Manga-MMTL: Multimodal Multitask Transfer Learning for Manga Character Analysis

Nhu-Van Nguyen, Christophe Rigaud, Arnaud Revel, Jean-Christophe Burie

Manga-MMTL: multimodal multitask transfer learning for manga character analysis*

Nhu-Van Nguyen[1] (✉)[0000−0003−2271−6918], Christophe Rigaud[1][0000−0003−0291−0078], Arnaud Revel[1][0000−0002−9498−392X], and Jean-Christophe Burie[1][0000−0001−7323−2855]

Laboratory L3i, SAIL joint lab
La Rochelle University, 17042 La Rochelle CEDEX 1, France

Abstract. In this paper, we introduce a new pipeline to learn manga character features with visual information and verbal information in manga image content. Combining these set of information is crucial to go further into comic book image understanding. However, learning feature representations from multiple modalities is not straightforward. We propose a multitask multimodal approach for effectively learning the feature of joint multimodal signals. To better leverage the verbal information, our method learn to memorize the content of manga albums by additionally using the album classification task. The experiments are carried out on Manga109 public dataset which contains the annotations for characters, text blocks, frame and album metadata. We show that manga character features learnt by the proposed method is better than all existing single-modal methods for two manga character analysis tasks.

Keywords: Manga image analysis · Multimodal learning · Auxiliary task learning · Transfer learning.

1 Introduction

Digital comics and manga (Japanese comics) wide spread culture, education and recreation all over the world. They were originally all printed but nowadays, their digital version is easier to transport and allows on-screen reading with computers and mobile devices. To deliver digital comics content with an accurate and user-friendly experience on all mediums, it might be necessary to adapt or augment the content [2]. This processing will help to create new digital services in order to retrieve very precise information in a corpus of images. For instance, comic character retrieval and identification would be useful as copyright owners need efficient and low-cost verification systems to assert their copyrights and detect possible plagiarisms [24]. In our paper, we use the term "character" as the person in comic books. From this point on, the term "character" in character

* This work is supported by the ResearchNational Agency (ANR) in the framework of the 2017 Lab-Com program (ANR 17-LCV2-0006-01)

retrieval, character classification, character clustering should be understood as "comic/manga character".

Current work [11,30] on comic/manga character analysis often rely on transfer learning [19]. When it is hard to perform a target task directly on comic character images due to the missing of labels, one can learn the character visual features basing on related tasks then use the learnt features to perform the target task. In these methods, a visual feature extractor (uni-modal) is learnt using the character classification task and then it is used to extract features to perform character retrieval, identification or clustering tasks.

However, single modality (e.g. graphics) is limited and can not profit from all the information we can extract or infer from static images. In comic/manga image analysis domain, recent works state that learning using only a single modality, as very often the image, can not cope with feature changes in some images and suggest to take advantage of another modality (text in their study) simultaneously with the first one [30]. Existing works have often ignored the implicit verbal information in the comic book images which are presented in form of speech text or narrative text. This source of verbal information is important, especially since the text recognition technology is strong nowadays.

Theoretically, multimodal analysis is able to retrieve more information and of an higher level than the uni-modal scenario, with an overall better performance. However, training from multiple modalities is actually hard and often results with lower performance than the best uni-modal model. The first reason is that it is easy to overfit: the learnt patterns from a train set that do not generalize to the target distribution [28]. Another reason is that one of the modality might interacts with the target task in a more complex way than the other [22]. For example, in our case, the image modality is relevant for characters retrieval/clustering because we can identify the character name and the character emotion by looking at the image. In contrast, the text modality is more adapted for album retrieval/clustering or character relationship analysis than the character retrieval/clustering. This is due to the text in each album that does not describe a specific character but instead tells a specific story involving many characters. While both modalities seem likely to be complementary, there are essential for other elements such as emotion analysis that can be jointly extracted from character face expression, speech balloon shape and speech text processing.

