Tensor Fusion Networks for Image Labeling

Adam Dahlgren[1][0000−0002−1112−2981]

Department of Computing Science, Umeå University, Sweden
dali@cs.umu.se

Abstract. Humans use multisensory input in their decision making, and this information is jointly processed in a seamless manner. Multimodal machine learning is the task of learning on structures that are represented to varying degrees in different data sources describing the same real world objects or concepts. This paper presents a Tensor Fusion Network model for labeling images with topics given the visual content and user generated tags. The model uses pretrained word and visual embeddings, as a way to investigate the transferability of semantic embeddings in the context of multimodal machine learning. The model is evaluated with experiments on the MIRFLICKR dataset. The results show that the proposed model gives state-of-the-art performance. It is shown that the fusing of semantic embeddings from different modalities gives performance that is greater than its unimodal parts. The model presented also opens up for interesting research questions regarding the transferability of semantic representations.

Keywords: Multimodal machine learning · semantic embedding · Tensor Fusion Networks · word embeddings · image labeling

1 Introduction

Humans perceive the world through many different senses, combining e.g. visual sensory data with olfactory sensory data to form an understanding of the world. These multiple sensory inputs are processed jointly to form an appropriate action. Utilizing the concept of multisensory input in artificially intelligent systems is central in many applications, but traditionally this is not done in a joint fashion. This problem is addressed with multimodal machine learning [2]. The field opens up for new avenues of research, and with increasing success [23] also improves on traditional machine learning tasks. A common example of multimodal data is video with modalities such as visual content and audio content (possibly extended to include metadata). Traditionally, these different modalities are examined separately (e.g. for face recognition the audio is discarded in the classification stage), where a lot of contextual information shared between the modalities might be lost. In this paper we are concerned with a fusion task, as outlined by Baltrušaitis et al., i.e. "join[ing] information from two or more modalities to perform a prediction" [2].

One concrete motivation to study multimodal machine learning is the ever-increasing demand to process the vast amounts of multimedia created every day

on the Internet [7, 9]. Such tasks include e.g. Visual Question Answering [10] or multimedia information retrieval [16]. Improvements have been shown in many machine learning tasks when using multimodal approaches, such as information retrieval [11], and speech-to-text [20]. Intuitively, both modalities are different and, each on its own, incomplete representations of the same underlying information. Therefore, if e.g. a loud noise is heard in a recorded video of a speech, visual information can offset such noisy data to still produce a good transcription. Learning joint representations also has other interesting applications such as generating shapes from natural language [5].

Another motivation for this study is that semantic representations in general are becoming increasingly useful [4]. Recent efforts towards human-understandable, *explainable*, AI focus on algorithms that produce results that can be explained using its internal representation, or meaning. In addition to being more easily interpretable by humans, semantic representations can be used to improve the performance as well [25].

This paper investigates how semantic embeddings can be used as an intermediate representation for multimodal machine learning. The main contributions of this paper are

- Constructing semantic embeddings of image and text that give good performance on image labeling.
- Joining these two by means of a Tensor Fusion Network to replicate previous results on TFNs on a new dataset.
- Evaluating the performance of the resulting system against its unimodal parts as well as against the current state-of-the-art approaches to image labeling.

Section 2 of the paper gives a rough overview of related work on multimodal machine learning and semantic embeddings. Section 3 outlines the proposed Tensor Fusion Network model that will be used in the experiments. Section 4 describes the experiment setup and the dataset. Finally, Section 5 presents the results and Section 6 reflects on them while outlining future work.

2 Related work

This section is divided into two parts. The first part describes Deep Boltzmann Machines and Tensor Fusion Networks, the two models that will be compared. The second part describes previous work on semantic embeddings, which will be the basis for the Tensor Fusion Network model used in this paper.

Multimodal machine learning Early approaches to multimodal machine learning adhere to one of two categories: early fusion or late fusion [26]. One big drawback of both methods is that neither approach directly models both intra-modality and inter-modality dynamics. Recent advancements employ more elaborate model architectures, such as Deep Boltzmann Machines [28], using deep

learning techniques to capture both these dynamics. Another similar method is the Tensor Fusion Networks, combining semantic embedding techniques with deep learning to show state-of-the-art performance in video sentiment analysis [30]. In addition to the usage of semantic embeddings in Tensor Fusion Networks, embeddings such as word2vec [19] and GloVe [22] have shown significant improvements in many NLP applications [12, 14].

