of human emotion that the model tries to predict. This gap is generally smaller in texts since words are more expressive and at a higher level of abstraction than images. In this paper we complete images and texts by tags, which are at a higher level than texts. Another challenge comes from the fact that different people can react differently to the same stimuli; this is called the "subjective perception problem".

More recently, multimodal emotion recognition received an increasing interest due to its potential to reduce the impact of both previous challenges. Most recent work explored the use of textual and visual data as complementary modalities in the analysis of social media users' posts [6, 8, 33]. However there are often other information available that can be used. On Flickr[2] for instance, many users also publish tags, which are high-level keywords that aim to facilitate the search by summarizing the content of the message (see Figure 1). Previous work mainly used these tags for other tasks such as building a dataset [5, 32], tag recommendation [3, 27], or for cross-modal retrieval [10]. In this paper, we use them as a third input modality and we show that they can significantly improve the results.

However, using several modalities as inputs in a machine learning model can also come with an additional problem, namely the problem of missing modality [1]. Indeed, traditional multimodal models are fixed to a predefined number of modality. For example an image-text bimodal model such as the one proposed by Chen et al. [6] cannot efficiently handle text-only or image-only data, which limits its flexibility at test time and training time.

In fact, before developing an image-text-tags trimodal model, we must be aware that social media users do not always publish content that is made of an image, text and tags. In this case, the labeled data that are such triplets can be limited for training. A simple solution to this problem is to develop independent unimodal

WCRML'19, June 10, 2019, Ottawa, ON, Canada
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6780-6/19/06…$15.00
https://doi.org/10.1145/3326459.3329165

[1] https://www.socialpilot.co/blog/social-media-statistics
[2] www.flickr.com

Image			
Text	"prisoner's sadness i felt so bad for this little fellow ... he gave me such a sad look that moment i looked at him :-(even his colors seem to vanish"	"student revolt shakes the condem coalition . the student protests very nearly defeated the coalition less than six months into its rule ."	let it come , and let it be . let it come , and let it be ...
Tags	prisoner, bird, cage, sad	studentprotest, righttowork, protest, demonstration	beach, colors, dramatic, florida, photography, real
Emotion	Sadness	Anger	Awe

Figure 1: Examples of image-text-tags.

models and average their predictions [1, 18]. However, in many situations this is suboptimal since the model cannot learn discriminative multimodal interactions. In the same vein than [18], we propose in this article to leverage a multitask framework to avoid the problem of missing modality while being able to learn multimodal representations. With this approach, it becomes possible to augment multimodal training datasets with data composed of an arbitrary number of modality, while the model can also perform predictions with missing modalities.

More specifically, our contribution is twofold. First, we consider the application of tags as a complement to images and texts to improve emotion recognition. Unlike previous work that either ignored the tags or used them as regular textual features [7], we instead learn tags embeddings with another neural network.

Second, we develop a multitask approach [18] that has several benefits for multimodal tasks, such as acting as a regularization mechanism and increasing the robustness to a missing modality at test time and training time.

Although less explored than the traditional multitask learning [23], our approach is multitask in the sense that it considers several multimodal and unimodal classification problems at the same time. In the same line than Vielzeuf et al. [25] and Pagé Fortin and Chaib-draa [18], we support that this multitask learning, which stems directly from the use of several modalities, has similar benefits to the more common multitask learning [23] without requiring additional labels, thus boosting the main task by subsidiary tasks as is the case of traditional multitask learning. We differ from these two work by the number and the nature of modalities that are used.

More specifically our model is made of one classifier for image-text-tags triplets; three bimodal classifiers for image-text, image-tags, and text-tags pairs; and three unimodal classifiers for when only the image, the text or the tags are available. With this approach, our model can offer a robustness to one or two missing modalities at test time, and it becomes easy to fine-tune some parts of it with unimodal and bimodal training data, which can further improve performance. Finally, our experiments support the fact that the proposed multitask approach acts as a regularization mechanism that improves generalization.

The content of this paper is as follows. The next section presents the related work. Section 3 describes the model and the multitask training procedure. Section 4 presents experiments, and finally Section 5 discusses the results and the proposed multimodal multitask approach.

2 RELATED WORK

Emotion Recognition. Emotion recognition can be seen as a fine-grained problem related to sentiment analysis [32]. The latter aims at recognizing positive or negative polarity in data, while emotion recognition distinguishes several basic emotion categories. Emotion recognition can be divided in two:

(1) recognizing the emotion that a person is feeling,
(2) recognizing the emotion that is conveyed by a support (e.g. image, text, painting, music, etc.).