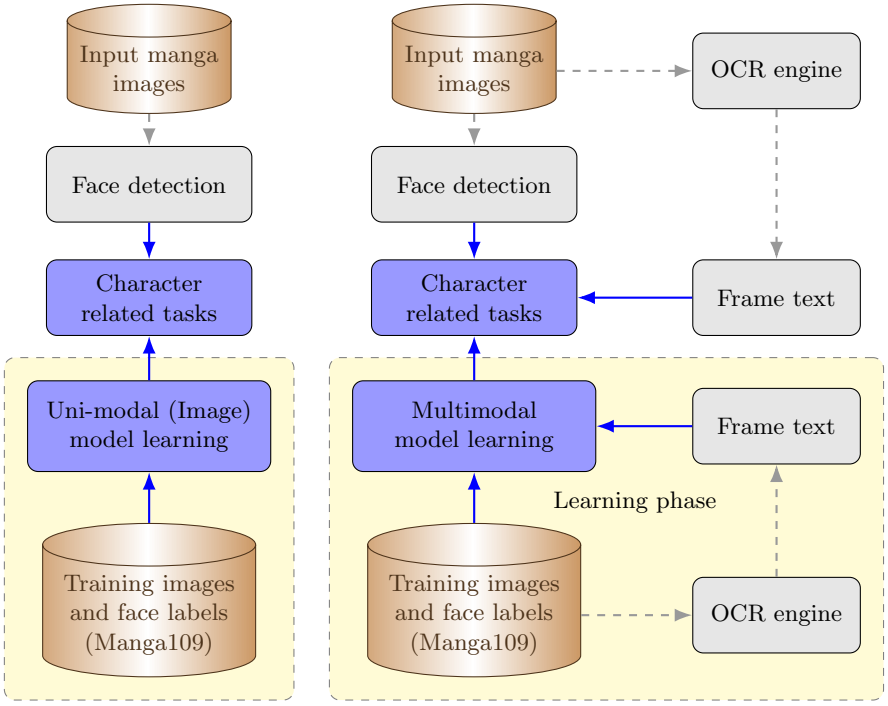
Multitasks analysis [22,18] is a powerful approach to train a deep model, especially when the target task, in a single task context, is difficult to train to differentiate between relevant and irrelevant features. In our work, we propose to use an auxiliary task to facilitate the multimodal training. The multitask multimodal methods we propose for learning the joint feature of multimodal signals (text and image) can be used subsequently in different analysis tasks for comic and manga image analysis such as character clustering, character retrieval and

emotion recognition.

The contributions of this paper are summarized as follows :

- Instead of using only visual information in comic/manga images, we propose a new pipeline which uses the implicit verbal information in the images to learn multimodal manga character features.
- Different methods for learning the comic/character feature are analysed. To achieve the best performance, we propose a self-supervised strategy with the album classification task. This strategy forces the model to remember the manga content which can improve the learnt multimodal features.
- The effectiveness of the learnt multimodal feature is demonstrated by transferring it to character retrieval and clustering tasks.

2 Related work



a. Current methods: Uni-modal manga character learning

b. Our pipeline: Multimodal manga character learning

Fig. 1: Current methods (a) compare with the proposed pipeline (b): Blue boxes and arrows show our contributions, other parts are not in the scope of this paper.

Our work focuses on comic character feature extraction to perform related tasks such as character retrieval, clustering and emotion recognition. Most of

the previous works have focused on these tasks independently and using only image-based features (uni-modal).

Regarding the character retrieval task, large handcrafted methods have started to be proposed in 2017 [12]. Later and similarly, convolution neural networks (CNN) and fine-tuning of two CNN on manga face images have been applied [15]. They used with-and-without screen-tone images to perform accurate query- and sketch- based image retrieval. Recently, the challenge of long-tail distribution data like manga character (face) images, inappropriate for using the usual cross-entropy softmax loss function, have been tackled [11]. To do so, the authors proposed a dual loss from dual ring and dual adaptive re-weighting losses which improved the mean average precision score by 3.16%.

Few methods have been proposed for character (face) clustering. Tsubota et al. [25] proposed to fine tune a CNN model using deep metric learning (DML) in order to get help from domain knowledge such as page and frame information. They include pseudo positive and negative pairs into the training process and use k-means clustering given the number of characters for the final clustering. The method applies its learning stage for each specific album of comic to learn the character features before realizing the character clustering for the album.

Yanagisawa et al. [30] analysed different clustering algorithms on the learnt features after training a CNN model on the character classification task. The HDBSCAN clustering algorithm [3] was selected because of its better performance and no requirement of the number of cluster priors (unlike k-means). The authors mention that "learning using only images cannot cope with feature changes in some character face images" and suggest to "utilize word information contained in manga simultaneously with images".

This suggestion makes the link with text analysis and natural language processing techniques that can be combined for retrieving character names, relationships, emotions etc. In 2018, a survey mentions that "research has been done on emotion detection based on facial image and physiological signals such as electroencephalogram (EEG) while watching videos [...]. However, such research has not been conducted while reading comics." [2].

From our knowledge, there is only one study about recognizing emotion from text and facial expression in manga images. This work is part of an end-to-end comic book analysis and understanding method able to text-to-speech a comic book image with respect of character identification, gender, age and emotion of each speech balloon [29]. It combines several modalities using computer vision, natural language processing and speech processing techniques.

Multimodal approaches are usually developed for image/video classification and retrieval where multiple modalities are available explicitly and strong for the task [16]. From our knowledge, there is no multimodal work on comic/manga book images where the verbal information is included in the image modality. Moreover, the verbal information in comic/manga images is very weak for the

character classification task so it is not trivial to apply multimodal learning methods to comic book images. To overcome this issue, we may think about finding another suitable task to learn the features from both modalities, but it is not a good choice. Firstly, because it is hard to find a suitable task with necessary labels for learning. Secondly, the character classification task is naturally a relevant task to learn the comic character features. Another solution is to add a useful task which will guide the model to learn relevant features to boost the performance [26].