Deep Boltzmann Machines (DBMs) [24] can be adapted to a multimodal setting by joining a DBM per modality with a fully connected layer. Given image and text, this means that one would have two subnetworks (DBMs) that connect via a final single output layer, and that these networks are trained jointly. Srivastava and Salakhutdinov [28] apply multimodal Deep Boltzmann Machines to the public dataset MIRFLICKR [8] to perform image labelling. MIRFLICKR contains 1 million images with user given tags taken from the website flickr.com. Out of those 1 million images, 25000 are annotated with one or more of 38 general labels. These labels both describe things seen in the image such as `people`, and scene information such as `vacation`. The authors pretrain parts of their model on all 1 million images, and proceed to use the 2000 most common tags together with the images to predict labels.

Another recent advancement in multimodal machine learning is the Tensor Fusion Networks (TFN) [30]. Zadeh et al. show state-of-the-art-performance on video sentiment analysis, with data from three modalities (audio, video, text). Tensor fusion networks have a three-part structure: (1) a semantic embedding layer with one embedding subnetwork per modality, (2) a tensor fusion layer, and (3) an inference subnetwork. In order to adapt a TFN to a specific problem, the semantic embeddings should capture important aspects of each modality (e.g. temporality). The semantic embeddings can be pre-trained networks (such as word2vec [19]). Semantic embeddings model the meaning of an input, rather than the raw values as done with traditional feature vectors. In the case of word2vec, this means that the semantic embedding for a word consists of the probability distribution (represented as a vector) of words that appear in the same context (i.e. closeby in a sentence). Semantic embeddings are sometimes used synonymously with latent features.

Semantic embeddings One recent advancement in semantic embeddings is the work on word embeddings. Instead of modelling text by its direct raw features (e.g. one-hot vectors), word embedding methods project this information into a latent feature space of a much lower dimension than the original feature space using unsupervised learning techniques. Word embedding vectors have shown interesting semantic properties, such as preserving semantic meaning under vector addition/subtraction. An example of this is word2vec, where adding e.g. the embeddings for *German* and *airline* results in a vector close to the embedding of the German airline *Lufthansa* [19]. Another modern word embedding technique is GloVe [22]. The main difference between them is that word2vec is a predictive model while GloVe is count-based, as outlined by Baroni et al. [3]. According to their results, the predictive approach performs better across a range of tasks.

Embedding visual content using convolutional neural networks (CNNs) pretrained on the ImageNet dataset [13] has been shown useful in several studies. Wei et al. show state-of-the-art performance across five different cross-modal (image and text) retrieval tasks [29], where the visual embedding is based on an ImageNet CNN model. DeViSE [6] is a deep visual-semantic embedding model that combines a skip-gram language model with an ImageNet based deep convolutional neural network, showing state-of-the-art results better than unimodal contemporaries. VGGnet [15] is a well established CNN model that was first applied to the ImageNet challenge with a significant bump to the state-of-the-art performance. Other models with even better performance have been proposed since then, but pretrained VGG16 models are readily available in common frameworks such as Tensorflow [1].

Utilizing pretrained models is sometimes referred to as transfer learning [21]. In the case of semantic embeddings, the idea is that the semantics of e.g. a word should be the same over different applications.

3 Proposed model

This section describes the proposed Tensor Fusion Network-based model used in the experiments of this paper. An overview is given in Figure 1. The figure illustrates how the semantic embeddings are produced by a subnetwork per modality, and how the product of these embeddings is fed into a classifier. Given that the

