Visual-Textual Late Semantic Fusion Using Deep Neural Network for Document Categorization

Cheng Wang(✉), Haojin Yang, and Christoph Meinel

Hasso Plattner Institute, University of Potsdam, Prof.-Dr.-Helmert-Str. 2-3,
14482 Potsdam, Germany
{cheng.wang,haojin.yang,christoph.meinel}@hpi.de

Abstract. Multi-modality fusion has recently drawn much attention due to the fast increasing of multimedia data. Document that consists of multiple modalities i.e. image, text and video, can be better understood by machines if information from different modalities semantically combined. In this paper, we propose to fuse image and text information with deep neural network (DNN) based approach. By jointly fusing visual-textual feature and taking the correlation between image and text into account, fusion features can be learned for representing document. We investigated the fusion features on document categorization, found that DNN-based fusion outperforms mainstream algorithms include K-Nearest Neighbor(KNN), Support Vector Machine (SVM) and Naive Bayes (NB) and 3-layer Neural Network (3L-NN) in both early and late fusion strategies.

Keywords: Categorization · Semantic feature · Late fusion · Deep neural network

1 Introduction

Over the past decade, the tremendous increasing of multimedia data (e.g.image, audio and video) brings difficulties to information processing. Traditional approach for representation learning, classification and retrieval tasks usually focus on singe modality. However, in reality, we receive data from different information channels, one of the most common scenarios is image-text paired document. It is worth to note that different data modalities actually carry different information at different semantic levels. As shown in Fig. 1, an example document which contains an image and a loosely related descriptive text. If image and text information can be semantically fused, the more expressive and representative features can be learned for representing this document, and further improve multimodal document classification accuracy. Realizing the importance of multimodal information, in this work, we propose to address this problem by fusing visual and textual information with deep neural network. Multi-modality joint modeling is an open problem in bridging "semantic gap" across modalities. The procedure for

S. Arik et al. (Eds.): ICONIP 2015, Part I, LNCS 9489, pp. 662–670, 2015.
DOI: 10.1007/978-3-319-26532-2_73

"Though adult lions have no natural predators, evidence suggests that the majority die violently from humans or other lions.Schaller, p. 183 This is particularly true of male lions, who, as the main defenders of the pride, are more likely to come into aggressive contact with rival males..." (—-from Wikipedia)

Fig. 1. An example document that paired with an image and a descriptive text

multimodal data modeling generally falls into two stages: (1) modality representation and (2) correlation learning. In modality representation, one popular approach for image representation is to represent images as "bag-of-visual-words" (BOVW) using scale-invariant feature transform (SIFT) [10] or Dense SIFT [16] descriptor. In text representation, text is represented as topic feature that derived from Latent Dirichet Allocation [2]. Recently, many approaches have been proposed to explore the correlation between different modalities, including Canonical Correlation Analysis (CCA) [14], Semantic Correlation Match (SCM)[12], Cross-Modal Topic Correlations (CMTC) [19].

Unfortunately, the problem of fusing and combining different modalities was rarely discussed for multimedia data classification. In this paper, from different perspective, we focus on multimodal fusion problem [1]. In [3] St. Clinchant et al. proposed semantic combination approach for late fusion and image re-ranking in multimedia retrieval. D.Liu et al. [9] proposed Sample Specific Late Fusion (SSLF) method, which learns the optimal sample-specific fusion weights and enforces the positive sample have the highest fusion scores. In [17] deep neural network has been proved effective at fusion video keyframe and audio information for video classification. Considering the powerful capability of late fusion in areas such as video analysis [18], image retrieval [5] and object recognition [13].

Our work is distinguished from previous works in two aspects. First, we investigated deep convolutional neural network(CNN) features as image feature, this is motivated by recent success of deep CNN feature in addressing various research questions such as speech recognition [6], image classification [8] and multimodal learning [11]. Compare to commonly used SIFT feature, we prove that deep CNN features are more robust and representative in multi-modality fusion. Second, we propose to use deep neural network (DNN) to capture the highly non-linear dependency between different modalities, besides, late fusion with linear interpolation rule is adopted to capture the semantic contribution of image and text. Our contributions can be summarized as follows:

1. We propose to represent image and text to higher level feature using deep CNN feature and topic feature respectively.
2. We propose a novel approach to learn visual-textual fusion feature, which is seen as a unified representation for document categorization.
3. Extensive experiments and discussion were provided to show the effectiveness of DNN based late semantic fusion.