3 Proposed method

3.1 Overview

The digital comic book datasets are often composed by images partially annotated [1,7,10,17]. This is the reason why most of researches in comic book images focus on exploiting the visual information to extract the features and apply to other tasks in this domain. In reality, the comic book may come with the album metadata which includes some basic information such as the book name, the author, published date, a short summary of the story etc. Thus, the information from metadata is limited. Some datasets provide also text transcriptions based on an Optical Character Recognition system (OCR) [10,17].

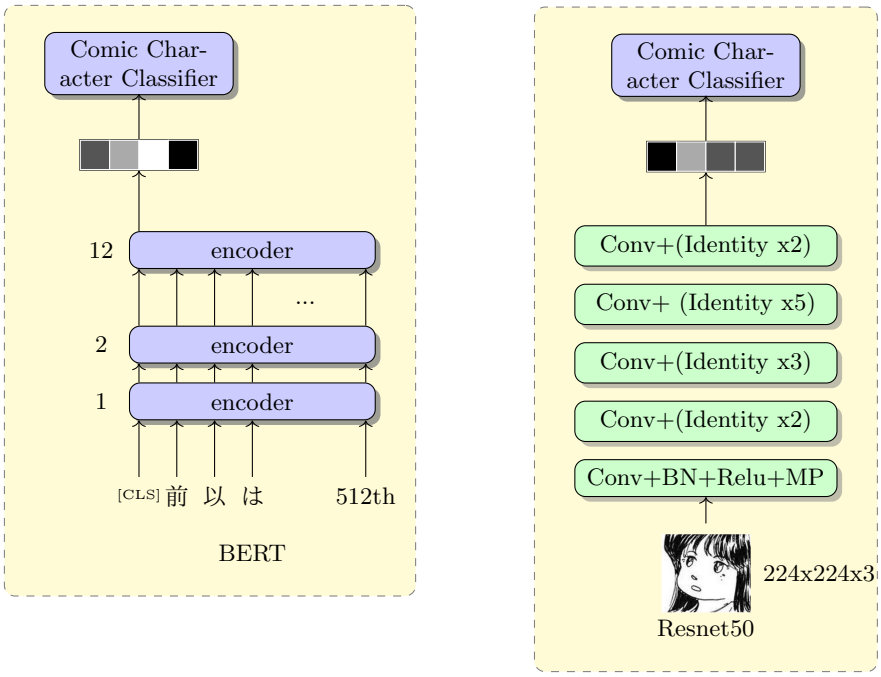
Concerning the feature extraction in comic book images, researchers often ignore the text in the comic book images, including speech text (the text that a character speaks) and narrative text (the text describing the flow of the situation in the comic story). However, text and images can be combined to form singular units of expressions in comics as mentioned by Cohn et al. [4], so it is very important to take into account the verbal information while performing comic analysis tasks, including character retrieval and recognition.

The high quality of text recognition technology nowadays gives us a big opportunity to take advantages of digitized text as a source of verbal information [14]. In this work, we consider the comic character face recognition and the comic text recognition to be performed previously. Hence, we have these two sources of information to study the related tasks. To simplify the experiments, and allow others to reproduce them, we directly use the ground-truth information of the face and the text from Manga109 dataset to learn the character feature and analyze our multimodal approach. To associate the face image with text and form a multimodal data sample, we associate a face image with all the text in the same comic frame.We assume that the text in the frame is related directly to the character whose the face is presented in the frame (speaking the text or receiving the speech). Our proposed pipeline is illustrated in Fig. 1.

Assuming that we have a set of comic character samples where each sample contains two modalities *text* and *image* defined as T and V, respectively. The existing works [30,11,18] use V as the only information to learn or manually

create the feature extractor which is then used to extract the features and realize the character related task such as multimodal and/or cross-modal character retrieval, character clustering, character recognition, character emotion recognition, etc. All of the learning approaches use character classification task to train the feature extractor which aims at classifying all samples (face image and associated text) in the database of multiple albums and authors.

In our work, the objective is to learn the character extractor from both T and V which are then used to perform the character related tasks. In single-modality setting, we learn the two modalities separately: the verbal feature extractor $E_T : T \rightarrow F_T$ and the visual feature extractor $E_V : V \rightarrow F_V$. In multimodality setting, we can combine the two sources of information to learn a combined feature extractor $E_{mix} : (T, V) \rightarrow F_{mix}$.



a. Text classification by BERT [5] b. Image classification by Resnet

Fig. 2: Learning two feature extractors separately

In the next sections, we analyze different strategies corresponding to several settings to learn from the two modalities of comic book images. We will focus on the late fusion strategy because in this case the early fusion strategy does not work well, which will be explained later. We will explain and prove by empirical experimentation that the multimodal training using the character classification does not work. Then we will propose a new way to learn the multimodal feature extractor.