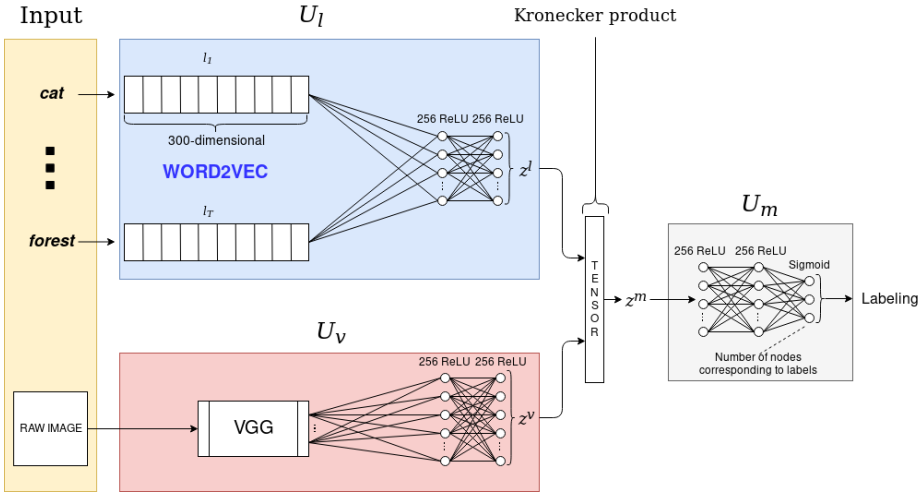

Fig. 1: Overview of proposed Tensor Fusion Network model.

problem studied in this paper is a multiclass multilabel classification problem, the model is trained using the binary cross entropy loss function with sigmoid

activation functions. Focal loss [17] was considered, but performed poorly during initial testing.

3.1 Text embedding subnetwork

The text embedding subnetwork $\mathcal{U}_l$ is a word2vec-based model trained on the tag data. The network takes as input a vector containing the vocabulary indices of the words in the input tags. These words are embedded as 300-dimensional word vectors $\boldsymbol{l} = l_1, l_2, \cdots, l_{T_l} \in \mathbb{R}^{300}$, where T_l is the number of words in the input. Dimensionalities ranging from 50 to 500 were considered during initial experiments, where 300 showed the most consistently good performance. This also aligns with the dimensions used by the Google News word embeddings. The flattened embedding is fed into two layers of fully connected nodes with 256 ReLU ($relu(x) = \max(0, x)$) nodes per layer. This layer produces the text embedding $\boldsymbol{z}^l$:

$$\boldsymbol{z}^l = \mathcal{U}_l(\boldsymbol{l}; W_l) \in \mathbb{R}^{256}$$

W_l denotes the weights of the word embedding. The network is later evaluated both trained from scratch and initialized with weights W_l from a word2vec model pre-trained on the Google News dataset [18] of about 100 billion words. The weights from the pre-trained model are then fixed during training.

For the unimodal experiments a final classification layer with the same number of nodes as the number of classes, with a sigmoid activation function and binary cross entropy loss.

3.2 Visual embedding subnetwork

The visual embedding subnetwork $\mathcal{U}_v$ is based on a VGGnet model pre-trained on the ImageNet dataset [15] provided in the machine learning framwork Keras [1]. The visual embedding subnetwork is initialized with weights W_{vgg} from the VGG16 model, and fine-tuned during training of the entire tensor fusion network. The VGG layer produces an initial visual embedding $\boldsymbol{z}^{v'} \in \mathbb{R}^{25000}$ that is fed into two layers of fully connected nodes with 256 ReLU nodes per layer. This produces a final semantic embedding $\boldsymbol{z}^v$:

$$\boldsymbol{z}^v = \mathcal{U}_v(\boldsymbol{z}^{\boldsymbol{v'}}; W_{vgg}) \in \mathbb{R}^{256}$$

Analogously with the text embedding subnetwork, a classification layer using a sigmoid activation function with a binary cross entropy loss is employed for the unimodal experiments.

3.3 Tensor Fusion Network

The central part of the Tensor Fusion Network is the tensor fusion layer. Figure 2 illustrates the Kronecker product of the semantic embeddings. The embeddings

[1] https://keras.io/applications/#vgg16

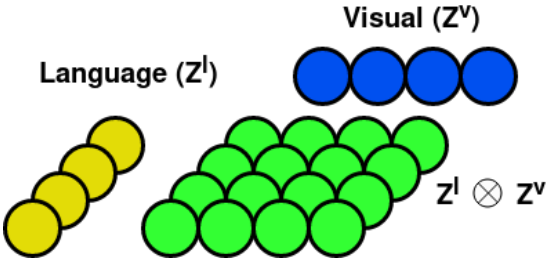

Fig. 2: Visualisation of tensor fusion product. Z^l represents the semantic embedding of the text, and Z^v the semantic embedding of the image. Each point $x_{ij} \in Z^l \otimes Z^v$ corresponds to the product $l_i v_j$ of semantic text feature l_i and image feature v_j.