The first is the most active area of research and it mainly focuses on physiological signals: for instance by analyzing the video, audio and speech of someone communicating an opinion [21], or by considering EEG and eye tracking data of people watching emotional videos [15].

The second category considers the task of recognizing what emotion is expressed in content such as artistic paintings, photos, or social media posts [8, 30, 31]. The two challenges of "affective gap"[11] and "subjective perception problem"[30] complicate this task and the annotation of datasets, which are therefore very few.

This paper falls in the second category of emotion recognition as we develop a model for analyzing social media posts. We propose to reduce the impact of the affective gap and the subjective perception problem with a multimodal approach and by making a more efficient use of available data with a multitask framework, as we describe in more details in the next section.

Multimodal machine learning. Multimodal emotion recognition offers an interesting solution to reduce the impact of both previous challenges, since it enables the exploitation of several complementary sources of information, for instance images with texts. These additional features can therefore help bridge the affective gap and reduce the ambiguity of social message posts. Several work showed that multimodality can significantly improve performance [6–8]. For instance, Chen et al. [6] used a Convolutional Neural Network (CNN) from Kim [14] as a textual feature extractor, with a visual CNN to analyze the images. Both vector representations are then concatenated and classified by another small neural network. This approach illustrates a simple and effective baseline for multimodal classification. However, with this framework another problem occurs, as it cannot efficiently handle the absence of a modality at test time. Our proposed model addresses that by using three modalities with a flexible method that is robust to the absence of one or two modalities. In fact the problem of missing data is one of the most important challenge in multimodal machine learning [1], since in real situations there is no guarantee that every modality will be available. This is especially true on social media where each modality is generally optional.

To avoid this problem of missing data, previous work proposed solutions coming from metric learning [8, 13] or generative models [17, 19]. However, these methods generally require much more complex training strategies. Also, all the works mentioned above are limited to multimodal training data, whereas these datasets can be hard to collect. For instance, from the total of 22K images in the dataset published by You et al. [29], only half of them are paired with a text. With typical multimodal methods, these image-only examples are wasted since they are not used in the training procedure [8]. Very few work has been done to approach this issue.

Wagner et al. [26] proposed a multimodal model that uses a weighted majority vote: when a modality is absent, the corresponding classifier is simply ignored during the poll. The results obtained by such decision-based fusion are often inferior to a feature-based fusion, since it considers each modality independently. Therefore the model cannot learn discriminative multimodal features [8, 9].

Pagé Fortin and Chaib-draa [18] recently proposed a feature-based fusion within a multitask framework to enable the handling of a missing modality at test and train time. Their model architecture is inspired from CentralNet [25], where the authors proposed to train unimodal classifiers together with multimodal ones mainly because it can act as a regularization mechanism and improve multimodal classification. In [18], it has been shown that the proposed framework can also be used to avoid the problem of missing modality.

Generally, most existing multimodal approaches for emotion recognition on social media only use images and texts [6, 8, 18], although more metadata such as tags are sometimes available. Previous work on multimodal emotion recognition either ignored these tags or used them as regular words by concatenating them with texts [7].

In this paper, we differ from the above multimodal approaches by using tags in addition to images and texts to further help bridge the affective gap and reduce the ambiguity coming from the problem of subjectivity. Notice that Corchs et al. [7] used words and tags together in a Bayesian model averaging paradigm. Our approach is however different from this since we learn two different embeddings for texts and tags. Indeed, tags are not used as normal words since they are high-level keywords that summarize a message rather than words in a sentence. Also, we approach the problem of missing modality at train and test time by using a multitask framework [18]. Our model can therefore leverage more information by using trimodal, bimodal, and unimodal classifiers, whereas the model from Pagé Fortin and Chaib-draa [18] is only made of one bimodal and two unimodal classifiers for images and texts.

3 APPROACH

Following previous work as [8, 29], we define the problem of emotion recognition on social media as predicting the label of a user's post, where it can take one of these values: *amusement*, *anger*, *awe*, *contentment*, *excitement*, *disgust*, *fear*, or *sadness*. We also approach the task of sentiment analysis, where social media posts are either labeled as *positive* or *negative*. Generally, a post is considered as a message composed of an image, a text, or a text-image pair [8]. In this paper, we extend this definition by considering that the tags can also be a third part of a post. These tags are high-level keywords that often accompany posts on social media. By using them as inputs, we leverage a supplementary source of information that can help to recognize the emotion.

Also, since texts and tags are optional on Flickr and therefore the users sometimes do not write one or both of them, we leverage a multimodal multitask approach to tackle the problem of missing modality at test time and training time. By doing so, our model does not require the presence of all three types of information (as long as there is at least one of the three), but they can help the classification when they are available.