3.2 Learning uni-modal feature extractor

In the single modality setting, we learn the feature representations separately (Fig. 2). Given a training set of a modality, for example the visual modal V: $C_v = \{X_1^v, ..., X_n^v, y_1, ..., y_n\}$, where X_i^v is the visual information of *i-th* training example and y_i is its true label (the character identity), we can learn the single modality feature extractor by training a neural network with respect to the classification task. The loss function can then be the cross entropy loss, an ubiquitous loss in modern deep neural networks.

$$L_v(C(\theta(X^v), y) = -\frac{1}{N}\left(\sum_{i=1}^{N} \log p(y = y_i|X_i^v)\right) \tag{1}$$

where $\theta^v(X^v)$ is usually a deep network and C is a classifier, typically one or more fully-connected (FC) layers with a parameter θ_c^v.

The same way, we can train another feature extractor on the text modality by optimizing the following cross entropy loss $L_t(C(\theta(X^t), y)$.

3.3 Learning multimodal feature extractor

A simple multimodal method is to jointly learn two modalities by training a multimodal feature extractor, based on character classification task with the late fusion technique. In neural network architecture, two modalities are processed by two different deep networks θ_X^t and θ_X^v, and then their outputs are fused and passed to a classifier C composed of FC layers with parameter θ_c. The loss function of the deep network is the cross entropy loss:

$$L_{mix}(C(\theta_X^t \oplus \theta_X^v, y)) = -\frac{1}{N}\left(\sum_{i=1}^{N} \log p(y = y_i|X_i^v, X_i^t)\right) \tag{2}$$

where $\oplus$ denotes a fusion operator (e.g. max, concatenation, addition, etc.).

One can argue that this multimodal feature extractor must be better than or equal to the best single-modal feature extractor because in the worst case the model will learn to mute all the parameters in one modality (θ_X^t or θ_X^v) and then becomes a single-modal model.

In our case, the multimodal approach is also worst than the single-modal based on the visual information (see Section 4). This bad performance is due to two problems. According to Wang et al. [28], the main problem is overfitting. When training any multimodal model, the deep network has nearly twice as many parameters as a single-modal model, and one may suspect that the overfitting is caused by the increased number of parameters.

We understand that another important problem comes from the target classification task used to train the multimodal deep network. All mentioned approaches do not work if the signals in the two modalities are very different in

term of the target classification task: one modality is dominant for the task while another one is weak or complex. For comic character analysis tasks, the signal in face image are very strong, we can identify the character name and the character emotion by looking at the image. The signal in the text is however more complex. To decide if the two texts come from the same character (in the context of a multiple-albums dataset), one needs to integrate a lot of knowledge to reason, including for each album, the story, the relationship between characters, names, places, events etc. So when we train the multimodal deep network with the two modalities on the character classification task, the model parameter θ_X^t often stucks at a bad local maximum because of the weak and complex signal in the text, which leads to the sub-optimum of the parameter θ_c. One can check the signal of a modality by training a single-modal model on that modality to see the performance. In our case, the performance of text-modal model on the character classification task is bad (see Section 4).

To overcome this difficulty and benefit from the advantages of the text modality, we propose to learn the text features by a self-supervised learning method. As mentioned before, we have to read all the manga albums to identify the character from a piece of text, so we are going to train the model to remember the contents of all the albums in the training dataset. In this work, we propose to use an already-available information to learn the strong signal in the text: *the album name*. The text signal is strong to identify the albums as the text in each book tells a different story. Training the model with the album classification task is an effective way to learn relevant information from manga albums. It is also helpful to differentiate the characters identity because the characters from different comic stories are different. We know the album names of each image from the comic book images dataset [6].

In the next section, we propose a multimodal multi-task learning (M-MTL) architecture which can learn a good feature extractor for character related tasks from the two modalities with two tasks: character classification and album classification.

3.4 Manga M-MTL modal for comic character analysis tasks

We propose a new multimodal architecture with three outputs, two for the main task character classification and one for the auxiliary album classification task. We use popular techniques to reduce overfitting such as Dropout [23], EarlyStopping, and Transfer learning as suggested in [28].