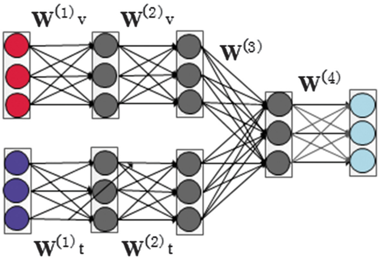


Fig. 2. DNN late fusion framework. Red nodes are visual (image) feature inputs and blue nodes are textual (text) feature inputs. The visual-textual fusion feature can be extracted from the output of 4^{th} layer (Color figure online).

The remainder of this paper is organized as follows. Section 2 states proposed approach for late semantic fusion and then we describe our implementation details in Sect. 3. Section 4 presents the experimental evaluation which illustrates the effectiveness of DNN-based late semantic fusion. Section 5 concludes this work.

2 DNN Late Semantic Fusion

This section introduces proposed DNN late semantic fusion. Given a set of N documents $S = \{D_n\}, \forall n = 1, 2...N$, where D_i is image-text paired document. We extracted the deep CNN feature and topic feature for each document D_n, then document D_n can be represented as $D_n = \{I_n, T_n\}$, $I_n \in \mathbb{R}^{d_i}$, $T_n \in \mathbb{R}^{d_t}$ at feature level, where d_i and d_t are the dimensionality of visual and textual feature respectively. Traditionally, if we combine visual feature I_n and textual feature T_n at feature level, called early fusion, formulated as

$$F_{(n)} = \alpha_v f_r(I_n) + (1 - \alpha_v)\, f_r(T_n),\ \forall n = 1, 2, .., N, r \in [0, 1] \tag{1}$$

where $F_{(n)}$ is early fused feature which is used to represent given document D_n and $f_r(\cdot)$ is normalization operator. $\alpha_v (0 \leq \alpha_v \leq 1)$ and 1-α_v denotes the fusion weight of visual and textual feature respectively. Another common approach is depicted in Eq. (2) called late fusion that performs fusion at decision level by combining the prediction scores of M pre-trained classifiers C_m.

$$P_{(n)} = \alpha_v C_m(I_n) + (1 - \alpha_v)\, C_m(T_n), \forall n = 1, .., N, \forall m = 1, .., M \tag{2}$$

In both approaches, α_v is usually assigned according to empirical experiments for demonstrating the importance of individual feature or classifier. Unfortunately, both fusion strategies do not take the correlations between visual and textual feature into account. A good fusion approach should consider the underlying shared semantic correlation between different modalities and take the advantage of the complementarity of modalities. To address the problem, besides heuristically assign α_v from 0 to 1 to capture the semantic contribution of each modality, we also learn latent fusion weights using deep neural network to capture the relationships across image and text. To achieve this goal, we propose a DNN fusion architecture which is shown as in Fig. 2. For a given single training sample $\{I_n, T_n, Y_n\}$, where I_n and T_n are input image and text feature respectively, Y_n is ground truth category label. The final output the global network can be represented as

$$\begin{cases} \hat{Y}^{(5)} = g^{(5)}(\hat{Y}^{(4)} W^{(4)} + b^{(4)}) \\ \hat{Y}^{(4)} = f^{(4)}((\alpha_v P_v + (1-\alpha_v) P_t) W^{(3)} + b^{(3)}) \end{cases} \tag{3}$$

where $\hat{Y}^{(l)}$ is the output of l^{th} layer and $W^{(l)}$ denotes the weights that connect to $(l-1)^{th}$ layer(also see from Fig. 2). $g(\cdot)$ and $f(\cdot)$ are activation functions and $b^{(l)}$ is bias item corresponding to l^{th} layer. P_v and P_t are prediction scores that computed from input feature I_n and T_n by

$$\begin{cases} P_v = f^{(3)}[f^{(2)}(I_n W_v^{(1)} + b_v^{(1)}) W_v^{(2)} + b_v^{(2)}] \\ P_t = f^{(3)}[f^{(2)}(T_n W_t^{(1)} + b_t^{(1)}) W_t^{(2)} + b_t^{(2)}] \end{cases} \tag{4}$$