$\boldsymbol{z}^l$ and $\boldsymbol{z}^v$ from the embedding subnetworks are forwarded to a tensor fusion layer that performs the Kronecker product:

$$\boldsymbol{z}^m = \begin{bmatrix} \boldsymbol{z}^l \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \boldsymbol{z}^v \\ 1 \end{bmatrix} \in \mathbb{R}^{257x257}$$

The semantic embeddings are expanded with an extra constant with value 1 in order to keep $\boldsymbol{z}^v$ and $\boldsymbol{z}^l$ in the resulting tensor. Otherwise, the resulting tensor would not contain the unimodal features but only the bimodal interactions. This ensures that the network will capture features that are perhaps only seen in one modality that would otherwise be lost. The resulting 2D tensor $\boldsymbol{z}^m$ is flattened and fed into a three layer fully connected network $\mathcal{U}_m$. The two first layers of $\mathcal{U}_m$ consists of 256 ReLU nodes, connected to the classification layer of 38 nodes. The final layer uses the sigmoid activation function and binary cross entropy loss.

Dropout layers with probability $p = 0.2$ are added in order to counteract overfitting [27]. These dropouts are placed inbetween each pair of ReLU layers in the entire network. Different p-values were considered, and $p = 0.2$ showed sufficient improvement while not risking underfitting as was the tendency with other configurations (e.g. p-values increasing from 0.2 to 0.4). That being said, these settings can be further investigated.

In the following section the subnetworks will be referred to as TFN-text and TFN-visual, and the entire model as TFN.

4 Experiments

The Tensor Fusion Network model described in Section 3 is trained end-to-end on the MIRFLICKR dataset. Specific details on e.g. loss and activation functions are found in Section 3. The proposed models is implemented in Keras, and trained on a Tesla P100 GPU. TFN-text, TFN-visual, and TFN had $3,591,064$,

21, 213, 030, and 59, 345, 170 trainable parameters respectively. The precision and recall will be used to compare the performance of the proposed model against its unimodal subnetworks. The Mean Average Precision (MAP) will be used to compare the performance of the proposed model against Multimodal Deep Boltzmann Machines and the unimodal subnetworks.

The TFN-text network is evaluated both when training from scratch, and when using a pre-trained Google News word2vec model. The latter is to evaluate the transferability of semantic embeddings in this context.

Dataset and training The MIRFLICKR dataset contains 1 million images from the social photography site flickr.com/. Each image has a set of user-generated tags, where 1386 of the tags occur in at least 20 images. There are on average 8.94 tags per image. Out of these 1 million images, 25 000 are labeled with a subset of 38 topics. These topics include object categories, such as *people, flower*, and scene categories, such as *clouds, sunset*. The dataset is divided into a training set of 20 000 samples and a validation set of 5 000 samples. A batch size of 128 is used during training, for a period of 30 epochs. Figure 3 shows the frequencies of which the labels appear in the 25 000 samples.

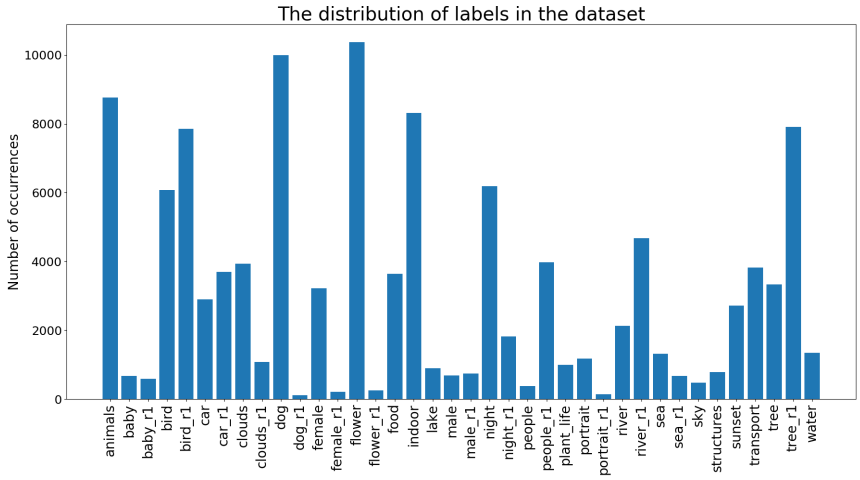

Fig. 3: Distribution of labels in the MIRFLICKR-25000 dataset. Each image can have multiple labels.