The Manga M-MTL architecture is shown in Fig. 3. It consists in two different feature components: the image network for learning visual features and the text network for learning the textual features. Three classifiers are added into our architecture. Each classifier, composed of FC layers, is used to perform the character classification task and the album classification task based on the visual feature, the textual feature, and the combined (concatenated or averaged) features. The character classification loss for the combined features is the main

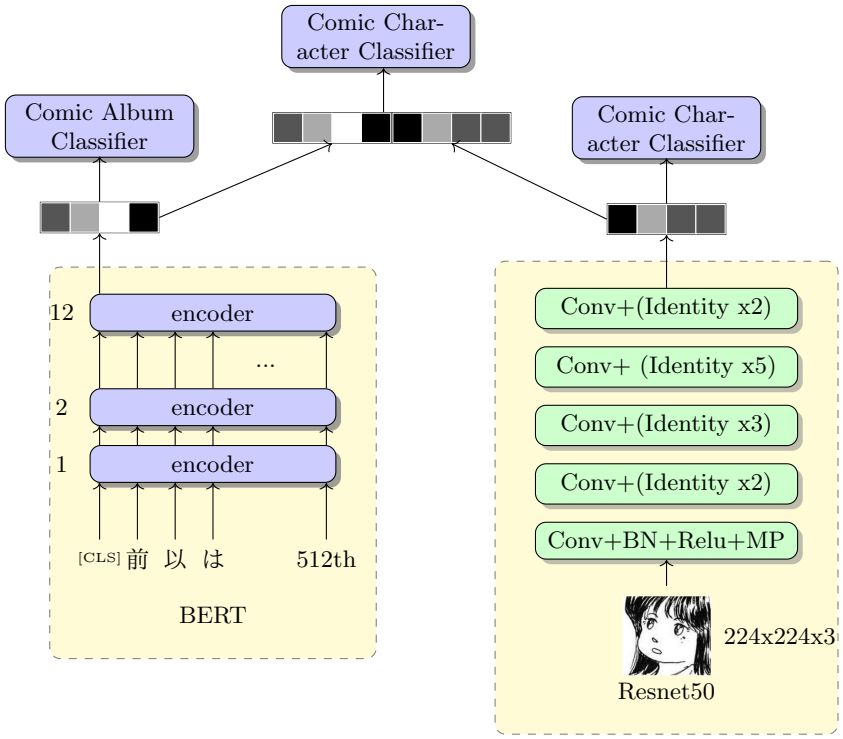


Fig. 3: Manga M-MTL architecture: multimodal learning with the auxiliary task.

loss of the model, while the two other losses are the auxiliary losses which can help the former to learn better combined features.

Text network: The transfer learning is well studied in natural language processing (NLP) domain. Strong pre-trained models such as BERT [5], GPT2 [20] are successfully used to solve many sub-tasks. We use the popular architecture Transformer and pre-trained weights BERT for the text network. Transformer is a strong architecture which can learn the feature representations of text based on its famous attention mechanism. The architecture we used is the BERT base architecture which has 12 attention layers, 768-dim hidden units, 12 heads.

Image network: As for the text network, we adopt a common CNN architecture in the vision domain for our image network. Resnet [8] is widely used in the vision domain because of the capacity of avoiding the gradient vanishing and the fast computation capacity. We use the 50 layers Resnet architecture.

3.5 Implementation details

Training feature extractor models We use WordPiece tokenizer [5] to process the text input before feeding them into the transformer network. All face images are resized to (224x224). We use image augmentation during training, including random crop (192x192) and rotate. The pre-trained weights using Japanese Wikipedia (performed by the Inui Laboratory, Tohoku University) are used for the BERT network. The ImageNet pre-trained Resnet weights are used for the CNN network. Dropout is applied to the BERT output layer and the Resnet output layer with a probability of 0.2. Early stopping is considered within 20 epochs. We train our models using SGD optimizer with a momentum of 0.9, a learning rate of 0.0001 for 100 epochs. For the multimodal models, we use the pre-trained weights of uni-modal models and the concatenation as the fusion operator, which is the best choice according to our experiments.

Extracting features to use in manga character related tasks We extract the second-last FC layer in the two classifiers: album classifier and the final character classifier. Then we apply L2 regularization before concatenate these two features for character clustering and character retrieval tasks. The character clustering task is applied for each album as described by Yanagisawa et al. [30]. It is a measure of the intra-album character identification capacity of the learnt features. The character retrieval task is applied for all character in the test albums which gives a measure of the inter-album character identification capacity of the learnt feature, as presented by Li et al. [11].

For realizing the character clustering, we need to reduce the dimensions of the extracted multimodal features of manga characters to reduce the computation cost in the clustering process. We follow the same setting and the parameters configuration as described by Yanagisawa et al. [30] where the dimension reduction algorithm is UMAP [13] and the clustering algorithm is HDBSCAN [3].

To do the character retrieval, we rank the retrieval results by the cosine similarity of the extracted multimodal features, same as described by Li et al. [11].