3.1 Model

Our whole model is composed of several sub-modules: three feature extractors (i.e. one for the image input, one for the text input, and one for the tags input), three unimodal classifiers, three bimodal classifiers and one trimodal classifier. Figure 2 illustrates our model that we now describe in details. The hyperparameters were evaluated on the validation set, but we found that training was robust to their values.

Image feature extractor. This network extracts a representation of the image input. In our experiments, we use DenseNet-121 [12] pretrained on ImageNet. The choice of DenseNet among other state-of-the-art CNNs is motivated by its ability to carry low- to high-level features. Psychological studies showed that emotions can be evoked by high-level semantics but also by low-level features such as colors, contrasts, shapes, etc. [2, 24]. We replace the last layer of DenseNet-121 by a fully-connected layer of 300 neurons to produce the visual representation $r_i \in R^{300}$.

Text feature extractor. We chose to use a text CNN similar to the one proposed by Kim [14] because it is faster and easier to train than the more sophisticated Recurrent neural networks (RNN) and yet it offers good performance. Since CNNs need a fixed size input, we limit the maximum number of words to 150. We truncate or pad the texts when necessary. The input text is first embedded

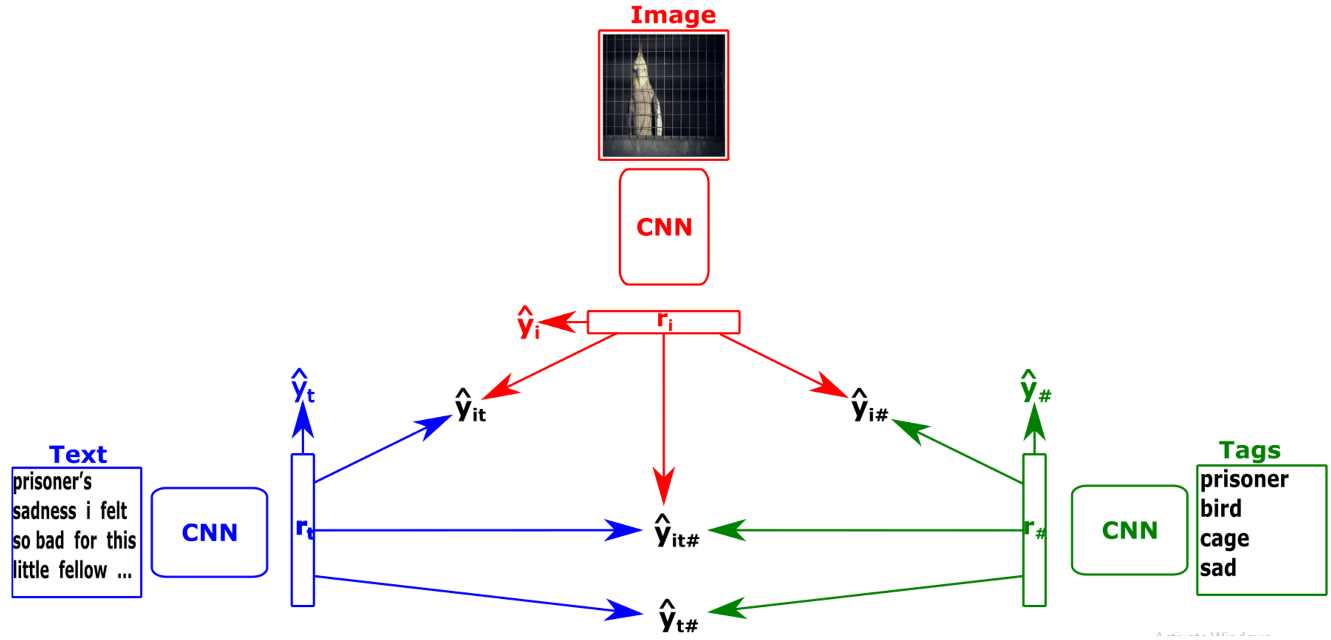


Figure 2: Overview of the proposed multimodal multitask model for emotion recognition.

using pre-trained representations of words in 300 dimensions from Word2Vec [16]. Dropout is applied with a probability of 0.25 to reduce overfitting. Following Chen et al. [6], we use three parallel convolutional layers of 100 filters to extract features with a window size of $h_i \times 300$ with $h_1 = 3, h_2 = 4, h_3 = 5$ to capture different *n-gram* patterns. After each convolutional layer i, the nonlinear activation ReLU is used to produce the feature maps c_i. Global max, average and min pooling are then computed, such that:

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

The three resulting $pool_i$ are concatenated and a fully-connected layer with 300 neurons produces the text representation $r_t \in R^{300}$.