4.1 Preprocessing

All the images are resized to 224×224 pixels, as this is the size the VGGnet model is pretrained on. The pixel values are normalized to the interval $[0, 1]$ instead

of the original $[0, 255]$. It is worth noting that tags are not filtered to use only the 2000 most commonly seen tags, in contrast to the authors of the benchmark DBMs [28]. Providing all of the tags gives the word embedding subnetwork a broader basis to give a good mapping in the embedding space. Initial experiments showed better performance when including all tags. 419 images were missing annotated labels and were removed from the dataset.

5 Results

Table 1 shows the comparision in Mean Average Precision (MAP) between TFN, its subnetworks, and the multimodal Deep Boltzmann Machines on the MIR-FLICKR dataset. The *TFN* model outperforms both *TFN-text* and *TFN-visual* by a significant margin, showing that the fusion of the semantic embeddings improve performance. The Deep Boltzmann Machines are outperformed by both the visual semantic embedding subnetwork and the Tensor Fusion model. The usage of the pre-trained Google News word embedding gives significantly worse performance when compared to training the embeddings from scratch. Further experiments using the pre-trained embedding were consistent with this result, and are therefore omitted.

Table 1: Showing the Mean Average Precision for all three models in comparision with Multimodal Deep Boltzmann Machines (as reported by Srivastava & Salakhutdinov [28]). It is important to note that by using the reported MAP, the DBM and TFN models are not evaluated on the same test set.

Model	MAP
Random	0.125
TFN-text-google	0.478
TFN-text-300	0.519
TFN-visual	0.648
TFN	**0.678**
DBM (as reported in [28])	*0.609*

Figure 4 shows two classifications made by the TFN model on the test set. Figure 4a is the prediction with the smallest error ($|e| = 3.6 \cdot 10^{-5}$), while Figure 4b is the prediction with the largest error ($|e| = 9.1$) in the test set. For Figure 4a the prediction was the exact labels listed, whilst only labels `animal`, `dog_r1` and `baby` were correctly identified in Figure 4b with all other labels having a probability close to zero.

Figure 5 shows the precision and recall at the top 5 predictions for all three models (TFN-text, TFN-visual, TFN). All three models have perfect or near-perfect precision for many labels, e.g. `animals` or `people`, but there is a great variance in the recall. For some of the labels all three models perform poorly,

(a) Tags: `silvia portrait retrato zuan 2007`. Labels: female_r1, female, indoor, people_r1, people, portrait_r1, portrait.

(b) Tags: `Rierra Ramona wisdoc dog labrador yellow lab friends SearchTheBest canon`. Labels: animals, baby, dog_r1, dog, flower, indoor, male_r1, male, people_r1, people, plant_life, portrait_r1, portrait.

Fig. 4: The best and the worst classifications by the TFN model, together with the original tags and ground truth labels.

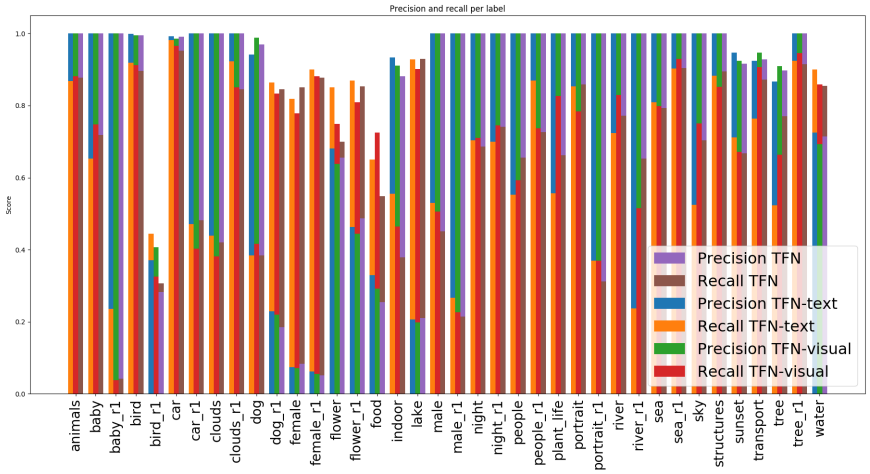

Fig. 5: Precision and recall per label for the Tensor Fusion Network Model (TFN), and each of its subnetworks (TFN-text and TFN-visual). Computed over the top 5 predictions.