4 Results

4.1 Experiment protocol

Dataset: We use the large-scale Manga109 dataset which is the largest comic/manga dataset with ground truth information for characters [1]. This dataset consists of 109 manga albums, 118,715 faces, 147,918 texts and 103,900 frames. Following the work of Yanagisawa et al. [30], we make the training dataset of face images and text of characters appearing in 83 manga each drawn by different authors and remaining 26 albums are used for testing. The characters who appear less than 10 times in each book are ignored. The total manga character in this training dataset is 76,653. It is then divided into two sets for training and validation.

The 26 test albums consist of 26,653 face images with or without associated texts (see Table 1). This multimodal character samples are used for testing the two tasks: characters clustering and character retrieval. For character retrieval, we use 2000 samples as queries and the 24,653 samples as retrieval dataset. For character clustering, we use the same 11 test albums as in the reference work [30] (a subset of the 26 test albums). We will open all the information for the community[1].

Table 1: Statistics of the training and test sets in our setting

#train album	#test album	#train face	#test face	#train character	#test character
83	26	76,653	26,653	1,114	319

Metric for evaluation: We evaluate the clustering performance by the V-measure, ARI, and AMI metrics [9,21,27]. These values all range from 0.0 to 1.0, and the better clustering result, the higher the value is. For each metric, we average the values over 11 selected test albums as used in the work of Yanagisawa et al.[30]. We evaluate the character retrieval performance by rank-1, rank-5 precision, and mean Average Precision (mAP) as in the work of Li et al.[11].

Different training models: We have trained 5 models with different configuration as shown in Table 2. Basing on these trained models, in the next sections, we will 1) compare the performance of multimodal models and uni-modal models on two manga character tasks: *character clustering* and *character retrieval* (including other feature extraction models of existing work [30,11]). 2) show that the manually concatenation of different learnt features (multimodal or uni-modal) is a simple yet effective method to improve the subsequent tasks in character comic analysis.

Table 2: Our different manga character feature models.

Model name	Image	Text	Training objectives (tasks)
M1 (Uni-modal Image)	✓		Manga character classification
M2 (Uni-modal Text)		✓	Manga character classification
M3 (Uni-modal Text)		✓	Manga album classification
M4 (Multimodal)	✓	✓	Manga character classification
M5 (Multimodal with auxiliary task)	✓	✓	Manga character classification and album classification

4.2 Multimodal vs. Uni-modal models

Character clustering We compare six different learnt features, using our five trained models and the uni-modal model reported in the work of Yanagisawa et

[1] https://gitlab.univ-lr.fr/nnguye02/paper-icdar2021-mmtl

al. [30], where the authors have used only image modality to train a feature extractor. In our experiments, we re-implement the method proposed in the work [30] and train the model M1 in the same way as described.

In Table 3, the multimodal multitask model (M5) outperforms all other models with big improvements (3.8% of ARI, 4.61% of AMI and 3.84% of V-measure). The M2 model using text alone performs poorly so it is not surprise that concatenating these noisy features to the features in the image domain will degrade the performance of the system (M4) compared to the image-only model M1. These results also show that our proposed auxiliary task is important, compared to the base task: it helps M5 to improve by 12.9% of ARI, 11.01% of AMI and 10.08% of V-measure, compared to M4.

The text-modal models (M2, M3) are the worst models but we can see that training text with album classification task (M3) can learn better features than training it with character classification task (M2). Intuitively, we can find that it is hard to distinguish manga characters inter-albums using only text. While text is good at distinguishing albums, we understand that the specific text features of each albums can be used to filter manga character inter-albums and intra-album.

Table 3: Character clustering results

Method	ARI	AMI	V-measure
Method in [30] (Image only)	0.5063	0.6160	0.6381
Image only (M1)	0.5104	0.5998	0.6225
Text only (M2)	0.0202	0.0478	0.0639
Text only (M3)	0.0678	0.2288	0.2919
Multimodal features (M4)	0.4071	0.5114	0.5383
Multimodal features (M5)	**0.5443**	**0.6621**	**0.6765**

Character retrieval We compare the results of different feature learning models for the character retrieval task. Li et al. [11] have applied their learnt features to the character retrieval task but they have not provided their list of 80 training albums yet. Therefore, we use the same five models presented previously (from M1 to M5). These models are trained using the list of 83 albums from the reference work [30] and the setting presented in Section 4.1. Although it is not the same as the training set of Li et al. [11] but as mentioned in the work of Yanagisawa et al. [30], these 83 albums come from 83 authors who are different from testing albums so it is safe to add the result of Li et al. [11] into our comparison table for reference.

In Table 4, we can see that the multimodal multitask model outperforms other uni-modal and multimodal models, thanks to the text modality. The multimodal learning without the auxiliary task is worse than the best uni-modal model (M1). This result confirms again that the base task character classification is not suitable to learn the verbal feature and our proposed auxiliary task can greatly improve the performance of the multimodal feature. Compare to the

Table 4: Character retrieval results. (*) Results in [11] is tested in unknown subset of Manga109 and a different setting compared to our experiments.