Tags feature extractor. To have a fixed size input, we limit the maximum number of input tags to 10. We only keep the first ten tags or we add zero-padding, when necessary. The tags are first embedded in 250 dimensions with an embedding matrix initialized randomly. We use one convolutional layer of 300 filters with a window size of 1×250, followed by ReLU. Global max-pooling is then used to obtain the tags representation $r_\# \in R^{300}$

Unimodal classifiers. Three unimodal classifiers use either r_i, r_t, or $r_\#$ as input to make a prediction $\hat{y}_i$, $\hat{y}_t$, or $\hat{y}_\#$ respectively. Each classifier is made of a two fully-connected layers of 256 and 128 neurons with ReLU activations, followed by a softmax classification layer.

Bimodal classifiers Three bimodal classifiers use either the pair (r_i, r_t), $(r_i, r_\#)$, or $(r_t, r_\#)$ as input. The first step consists of performing the fusion of representations, for which we use the concatenation. Then, an architecture similar to the unimodal classifiers is used to make a prediction $\hat{y}_{it}$, $\hat{y}_{i\#}$, or $\hat{y}_{t\#}$, respectively.

Trimodal classifier. This central network uses the three representations $(r_i, r_t, r_\#)$ as inputs. These three feature vectors are also fused by concatenation. Then, an architecture similar to the other classifiers is used to make a prediction $\hat{y}_{it\#}$.

3.2 Multitask training procedure

Our multimodal multitask framework enables the model to be fed with unimodal and bimodal training data (in addition to trimodal data). Since there is a classifier for each combination of modality, including unimodal input, we can update the corresponding classifier(s) and feature extractor(s) according to the following loss function:

$$\begin{aligned} L \quad = \quad & L_{it\#}(\hat{y}_{it\#}, y) \\ & + L_{it}(\hat{y}_{it}, y) + L_{t\#}(\hat{y}_{t\#}, y) + L_{i\#}(\hat{y}_{i\#}, y) \\ & + L_i(\hat{y}_i, y) + L_t(\hat{y}_t, y) + L_\#(\hat{y}_\#, y), \end{aligned}$$

where $L(\cdot)$ is the cross-entropy loss and the subscripts $i, t, \#$ respectively indicate the presence of the image, text, or tag information to make the prediction $\hat{y}$. As we can see our approach is multitask in the sense that it is trained to simultaneously perform seven classification tasks with all the combinations of modalities. It enables the training of some parts of the model with data that lack one or two modalities since we can still make a prediction and compute a signal to update the feature extractors and classifiers that are considered. When a modality is absent, the corresponding terms in the equation are simply set to zero.

We use Adam with a learning rate of 0.01 and a batch size of 64, and we do early-stopping based on the loss mentioned in section 3.2.

Emotion	Quantity
Amusement	4241
Anger	1012
Awe	2677
Contentment	4626
Disgust	1405
Excitement	2526
Fear	829
Sadness	2456
Total	19772

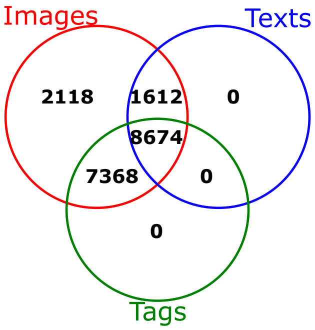


Figure 3: Summary of Flickr Emotion dataset. Left: number of examples for each emotion. Right: Number of examples in each modality.

4 EXPERIMENTS

4.1 Datasets

Flickr Emotion dataset [29] has been built by collecting images published on Flickr. The images have been annotated by five workers from Amazon Mechanical Turks, where they were asked to attribute one of these emotion labels: *amusement*, *anger*, *awe*, *contentment*, *excitement*, *disgust*, *fear*, or *sadness*. Similar to Duong et al. [8], we enriched this image-only dataset by using the given URLs to crawl, with Flickr API, the images, the texts and the tags, when available. We removed the examples in which strictly less than three workers agreed for a particular label. We divided the dataset into 80% training, 10% validation and 10% test sets. Also, to reduce the ambiguity of the evaluation data, we selected the validation and test data only from the subset of examples that received five votes for the same label. Figure 3 summarizes the composition of Flickr Emotion dataset. Note that since Flickr Emotion is originally an image-only dataset, there are no text-tags pairs nor text-only and tags-only examples.