which could be attributed to the number of samples in the dataset for each of those labels. The label `river_r1` exemplifies the precision being same for all models, but where recall is improved by the fusion of modalities. Other labels, e.g. `food`, show that the precision and recall can be worsened by the fusion.

The three proposed models were trained for 10 epochs for TFN-text, and 30 epochs for both TFN-visual and TFN. The training procedure showed a resonable execution time for the TFN model compared to the other two. On average of TFN-text took 3s per epoch, TFN-visual took 150s per epoch, and TFN took 153s per epoch.

6 Discussion

The experiments presented in this paper show a couple of interesting results. It is shown that Tensor Fusion Networks can be a more powerful tool than Deep Boltzmann Machines in multimodal machine learning. More importantly, the proposed model strengthens the claim that combining information from different modalities gives performance greater than the each of its parts. Another interesting observation is that the pretrained Google News word embedding did not show better performance for the text-only classification even though it is based on a much larger corpus. Moreover, even though the TFN model intuitively seems much more complex than its subnetworks, the results in terms of the time to train indicate a very small overhead when combining them. This could be attributed to the TFN-visual network (and by extension VGGnet) doing the heavy lifting. However, it is necessary to make more comprehensive experiments and comparisons to establish a fair judgement of Tensor Fusion Networks versus Deep Boltzmann Machines for image labeling. The current setup where the DBM and TFN models are not evaluated on the same test set is the biggest flaw of these experiments, and might give skewed results unfair to the DBM model. Another flaw of the experiments is that currently there is no correction for the difference in suppport for the different labels. Taking the support into account during the training phase could improve the results even further, since the support varies greatly (as presented in Section 4). It is also necessary to compare these results with those of early and late fusion approaches.

The best and worst classifications of the TFN model show that even the worst predictions still manage to pick out the most prominent of the labels. This is not necessarily true for all predictions done on the test data and further analysis of this phenomenon is needed to draw strong conclusions, but it is interesting either way. It is also worth noting that the labels missing in the worst classification are not clear from the image nor the labels. For example, the flower in the top-left corner is out of focus and very much in the background, the gender of the baby is hard to tell, and it is not clear if this should be classified as a portrait or people given the content of the image.

The results of this paper open a number of interesting lines of future research. Firstly, the subnetworks of the proposed model are rather simplistic in their construction. The model could utilize many of the state-of-the-art methods for deep learning (e.g. normalization), and the classifier subnetwork could employ a more advanced architecture. Another way to improve the performance of the model could be to pretrain the word embeddings on the entire 1 million image dataset instead of using the Google News embeddings. Secondly, initializing the Tensor