Method	rank-1(%)	rank-5(%)	mAP(%)
[11] (*)	70.55	84.30	38.88
Image only (M1)	71.70	87.65	39.15
Text only (M2)	0.30	2.05	0.79
Text only (M3)	68.85	83.25	24.62
Multimodal features (M4)	63.40	82.10	28.13
Multimodal features (M5)	**85.78**	**94.65**	**43.17**

best uni-modal model M1, our multimodal model M5 gives big improvements (15.23% of rank@1, 10.35% of rank@5 and 4.29% of mAP).

The signal in verbal information is weak to train a character classifier but strong to train an album classifier. It is confirmed by the performance of text-modal models M2 where we learnt almost zero knowledge from text using the character classification task (0.79% mAP). We understand that the verbal information is possibly good to distinguish character intra-album but it is weak to distinguish manga character inter-albums because the number of characters in an album is small while it is big in a comic/manga dataset. Using the auxiliary album classification task forces the model to remember the contents of manga albums so it can learn to distinguish the albums which can be used to distinguish characters inter-albums. It is to worth noting that, in the multitask multimodal model, the text modal is used to optimize the features by training both album and character classification tasks so it can learn to distinguish both characters and albums.

4.3 Boosting with simple learnt features combination

A part from the features learnt from end-to-end trainings, one can think about manually combine any different learnt features then apply to the related tasks. We have experimented different combinations of the learnt features to further analyse the importance and the relation of learnt features. We use the concatenation operator as the baseline combination method for different feature vectors. The setting for this evaluation keeps intact as in the previous evaluations.

We can see in Table 5 the results for character clustering task and character retrieval tasks follow the same pattern. The best complementary pair of learnt features is M3 and M5 for both tasks. For example, this combination gives improvements in all three metrics for character retrieval task, compared to the best single learnt feature (M5) : 4.27% of rank-1, 2.2% of rank-5, and 5.16% of mAP. The combination of any two uni-modal features is worse than the best multimodal feature of M5. It is easy to see that M2 feature is very weak so it

Table 5: Character retrieval and character clustering results: concatenations of different learnt features.

	Character clustering			Character retrieval		
Combination	ARI	AMI	V-measure	rank-1(%)	rank-5(%)	mAP(%)
Best single (M5)	**0.5443**	**0.6621**	**0.6765**	**85.78**	**94.65**	**43.17**
M1 + M2	0.4277	0.5278	0.5543	71.00	87.25	26.55
M1 + M3	0.5439	0.6540	0.6742	83.95	94.25	40.51
M1 + M4	0.5239	0.6158	0.6394	76.25	89.90	38.75
M1 + M5	0.5074	0.6188	0.6369	77.90	91.00	40.30
M3 + M4	0.4817	0.5868	0.6098	83.40	94.45	41.83
M3 + M5	**0.5853**	**0.6791**	**0.6965**	**90.05**	**96.85**	**49.33**
M1+ M3 + M5	0.4999	0.6197	0.6419	85.90	95.20	44.05

will pull down the performance of other features. Uni-modal text feature M3 gives improvements for any other learnt features, even the learnt multimodal feature ones (see M1+M3, M3+M4 or M3+M5). Our multimodal multitask feature (M5) is the best compared to single models but it also gives the best result while combining with the uni-modal text feature (M3). Combining more features does not help, as illustrated in the last row: the combination of 3 best features M1, M3 and M5 is worse than the combination of M3 and M5.

5 Conclusion

We have proposed a pipeline to leverage the implicit verbal information in the manga images. Our analysis shows that the multimodal manga character feature is better than the best uni-modal feature (visual feature). The results show that the self-supervised learning strategy with the album classification task is important. Without learning to memorize the content of the manga albums, the multimodal model gives lower performance than the best uni-modal feature which is trained using image model only.

One effective method to leverage both visual and verbal information is the simple manual combination of features from multiple single models. We have shown that this technique gives some improvements of the performance in character clustering/retrieval tasks. Therefore, the multimodal learning in this domain still need further studies. And more character related tasks need to be analysed in order to get more insights of multimodal features in this comic/manga domain, such as the emotion recognition tasks.