VSO. Visual Sentiment Ontology (VSO) [4] has also been built by querying Flickr. It contains several hundred thousand examples that were annotated automatically as *positive* or *negative*, which is therefore much more noisy than Flickr Emotion dataset. We used the Flickr API and the given URLs to download the available image-text-tags triplet examples, which gave us 96,999 positive and 54,965 negative examples for a total of 151,964 image-text-tags triplets. We randomly split the dataset into 80% training, 10% validation and 10% test sets

For both datasets, we only kept the English texts of length between 5 and 150 words. The words and tags that appear less than 5 times in the training set are replaced by the token <UNK>.

4.2 Results

4.2.1 Experiment 1. As a first experiment, we evaluated the predictive power of each individual modality and each combination of modalities on the small and clean Flickr Emotion dataset and on the large and noisy VSO dataset. We trained three unimodal models to evaluate the performance of image-based, text-based and tags-based classification; three bimodal models for each pair of modalities; and a trimodal model. Note that these bimodal and trimodal baselines are trained in the traditional single-task fashion as done by Chen et al. [6].

Table 1: Accuracies (%) of experiment 1 on the trimodal subset of Flickr Emotion. (I: Image, T: Text, #: Tags)

Modality	Single-Task	Multitask (Ours)
I	70.61±0.57	69.74±2.50
T	78.32±2.59	77.69±0.87
#	78.32±1.51	78.76±0.76
I+T	80.26±0.50	85.06±1.13
I+#	85.09±0.47	85.90±0.65
T+#	89.41±0.60	91.10±0.39
I+T+#	92.48±0.53	**93.61**±0.50

Table 2: Accuracies (%) of experiment 1 on VSO dataset. (I: Image, T: Text, #: Tags)

Modality	Single-Task	Multitask (Ours)
I	70.40±0.18	70.48±0.38
T	75.45±0.31	73.87±0.71
#	79.13±0.09	77.80±0.34
I+T	77.31±0.91	77.46±0.26
I+#	79.31±0.51	79.77±0.25
T+#	82.22±0.13	82.10±0.93
I+T+#	83.27±0.52	**83.55**±0.43

Also, note that on Flickr Emotion dataset the unimodal and bimodal baselines could be trained with their corresponding subset of data. However, to make a fair comparison between each experiment regarding the amount of training data, we only used the subset of 8674 examples that are triplets of image-text-tags (see Figure 3) for each model. While training the unimodal and bimodal baselines, we simply ignored the other modalities from the dataset.

We trained our multimodal multitask model, following the training procedure introduced in section 3.2. Since our model is made of seven classifiers to handle one or two missing modalities at test time, we report the results of each one. To reduce variance, we repeat each experiment three times. The mean ±standard deviation of the accuracies on Flickr Emotion and VSO test sets are shown in Table 1 and 2, respectively.

First, the emotion/sentiment classification from images only confirms that it is the hardest problem, with an accuracy of 70.61±0.57% and 70.40±0.18%, respectively. This is mainly due to the affective gap [11] that is especially large in images. The gap is generally lower in texts, which is empirically confirmed by the higher results of text classification that obtained 78.32±2.59% and 75.45±0.31% of accuracy. The tags-based classification shows very similar performance to text-based classification on Flickr Emotion dataset with an accuracy of 78.32±1.51%. On VSO dataset tags classification is higher than text classification, with an accuracy of 79.13±0.09%. These results are consistent with the fact that tags serve as high-level keywords that summarize the post, which should reduce the affective gap.

As expected, the results significantly improve with every additional modality, especially on Flickr Emotion dataset. Interestingly, although text and tags classification obtained similar results of

Table 3: Accuracies (%) of unimodal baselines with late-fusion and our multimodal multitask model with additional unimodal and bimodal training data on Flickr Emotion dataset. (I: Image, T: Text, #: Tags)

Modality	Baselines	Multitask (Ours)
I	72.93±0.67	72.74±1.81
T	78.20±1.53	78.13±1.60
#	84.59±1.60	85.59±0.89
I+T	82.89±1.72	82.96±1.98
I+#	89.29±2.28	90.60±0.50
T+#	91.35±0.64	91.29±0.74
I+T+#	93.42±0.22	**94.80**±0.43