We unitized sigmoid function $f^{(2)}(x)$=$f^{(3)}(x)$=$f^{(4)}(x)$=$\frac{1}{1+e^{-x}}$ and softmax function $g^{(5)}(x)$=$e^{(x-\varepsilon)} / \sum_{k=1}^{K} e^{(x_k-\varepsilon)}$ where $\varepsilon = max(x_k)$. To learn optimal weight set $\mathbf{W}$=$\{W^{(l)}\}$ and $\mathbf{b}$=$\{b^{(l)}\}$, $\forall l = 1, 2, 3, 4$, with all training samples, the objective is to minimize following loss function

$$\underset{\mathbf{W},\mathbf{b}}{\text{argmin}} \quad \frac{1}{2N} \sum_{n=1}^{N} \parallel \hat{Y}_n^{(5)} - Y_n \parallel^2 + \frac{\lambda}{2} \sum_{l=1}^{L-1} \parallel \boldsymbol{W_l} \parallel^2 \tag{5}$$

Where the second part is weight decay item for preventing overfitting. In learning procedures, the weights $W_m^{(1)}$ and $W_m^{(2)}$, m={v,t} are first learned by intra-modality training. Those weights can be regarded as local weights for achieving better prediction results. $W^{(3)}$ and $W^{(4)}$ are learned globally by fusing the scores of predicting image and text feature. The output of the 4^{th} layer are fusion features which combines visual and textual predictions.

3 Implementation

Our experimental configuration are Ubuntu 12.04, Nvidia GTX 780 GPU with 3G memory for image feature extraction. Ubuntu 12.04, Intel 3.20GHz×4 CPU, 8G RAM for text model training and feature extraction. And Window 8, Intel 3.20GHz×4 CPU, 8G RAM for training visual-textual joint model on Matlab.

Dataset: Our experiments were conducted on open benchmark Wikipedia dataset[1], which contains 2886 documents (2173 for training and 693 for test).

[1] http://www.svcl.ucsd.edu/projects/crossmodal/.

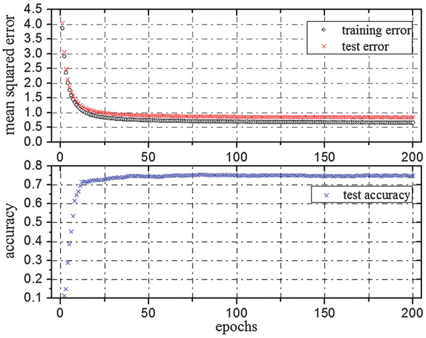


Fig. 3. Top: The mean squared error of training and test against epochs. Bottom: Test accuracy against epochs

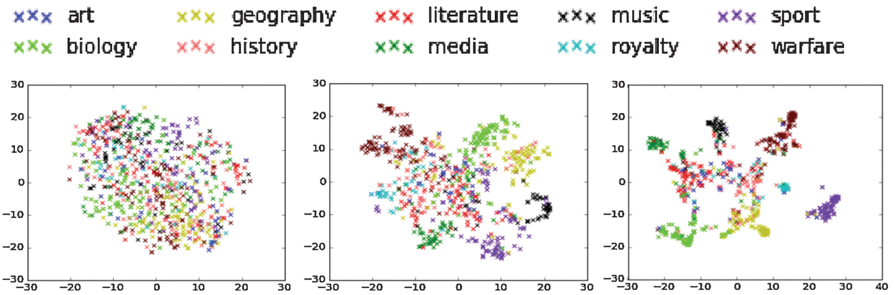
This dataset has 10 semantic categories such as "biology","geography". Each document is comprised of an image and a short descriptive text as the example we given in Fig. 1.

Image Representation: We use deep convolutional neural network (deep CNN) [8] that has been proved its effectiveness in image representation in recent years. Based on Caffe framework [7] we extracted the image feature with Caffe model that on ImageNet [4] ILSVRC2012 dataset (more than 1.2M training images). By extracting the output of the 7^{th} layer(F7), each image can be represented as a 4096-dimension vector, that is, visual feature $I \in \mathbb{R}^{4069}$. Due to image features are highly learned by deep CNN, it can be considered as kind of high level semantic feature.

Text Representation: To represent text as semantic feature, Latent Dirichlet Allocation(LDA) [2] was used to generate 20 topics. We compute the topic distribution of given text document d over 20 topics and finally obtain a 20-dimension vector, that is, textual feature $T \in \mathbb{R}^{20}$.