Fusion Network model with the weights obtained when using the subnetworks unimodally for classification could prove fruitful. This would be another way of investigating semantic embeddings in the context of transfer learning.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015), https://www.tensorflow.org/, software available from tensorflow.org
2. Baltrusaitis, T., Ahuja, C., Morency, L.: Multimodal machine learning: A survey and taxonomy. CoRR **abs/1705.09406** (2017), http://arxiv.org/abs/1705.09406
3. Baroni, M., Dinu, G., Kruszewski, G.: Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 238–247 (2014)
4. Chang, X., Ma, Z., Yang, Y., Zeng, Z., Hauptmann, A.G.: Bi-level semantic representation analysis for multimedia event detection. IEEE Transactions on Cybernetics **47**(5), 1180–1197 (May 2017). https://doi.org/10.1109/TCYB.2016.2539546
5. Chen, K., Choy, C.B., Savva, M., Chang, A.X., Funkhouser, T.A., Savarese, S.: Text2shape: Generating shapes from natural language by learning joint embeddings. CoRR **abs/1803.08495** (2018), http://arxiv.org/abs/1803.08495
6. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M.A., Mikolov, T.: Devise: A deep visual-semantic embedding model. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 26, pp. 2121–2129. Curran Associates, Inc. (2013), http://papers.nips.cc/paper/5204-devise-a-deep-visual-semantic-embedding-model.pdf
7. Gandomi, A., Haider, M.: Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management **35**(2), 137–144 (2015)
8. Huiskes, M.J., Lew, M.S.: The mir flickr retrieval evaluation. In: MIR '08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval. ACM, New York, NY, USA (2008)
9. IBM: What is big data? http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html
10. Ilievski, I., Feng, J.: Multimodal learning and reasoning for visual question answering. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 551–562. Curran Associates, Inc. (2017), http://papers.nips.cc/paper/6658-multimodal-learning-and-reasoning-for-visual-question-answering.pdf
11. J. Huiskes, M., Thomee, B., S. Lew, M.: New trends and ideas in visual concept detection: The mir flickr retrieval evaluation initiative. MIR 2010 - Proceedings of the 2010 ACM SIGMM International Conference on Multimedia Information Retrieval pp. 527–536 (01 2010). https://doi.org/10.1145/1743384.1743475

12. Kim, Y.: Convolutional neural networks for sentence classification. CoRR **abs/1408.5882** (2014), http://arxiv.org/abs/1408.5882
13. Krizhevsky, A., Sutskever, I., E. Hinton, G.: Imagenet classification with deep convolutional neural networks. Neural Information Processing Systems **25** (01 2012). https://doi.org/10.1145/3065386
14. Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., Zhong, V., Paulus, R., Socher, R.: Ask me anything: Dynamic memory networks for natural language processing. In: International Conference on Machine Learning. pp. 1378–1387 (2016)
15. Lenc, K., Vedaldi, A.: Understanding image representations by measuring their equivariance and equivalence. International Journal of Computer Vision (May 2018). https://doi.org/10.1007/s11263-018-1098-y, https://doi.org/11.1007/s11263-018-1098-y
16. Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: State of the art and challenges. ACM Trans. Multimedia Comput. Commun. Appl. **2**(1), 1–19 (Feb 2006). https://doi.org/10.1145/1126004.1126005, http://doi.acm.org/10.1145/1126004.1126005
17. Lin, T., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. CoRR **abs/1708.02002** (2017), http://arxiv.org/abs/1708.02002
18. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Googlenews-vectors-negative300.bin.gz - efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013), https://code.google.com/archive/p/word2vec/
19. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 26, pp. 3111–3119. Curran Associates, Inc. (2013), http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf
20. 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-11). pp. 689–696 (2011)
21. Pan, S.J., Yang, Q., et al.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering **22**(10), 1345–1359 (2010)
22. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543 (2014), http://www.aclweb.org/anthology/D14-1162
23. Poria, S., Cambria, E., Bajpai, R., Hussain, A.: A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion **37**, 98 – 125 (2017). https://doi.org/https://doi.org/10.1016/j.inffus.2017.02.003, http://www.sciencedirect.com/science/article/pii/S1566253517300738
24. Salakhutdinov, R., Hinton, G.: Deep boltzmann machines. In: van Dyk, D., Welling, M. (eds.) Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 5, pp. 448–455. PMLR, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA (16–18 Apr 2009), http://proceedings.mlr.press/v5/salakhutdinov09a.html
25. Shen, Y., He, X., Gao, J., Deng, L., Mesnil, G.: Learning semantic representations using convolutional neural networks for web search. In: Proceedings of the 23rd