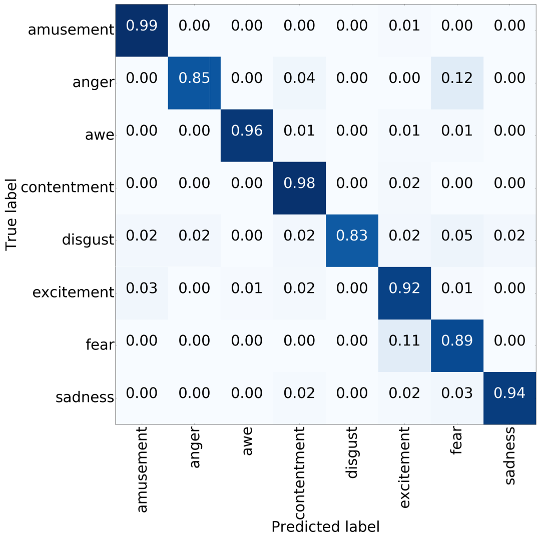
78.32% when taken separately on Flickr Emotion dataset, text-tags bimodal classification improved to 89.41±0.60%. Although less pronounced, this effect is also observed on VSO dataset, with a text-tags bimodal classification accuracy of 82.22±0.13. These results support that texts and tags convey complementary information. Without surprise, the best single-task classifier is the one that uses all three types of information, with an accuracy of 92.48±0.53% and 83.27±0.52% for Flickr Emotion and VSO datasets, respectively.

Then for our multimodal multitask approach, the results obtained by the unimodal and bimodal classifiers show that our model can handle unimodal and bimodal data at test time, even if it was trained with trimodal data. Interestingly, the results also show that our multitask approach improves the generalization performance of trimodal classification on both datasets.

Our trimodal classifier trained with multitask learning obtained an accuracy of 93.61%, as compared to 92.48% obtained by the trimodal classifier trained individually on the first dataset. On VSO dataset this increase is lower, with a trimodal accuracy of 83.55±0.43% when trained in a multitask fashion, whereas the single-task trimodal classifier obtained 83.27±0.52%. Moreover, the results are also improved on Flickr Emotion for bimodal classification. This increase varies from modest to high, with respectively a 0.81% and 4.8% absolute increase of accuracy for I+# and I+T classification. Our improved results of multimodal classification are consistent with previous works that leveraged a similar multitask framework as a regularization mechanism [18, 25].

4.2.2 Experiment 2. Next, we aim to demonstrate that our multitask approach also presents a very important feature: the possibility to do training with unimodal and bimodal data. To do so, we follow the multitask training procedure explained in section 3.2, and we augment the previous training set of Flickr Emotion dataset of 8674 examples with the 1612 image-text pairs, 7368 image-tags pairs, and 2118 image-only examples (see Figure 3). The validation and test sets remain the same as in the previous experiment. As a comparison, we train unimodal classifiers on all available unimodal data and we combine their prediction with late-fusion, i.e. by averaging each unimodal score. We report the results in Table 3.

With more than twice as much training data, almost every result improves when compared to Table 1. Note that late-fusion obtains similar results to our model for bimodal classification. However our model obtains an absolute 1.38% increase in accuracy to reach 94.80% for trimodal classification, as compared to 93.42% for late-fusion. With these additional unimodal data, image-tags pairs classification increased from 85.90% (Table 1) to 90.60%, an absolute improvement of 4.70%. To detail the results, we show the confusion matrix obtained by our trimodal classifier in Figure 4.



	amusement	anger	awe	contentment	disgust	excitement	fear	sadness
amusement	0.99	0.00	0.00	0.00	0.00	0.01	0.00	0.00
anger	0.00	0.85	0.00	0.04	0.00	0.00	0.12	0.00
awe	0.00	0.00	0.96	0.01	0.00	0.01	0.01	0.00
contentment	0.00	0.00	0.00	0.98	0.00	0.02	0.00	0.00
disgust	0.02	0.02	0.00	0.02	0.83	0.02	0.05	0.02
excitement	0.03	0.00	0.01	0.02	0.00	0.92	0.01	0.00
fear	0.00	0.00	0.00	0.00	0.00	0.11	0.89	0.00
sadness	0.00	0.00	0.00	0.02	0.00	0.02	0.03	0.94

Figure 4: Confusion matrix on Flickr Emotion test set produced by our trimodal classifier trained with additional unimodal data.

To further demonstrate the advantage of leveraging unimodal training data, we conduct an experiment with fewer multimodal examples on Flickr Emotion dataset. More specifically, we randomly pick a given fraction p of the multimodal subset of 8674 examples (see Figure 3), while the remaining portion $1 - p$ is divided in three equal parts: in the first tier we only keep the images, in the second the texts, and in the third the tags. All splits are stratified to ensure that each class is properly represented. We train our trimodal multitask model with the multimodal and the three unimodal subsets. As a comparison, a trimodal single-task model is trained only with the multimodal fraction. We report the accuracy for different proportions of multimodal data in Figure 5.