Training: In DNN learning, the first three layers are designed for intra-modal regularization which optimizes the weights within each modality to improve performance firstly. Thus we named our fusion framework as RE-DNN. The networks are designed as [4096/100/10] and [20/100/10] for image and text receptively, and the last three layers is set as [20/100/10]. In our experiments, learning rate α=0.001, momentum=0.9 achieved the best performance. According the scale of our training data (2173 training samples), we adopted the mini batch gradient descent with batch size 41. The epoch number fixed at K=200. Figure 3 shows the change of mean squared error of training and test during training procedure as well as the increasing of test accuracy against training epochs. We obtained the final test accuracy is 74.6 %.

Table 1. Comparison between unimodal and multimodal fusion feature. Top: visualization of visual feature(a(1)), textual feature(a(2)) and fusion features (a(3)) from test examples. Bottom: classification results include precision (P), recall (R), F1-score (F1) and Accuracy (A).



a(1).Visual Feature

a(2).Textual Feature

a(3).RE-DNN Fusion Feature

classifier	P	R	F1	A
KNN	0.43	0.42	0.42	0.417
SVM	0.44	0.44	0.40	0.440
NB	**0.48**	**0.46**	**0.45**	**0.463**
3L-NN	0.42	0.44	0.42	0.446

b(1).Classification Result

classifier	P	R	F1	A
KNN	0.67	0.68	0.67	0.678
SVM	0.69	0.65	0.66	0.653
NB	0.64	0.59	0.59	0.594
3L-NN	**0.69**	**0.70**	**0.70**	**0.695**

b(2).Classification Result

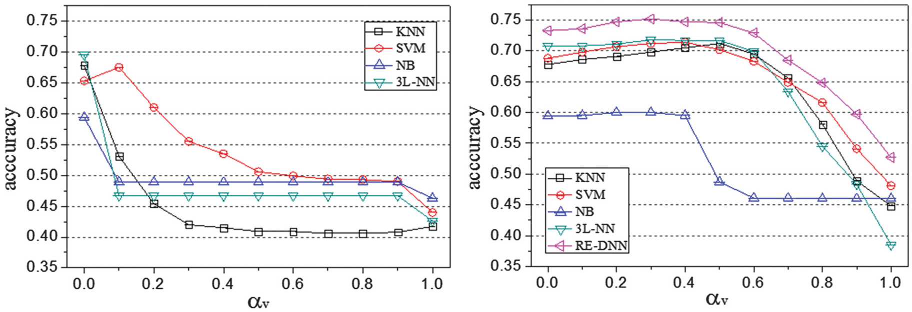
classifier	P	R	F1	A
RE-DNN	**0.74**	**0.74**	**0.73**	**0.746**

b(3).Classification Result

4 Experimental Evaluation

To validate the effectiveness of proposed RE-DNN approach for multimodal feature fusion. Our experiment first consider unimodal (visual or textual feature separately) to perform document categorization task and then compared with RE-DNN approach. In this work, we also explored early fusion and late fusion on some mainstream classifiers such as K-Nearest Neighbor(KNN), Support Vector Machine (SVM), Naive Bayes (NB) and Neural Network(NN).

Table 1 shows the comparison between unimodal feature and multimodal fusion feature based classification. We visualized visual feature $I \in \mathbb{R}^{4096}$ and textual feature $T \in \mathbb{R}^{20}$ to 2D by using t-SNE [15] as shown in a(1) and a(2) respectively. By visually comparing visual and textual feature from a(1) and a(2) find that the margin of textual feature tend to be clearer. Meanwhile, we applied those features to classification. We note that text-based classification outperforms image-based classification for all employed classifiers. This confirms previous research that text information is easier to be perceived and recognized by machines compare to image information. The best performed classification accuracy of visual feature is achieved by NB with 0.463, and a 3L-NN achieved the best classification accuracy 0.695 for textual feature. The configuration of 3L-NN are {4096/100/10} for visual feature and {20/100/10} for textual feature. The learning rate is adjusted as 0.001 and momentum=0.9. However, the further improvements are made by fusing visual and textual feature with deep neural network. This relies on the fact that paired image and text are perceived by machines that they belong to same semantic and the latent relationships



(a) visual-textual early fusion (b) visual-textual late fusion

Fig. 4. Visual-textual early and late fusion

between visual and textual features are captured by network. At this stage, we set α_v=0.5, it means the semantic contribution of each modality are equal so that we can observe the capability of RE-DNN in fusing features. Our final classification accuracy is 74.6 %. Here we extracted the late fusion feature from the output of the 4^{th} layer in RE-DNN and visualized as in a(3). It is clear to see, the fusion features tend to more discriminative than both textual and visual features. Compare Table 1 b(1)–b(3) we see that the overall performance including precision, recall, F1 and accuracy of RE-DNN approach are higher than unimodal based classification. The result shows that late fusion based RE-DNN improves on the approaches "3L-NN for textual" and "NB for visual" by 5.1 % and 28.3 % respectively.