References

1. Aizawa, K., Fujimoto, A., Otsubo, A., Ogawa, T., Matsui, Y., Tsubota, K., Ikuta, H.: Building a manga dataset "Manga109" with annotations for multimedia applications. IEEE MultiMedia **27**(2), 8–18 (2020)

2. Augereau, O., Iwata, M., Kise, K.: A survey of comics research in computer science. J. Imaging **4**(7), 87 (2018). https://doi.org/10.3390/jimaging4070087, `https://doi.org/10.3390/jimaging4070087`
3. Campello, R.J.G.B., Moulavi, D., Sander, J.: Density-based clustering based on hierarchical density estimates. In: Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G. (eds.) Advances in Knowledge Discovery and Data Mining, 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part II. Lecture Notes in Computer Science, vol. 7819, pp. 160–172. Springer (2013)
4. Cohn, N.: Comics, linguistics, and visual language: The past and future of a field. In: Linguistics and the Study of Comics, pp. 92–118. Springer (2012)
5. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics (2019)
6. Fujimoto, A., Ogawa, T., Yamamoto, K., Matsui, Y., Yamasaki, T., Aizawa, K.: Manga109 dataset and creation of metadata. In: Proceedings of the 1st international workshop on comics analysis, processing and understanding. pp. 1–5 (2016)
7. Guérin, C., Rigaud, C., Mercier, A., Ammar-Boudjelal, F., Bertet, K., Bouju, A., Burie, J.C., Louis, G., Ogier, J.M., Revel, A.: eBDtheque: a representative database of comics. In: Proc. of the 12th Int. Conf. on Doc. Ana. and Rec. (ICDAR). pp. 1145–1149 (2013)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR). pp. 770–778 (2016)
9. Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification **2**, 193–218 (1985)
10. Iyyer, M., Manjunatha, V., Guha, A., Vyas, Y., Boyd-Graber, J.L., III, H.D., Davis, L.S.: The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 6478–6487. IEEE Computer Society (2017)
11. Li, Y., Wang, Y., Qin, X., Tang, Z.: Dual loss for manga character recognition with imbalanced training data. In: 2020 25th ICPR International Conference on Pattern Recognition (ICPR). pp. 2166–2171 (2020)
12. Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T., Aizawa, K.: Sketch-based manga retrieval using Manga109 dataset. Multimedia Tools and Applications **76**(20), 21811–21838 (2017)
13. McInnes, L., Healy, J.: Umap: Uniform manifold approximation and projection for dimension reduction. ArXiv **abs/1802.03426** (2018)
14. Memon, J., Sami, M., Khan, R.A.: Handwritten optical character recognition : A comprehensive systematic literature review. IEEE Access **8**, 142642–142668 (2020)
15. Narita, R., Tsubota, K., Yamasaki, T., Aizawa, K.: Sketch-based manga retrieval using deep features. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 3, pp. 49–53. IEEE (2017)
16. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of the 28th Int. Conf. on Machine Learning. p. 689–696. ICML'11, Omnipress, Madison, WI, USA (2011)
17. Nguyen, N., Rigaud, C., Burie, J.: Digital comics image indexing based on deep learning. J. Imaging **4**(7), 89 (2018)

18. Nguyen, N., Rigaud, C., Burie, J.: Comic MTL: optimized multi-task learning for comic book image analysis. Int. J. Document Anal. Recognit. **22**(3), 265–284 (2019)
19. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. on Knowl. and Data Eng. **22**(10), 1345–1359 (Oct 2010)
20. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2019)
21. Rosenberg, A., Hirschberg, J.: V-measure: A conditional entropy-based external cluster evaluation measure. In: EMNLP-CoNLL. pp. 410–420. Prague, Czech Republic (Jun 2007)
22. Ruder, S.: An overview of multi-task learning in deep neural networks. ArXiv **abs/1706.05098** (2017)
23. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research **15**(56), 1929–1958 (2014)
24. Sun, W., Kise, K.: Detection of exact and similar partial copies for copyright protection of manga. Int. J. Document Anal. Recognit. **16**(4), 331–349 (2013)
25. Tsubota, K., Ogawa, T., Yamasaki, T., Aizawa, K.: Adaptation of manga face representation for accurate clustering. In: SIGGRAPH Asia Posters, pp. 1–2 (2018)
26. Vafaeikia, P., Namdar, K., Khalvati, F.: A brief review of deep multi-task learning and auxiliary task learning. ArXiv **abs/2007.01126** (2020)
27. Vinh, N.X., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. J. Mach. Learn. Res. **11**, 2837–2854 (Dec 2010)
28. Wang, W., Tran, D., Feiszli, M.: What makes training multi-modal classification networks hard? In: IEEE Conf. on Comp. Vis. and Pat. Recognition (CVPR). pp. 12692–12702. IEEE Computer Society, Los Alamitos, CA, USA (jun 2020)
29. Wang, Y., Wang, W., Liang, W., Yu, L.F.: Comic-guided speech synthesis. ACM Transactions on Graphics (TOG) **38**(6), 1–14 (2019)
30. Yanagisawa, H., Kyogoku, K., Ravi, J., Watanabe, H.: Automatic classification of manga characters using density-based clustering. In: Int. Work. on Ad. Im. Tech. (IWAIT) 2020. vol. 11515, p. 115150F. Int. Society for Optics and Photonics (2020)