As we can see, even with very few multimodal training data, our model still obtains decent results thanks to the additional unimodal examples. When 1% of the training set is multimodal, the gap between the accuracy obtained by our model and the baseline is 19.5%. Interestingly, our model trained with 25% multimodal and 75% unimodal training data obtains a similar accuracy to that obtained with the baseline trained with 100% multimodal data. Since multimodal datasets can be hard to collect and they are therefore often small, this supports the possibility to augment them with unimodal data. This result suggests that with a model capable of being trained with unimodal data such as ours, it reduces the importance of training data composed of all the modalities for emotion recognition. This contrasts with the more common approach of keeping only multimodal examples from datasets.

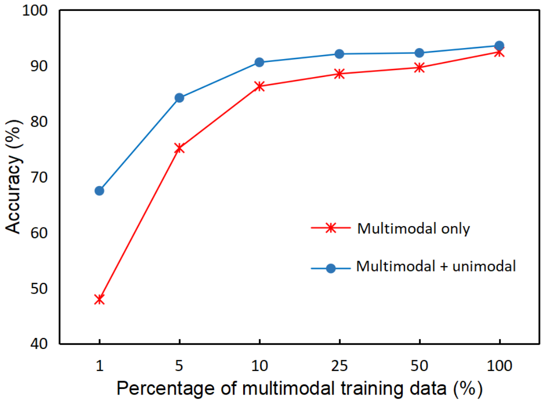


Figure 5: Classification accuracy for different proportions of multimodal training data.

5 DISCUSSION AND CONCLUSION

We explored the use of Flickr tags as an additional supervision for emotion recognition and sentiment analysis. The results supported that they can convey complementary and discriminative information, especially on the small Flickr Emotion dataset. Also, since texts and tags are optional when a user shares a message, we leveraged a multitask approach to avoid the problem of missing modality that occurs in real situations where all modalities are not always available. We showed that our model enables the presence of each input modality to be optional. When images, texts and tags are available, performance measures are significantly higher, but if one or two of them are absent, the model still obtains good results. In fact, the experiments even supported that this multitask training improved the generalization of multimodal classification. For instance the more common task of emotion recognition from image-text pairs improved by 4.8% in accuracy when the training was done with multitask learning. This benefit was more important on the small Flickr emotion dataset than on the large VSO dataset. Since overfitting is more problematic with small training sets, this is consistent with the idea that this framework acts as a regularization mechanism. Previous work already discussed that this multitask learning acts as a regularization mechanism by forcing each unimodal representation to be discriminative before multimodal fusion, which improves generalization performance [25].

We also went further by showing that the proposed multimodal multitask framework also enables the model to be trained with data that lack one or two modalities as shown by the results of Table 3. Adding these data further improved the results of most sub-tasks (i.e. unimodal, bimodal, trimodal classification).

We also showed in Figure 5 that even with very few multimodal training data, our model can compensate with unimodal examples to obtain decent results. This last observation supports the opportunity to augment multimodal datasets with unimodal examples instead of limiting the training set to multimodal examples.

More generally, the problem of handling unimodal data concerns all multimodal tasks that are confronted to the problem of missing modality at test time, one of the most important challenge in multimodal machine learning [1]. In this work we showed that a multitask approach can constitutes a viable solution for multimodal emotion recognition and sentiment analysis.

REFERENCES

[1] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2019), 423–443.

[2] Moshe Bar and Maital Neta. 2006. Humans prefer curved visual objects. *Psychological Science* 17, 8 (2006), 645–648.

[3] Fabiano M Belém, Jussara M Almeida, and Marcos A Gonçalves. 2017. A survey on tag recommendation methods. *Journal of the Association for Information Science and Technology* 68, 4 (2017), 830–844.

[4] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM International Conference on Multimedia*. ACM, 223–232.

[5] Erion Çano and Maurizio Morisio. 2017. Music Mood Dataset Creation Based on Last. fm Tags. In *2017 International Conference on Artificial Intelligence and Applications, Vienna, Austria*.

[6] Xingyue Chen, Yunhong Wang, and Qingjie Liu. 2017. Visual and textual sentiment analysis using deep fusion convolutional neural networks. In *Image Processing (ICIP), 2017 IEEE International Conference on*. IEEE, 1557–1561.

[7] Silvia Corchs, Elisabetta Fersini, and Francesca Gasparini. 2017. Ensemble learning on visual and textual data for social image emotion classification. *International Journal of Machine Learning and Cybernetics* (2017), 1–14.