Further experiments were conducted to explore visual-textual early fusion and late fusion by taking the semantic contribution of each modality into consideration. In both fusion strategies, according to Eqs. (1) and (2) we heuristically assign α_v(image modality weight) from 0 to 1. For early fusion, the inputs are raw image and text features. For late fusion, the inputs are prediction scores of different classifiers. Figure 4(a) shows the accuracy changes in early fusion and Fig. 4(b) describes late fusion results. It is observed that late fusion outperforms early fusion at most of levels of α_v. In early fusion approach, almost the accuracy for all classifiers decreasing along with the increasing of α_v. When we impose linear interpolation on RE-DNN, we note that for all levels of α_v, RE-DNN late fusion with linear interpolation further improved the classification accuracy to 75.3 % at α_v=0.3. It proves the effectiveness of our approach.

5 Conclusions

In this paper, we have proposed a DNN framework for fusing visual and textual features. By imposing linear interpolation on DNN, more discriminative and representative fusion feature can be extracted. Our experiments on document categorization show that our proposed approach outperforms mainstream classifiers in both early fusion and late fusion.

References

1. Atrey, P.K., Hossain, M.A., El-Saddik, A., Kankanhalli, M.S.: Multimodal fusion for multimedia analysis: a survey. Multimedia Syst. **16**(6), 345–379 (2010)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
3. Clinchant, S., Ah-Pine, J., Csurka, G.: Semantic combination of textual and visual information in multimedia retrieval. In: Proceedings of the 1st ACM International Conference on Multimedia Retrieval, p. 44. ACM (2011)
4. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, 2009 (CVPR 2009), pp. 248–255. IEEE (2009)
5. Escalante, H.J.: Late fusion of heterogeneous methods for multimedia image retrieval (2008)
6. Hinton, G., Deng, L., Dong, Y., Dahl, G.E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Sig. Process. Mag. **29**(6), 82–97 (2012)
7. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
9. Liu, D., Lai, K.-T., Ye, G., Chen, M.-S., Chang, S.-F.: Sample-specific late fusion for visual category recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 803–810. IEEE (2013)
10. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
11. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of the 28th International Conference on Machine Learning (ICML 2011), pp. 689–696 (2011)
12. Rasiwasia, N., Pereira, J.C., Coviello, E., Doyle, G., Lanckriet, G.R.G., Levy, R., Vasconcelos, N.: A new approach to cross-modal multimedia retrieval. In: Proceedings of the International Conference on Multimedia, pp. 251–260. ACM (2010)
13. Terrades, O.R., Valveny, E., Tabbone, S.: Optimal classifier fusion in a non-bayesian probabilistic framework. IEEE Trans. Pattern Anal. Mach. Intell. **31**(9), 1630–1644 (2009)
14. Thompson, B.: Canonical correlation analysis. In: Everitt, B., Howell, D. (eds.) Encyclopedia of Statistics in Behavioral Science. Wiley, New York (2005)
15. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. J. Mach. Learn. Res. **9**(2579–2605), 85 (2008)
16. Vedaldi, A., Fulkerson, B.: Vlfeat: An open and portable library of computer vision algorithms. In: Proceedings of the International Conference on Multimedia, pp. 1469–1472. ACM (2010)
17. Wu, Z., Jiang, Y.-G., Wang, J., Pu, J., Xue, X.: Exploring inter-feature and inter-class relationships with deep neural networks for video classification. In: Proceedings of the ACM International Conference on Multimedia, pp. 167–176. ACM (2014)

18. Ye, G., Liu, D., Jhuo, I.-H., Chang, S.-F.: Robust late fusion with rank minimization. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3021–3028. IEEE (2012)
19. Yu, J., Cong, Y., Qin, Z., Wan, T.: Cross-modal topic correlations for multimedia retrieval. In: 2012 21st International Conference on Pattern Recognition (ICPR), pp. 246–249. IEEE (2012)