International Conference on World Wide Web. pp. 373–374. WWW '14 Companion, ACM, New York, NY, USA (2014). https://doi.org/10.1145/2567948.2577348, http://doi.acm.org/10.1145/2567948.2577348
26. Snoek, C.G.M., Worring, M., Smeulders, A.W.M.: Early versus late fusion in semantic video analysis. In: Proceedings of the 13th Annual ACM International Conference on Multimedia. pp. 399–402. MULTIMEDIA '05, ACM, New York, NY, USA (2005). https://doi.org/10.1145/1101149.1101236, http://doi.acm.org/10.1145/1101149.1101236
27. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., Bengio, Y.: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research (1929)
28. Srivastava, N., Salakhutdinov, R.R.: Multimodal learning with deep boltzmann machines. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 25, pp. 2222–2230. Curran Associates, Inc. (2012), http://papers.nips.cc/paper/4683-multimodal-learning-with-deep-boltzmann-machines.pdf
29. Wei, Y., Zhao, Y., Lu, C., Wei, S., Liu, L., Zhu, Z., Yan, S.: Cross-modal retrieval with cnn visual features: A new baseline. IEEE Transactions on Cybernetics **47**(2), 449–460 (Feb 2017). https://doi.org/10.1109/TCYB.2016.2519449
30. Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.: Tensor fusion network for multimodal sentiment analysis. CoRR **abs/1707.07250** (2017), http://arxiv.org/abs/1707.07250

A Review comments

In this appendix the reviews presented in their full form with comments on how they were implemented or discarded.

A.1 Review 1

```
+ Strengths:
++ The subject is timely.
++ The paper is well-written and easy to follow.

- Weakness:
--In Section 3, there is no information how the author
    picked the numeric values for parameters of the
    network (for example, the word vector dimension is
    300, the dropout layer with probability 0.2?)

--In Section 5, the author mention number of samples in
    the original dataset is the reason of poor
    performances of three models for some labels. Is it
    because too many or too few number of samples?, could
    the author discuss more detail on this by giving some
    statistical information?
```

——Figure 3 is unclear (the x–axis and y–axis labels are too small, and the colors are overlapped), hence difficult to understand.

——The author should also consider the complexity and running time of TFN and DBM in the comparison between these two algorithms.

——Citation is not consistent (i.e., some citations have links, some without links).

All the comments have been taken into consideration. Explainations on parameters and elaborations on the running times et c. have been added. However, the author did not have time to fix inconsistencies in citation links, and Fig. 3 is still a bit unclear. The figure should have been divided into multiple subfigures, or completely replaced with a better representation of the information.

A.2 Review 2

This paper proposes a multimodal tensor fusion network for the task of image labelling. The author compared the proposed approach with unimodal networks (i.e., text based and visual based) and an existing multimodel DBMs. MAP (mean averge precision), precision, and recall were used as evaluation measurements. The results show that the multimodal network outperforms others in terms of MAP. The data set used in this paper is MIRFLICKR with 1 million labelled Flickr images with labels. In sum, this paper is resonably written, howeer, technique contributions could be improved.

Section 1: contribution 1 and contribution 2 could be merged.
The current draft did not present significant technique contribution on fusion, even if the challenges do exist (e.g., hyperparameter optimization).

Section 3: the authors explained the text embedding subnetwork (a word2vector based model), and the visual embedding subnetwork (VGGnet based model), and the fused model TFN (Tensor Fusion Network).

Section 4.1: the author stated "It is worth noting that tags are not filtered to use only the 200 most commonly seen tags, in contrast to the authors of the

benchmark DBMs [28]. The author should clarify the advantage and limitations of using the image selection methods in this paper or in DBMs.

Table 1: the author displayed MAP result of DBMs as reference. However, in this paper, it seems DBMs and the proposed TFN was evaluated on different test data, if so, more explanations are required.

Section 4 and 5 could be merged.
Section 6: the discussion part is valuable with a few pointed future directions.

Table 1 and Figure 3 show the evaluation results.
An intuitive thought regarding model evaluation would be how about the tradeoffs of execution efficiency and precision/reall.

When analyze the results, it would be interesting and intuitive to display some exmples (images with resulting labels), to show different models may perform well on different types of categories (people, animal, river etc.)

Typo: "have ben" -> "have been"

All but one comment have been taken into consideration and implemented to the best of the authors ability given the time at hand. It turned out that in order to provide some of the further information requested (and rightly so), the experiments had to be rerun to save that additional information. This took more time than expected and resulted in the author not being able to address the issues properly (i.e., clearer figures would have been nice and a better analysis, and example classifications for all modalities).

The suggestion that Section 4 and 5 could be merged was not implemented for two reasons. The first reason is that it is common within the field to have a separate section for the experiment setup. Secondly, the sections in the paper are divided according to semantically grouped content, and while the experiment section is not that long it is definitely a clear such group. Additions following the reviewers comments made the section longer as well, further supporting the decision.