[8] Chi Thang Duong, Remi Lebret, and Karl Aberer. 2017. Multimodal Classification for Analysing Social Media. *arXiv preprint arXiv:1708.02099* (2017).

[9] Michael Glodek, Stephan Tschechne, Georg Layher, Martin Schels, Tobias Brosch, Stefan Scherer, Markus Kächele, Miriam Schmidt, Heiko Neumann, Günther Palm, et al. 2011. Multiple classifier systems for the classification of audio-visual emotional states. In *Affective Computing and Intelligent Interaction*. Springer, 359–368.

[10] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. 2014. A multi-view embedding space for modeling internet images, tags, and their semantics. *International Journal of Computer Vision* 106, 2 (2014), 210–233.

[11] Alan Hanjalic. 2006. Extracting moods from pictures and sounds: Towards truly personalized TV. *IEEE Signal Processing Magazine* 23, 2 (2006), 90–100.

[12] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 1. 3.

[13] Marie Katsurai and Shin'ichi Satoh. 2016. Image sentiment analysis using latent correlations among visual, textual, and sentiment views. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2837–2841.

[14] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).

[15] Yifei Lu, Wei-Long Zheng, Binbin Li, and Bao-Liang Lu. 2015. Combining Eye Movements and EEG to Enhance Emotion Recognition.. In *International Joint Conference on Artificial Intelligence (IJCAI)*, Vol. 15. 1170–1176.

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

[17] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. 689–696.

[18] Mathieu Pagé Fortin and Brahim Chaib-draa. 2019. Multimodal Sentiment Analysis: A Multitask Learning Approach.. In *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods*.

[19] Lei Pang, Shiai Zhu, and Chong-Wah Ngo. 2015. Deep multimodal learning for affective analysis and retrieval. *IEEE Transactions on Multimedia* 17, 11 (2015), 2008–2020.

[20] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37 (2017), 98–125.

[21] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE, 439–448.

[22] Soujanya Poria, Amir Hussain, and Erik Cambria. 2018. *Multimodal Sentiment Analysis*. Springer International Publishing.

[23] Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098* (2017).

[24] Patricia Valdez and Albert Mehrabian. 1994. Effects of color on emotions. *Journal of Experimental Psychology: General* 123, 4 (1994), 394.

[25] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. 2018. CentralNet: a Multilayer Approach for Multimodal Fusion. *arXiv preprint arXiv:1808.07275* (2018).

[26] Johannes Wagner, Florian Lingenfelser, and Elisabeth André. 2015. Building a robust system for multimodal emotion recognition. In *Emotion Recognition: A Pattern Analysis Approach*. Wiley, 379–410.

[27] Hao Wang, Xingjian Shi, and Dit-Yan Yeung. 2015. Relational Stacked Denoising Autoencoder for Tag Recommendation.. In *Association for the Advancement of Artificial Intelligence (AAAI)*. 3052–3058.

[28] Fangzhao Wu and Yongfeng Huang. [n. d.]. Personalized microblog sentiment classification via multi-task learning. In *Association for the Advancement of Artificial Intelligence (AAAI)*. 3059âĂŞ–3065.

[29] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. Building a Large Scale Dataset for Image Emotion Recognition: The Fine Print and The Benchmark.. In *Association for the Advancement of Artificial Intelligence (AAAI)*. 308–314.

[30] Sicheng Zhao, Guiguang Ding, Qingming Huang, Tat-Seng Chua, Björn W Schuller, and Kurt Keutzer. 2018. Affective Image Content Analysis: A Comprehensive Survey.. In *International Joint Conference on Artificial Intelligence (IJCAI)*. 5534–5541.

[31] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun. 2014. Exploring principles-of-art features for image emotion recognition. In *Proceedings of the 22nd ACM International Conference on Multimedia*. ACM, 47–56.

[32] Sicheng Zhao, Hongxun Yao, Yue Gao, Rongrong Ji, and Guiguang Ding. 2017. Continuous probability distribution prediction of image emotions via multitask shared sparse regression. *IEEE Transactions on Multimedia* 19, 3 (2017), 632–645.

[33] Xuelin Zhu, Biwei Cao, Shuai Xu, Bo Liu, and Jiuxin Cao. 2019. Joint Visual-Textual Sentiment Analysis Based on Cross-Modality Attention Mechanism. In *International Conference on Multimedia Modeling*. Springer, 264–276.

[34] Chiara Zucco, Barbara Calabrese, and Mario Cannataro. 2017. Sentiment analysis and affective computing for depression monitoring. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 1988–1995.