Exploring Deep Multimodal Fusion of Text and Photo for Hate Speech Classification

Fan Yang **Xiaochang Peng** **Gargi Ghosh**
Reshef Shilon **Hao Ma** **Eider Moore** **Goran Predovic**
Facebook Inc.
{flymonkey,xiaochang,gghosh,reshefshilon,haom,idr,predovic}@fb.com

Abstract

Interactions among users on social network platforms are usually positive, constructive and insightful. However, sometimes people also get exposed to objectionable content such as hate speech, bullying, and verbal abuse etc. Most social platforms have explicit policy against hate speech because it creates an environment of intimidation and exclusion, and in some cases may promote real-world violence. As users' interactions on today's social networks involve multiple modalities, such as texts, images and videos, in this paper we explore the challenge of automatically identifying hate speech with deep multimodal technologies, extending previous research which mostly focuses on the text signal alone. We present a number of fusion approaches to integrate text and photo signals. We show that augmenting text with image embedding information immediately leads to a boost in performance, while applying additional attention fusion methods brings further improvement.

1 Introduction

While social network platforms give people the voice to speak, they also have a need to moderate abusive and objectionable content that is harmful for their communities. Most social platforms have explicit policy against hate speech (e.g. https://www.facebook.com/communitystandards/hate_speech) because such content creates an environment of intimidation, exclusion, and in some cases promote real-world violence.

The automatic identification of hate speech has been mostly formulated as a natural language processing problem (e.g. Mishra et al., 2018; Gunasekara and Nejadgholi, 2018; Kshirsagar et al., 2018; Magu and Luo, 2018; Sahlgren et al., 2018). The signal from text, however, sometimes is not sufficient for determining whether a piece of content (such as a post) on the social network platforms constitutes hate speech. There is a need to take into account signals from multiple modalities in order to have a full comprehension of the content for hate speech classification. For example, "these are disgusting parasites", the sentence itself can be either benign or hateful, depending on what "these" refer to; and when it is combined with a photo of people or symbols in a post, it is very likely to be hate speech. We have seen many cases where the text itself is benign, but the whole post is hateful if we consider the context of the image.

There has been a number of research on multimodal fusion in the deep learning era. For example, Tong et al. (2017) apply an outer product fusion method to combine text and photo information for the task of detecting human trafficking. For the task of user profiling, formulated as a multi-tasking classification problem, Vijayaraghavan et al. (2017) propose a hierarchical attention model; and Farnadi et al. (2018) propose the UDMF framework, a hybrid integration model that combines both early feature fusion and later decision fusion using both stacking and power-set combination. Zhong et al. (2016) also studied the combination of image and captions for the task of detecting cyberbullying. For the task of name tagging, formulated as a sequence labeling problem, Lu et al. (2018) apply a visual attention model to put the focus on the sub-areas of a photo that are more relevant to the text encoded by a bi-LSTM model. For the task of image-text matching, Wang et al. (2017) compare an embedding network that projects texts and photos into a joint space where semantically-similar texts and photos are close to each other, with a similarity network that fuses text embeddings and photo embeddings via element multiplication. For the task of sentiment analysis, Zadeh et al. (2017); Ghosal et al. (2018);

Proceedings of the Third Workshop on Abusive Language Online, pages 11–18
Florence, Italy, August 1, 2019.

Bagher Zadeh et al. (2018); Liu et al. (2018) propose several models, namely contextual inter-modal attention, dynamic fusion graph, and low-rank multimodal fusion, for integrating visual, audio, and text signals on the CMU-MOSEI data set. There is also research initiative in multimodal summarization (Li et al., 2017) and multimodal translation (Calixto et al., 2017; Delbrouck and Dupont, 2017). These works have demonstrated the effectiveness of multimodal fusion methods in problems where non-text signals play an important role in disambiguating the text.

In this research, we explore deep multimodal fusion of text and photo for the task of hate speech classification on social networks, where hate speech posts frequently appear with images. We experiment with many fusion techniques, including simple concatenation, bilinear transformation, gated summation, and attention mechanism. We find that concatenation with photo information in the convolution text classifier immediately gives us a nice gain, while fusion with attention offers further improvement. Specifically attention with deep cloning, sparsemax, and symmetric_gate provides the best performance. These results shall shed light on better identifying hate speech to provide a safer community of online social networks.

2 Text And Photo Fusion

In this section we first describe our baseline convolutional text classifier, and the image features of photos. We then describe many approaches of fusing texts and photos, including basic concatenation, gated summation, bilinear transformation, and attention with different alternations.

2.1 Convolutional text model

We adopt the convolutional sentence classification architecture by Kim (2014) as our baseline text model, as illustrated on the left hand side in Figure 1.

1. For each word in a piece of text, we retrieve the pre-trained embeddings $[v_1, v_2, ..., v_n]$. These embeddings are fixed during our model training. We then apply a word-level MLP on each of the word embeddings, creating the new word embeddings $[v'_1, v'_2, ..., v'_n]$. This word-level MLP serves as a solution of fine-tuning the word embeddings towards the hate speech domain, by applying a systematic transform to the whole embeddings space, which has the benefit of also taking care of words that do not appear in the training data. We then apply a dropout layer on the word-level so that the model is more robust against word embeddings features.

2. We next apply a 1D-convolution to the words. With proper padding, we ensure that the output of the convolution matches the length of the input for different ngram-window sizes (Gehring et al., 2017). This offers the convenience for executing attention operation (see Section 2.5). The output of the convolution is a list of vectors $[c_1, c_2, ..., c_n]$.

3. We then apply max-pooling and tanh to create a fixed-size vector representation for the piece of text, denoted as t.

4. Finally we apply dropout, MLP and softmax on the vector t to discriminate between *hate* vs *benign*.

2.2 Photo features

We first pre-train a deep neural network for image classification, similar to the deep ResNet neural architecture (He et al., 2016) for ImageNet (Deng et al., 2009), with hundreds of millions of photos on a social network platform (not limited to the domain of hate speech). For each photo, we then extract the features from the second last layer, which is a float vector of 4096 dimensions. Finally we run iterative quantization to convert this vector into a hash of 256-bit binary vector (Gong et al., 2013). We store the photo hashes for efficient photo indexing, searching, and clustering.

In this research, we conveniently represent each photo with its hash (Sablayrolles et al., 2018). The hash takes advantage of the deep pre-trained image network which offers discriminative semantic representations. It preserves the similarity between original photos: the photos with smaller Hamming distance between their hashes look similar to each other. While it is sub-optimal as the iterative quantization might be information-lossy, the photo hashing technique provides an infrastructure-economic solution to compactly store and promptly retrieve the information of billions of photos on the platform.

Note that the hash comes from the second last layer representations of the deep ImageNet-like network. This has the flavor of transfer learning

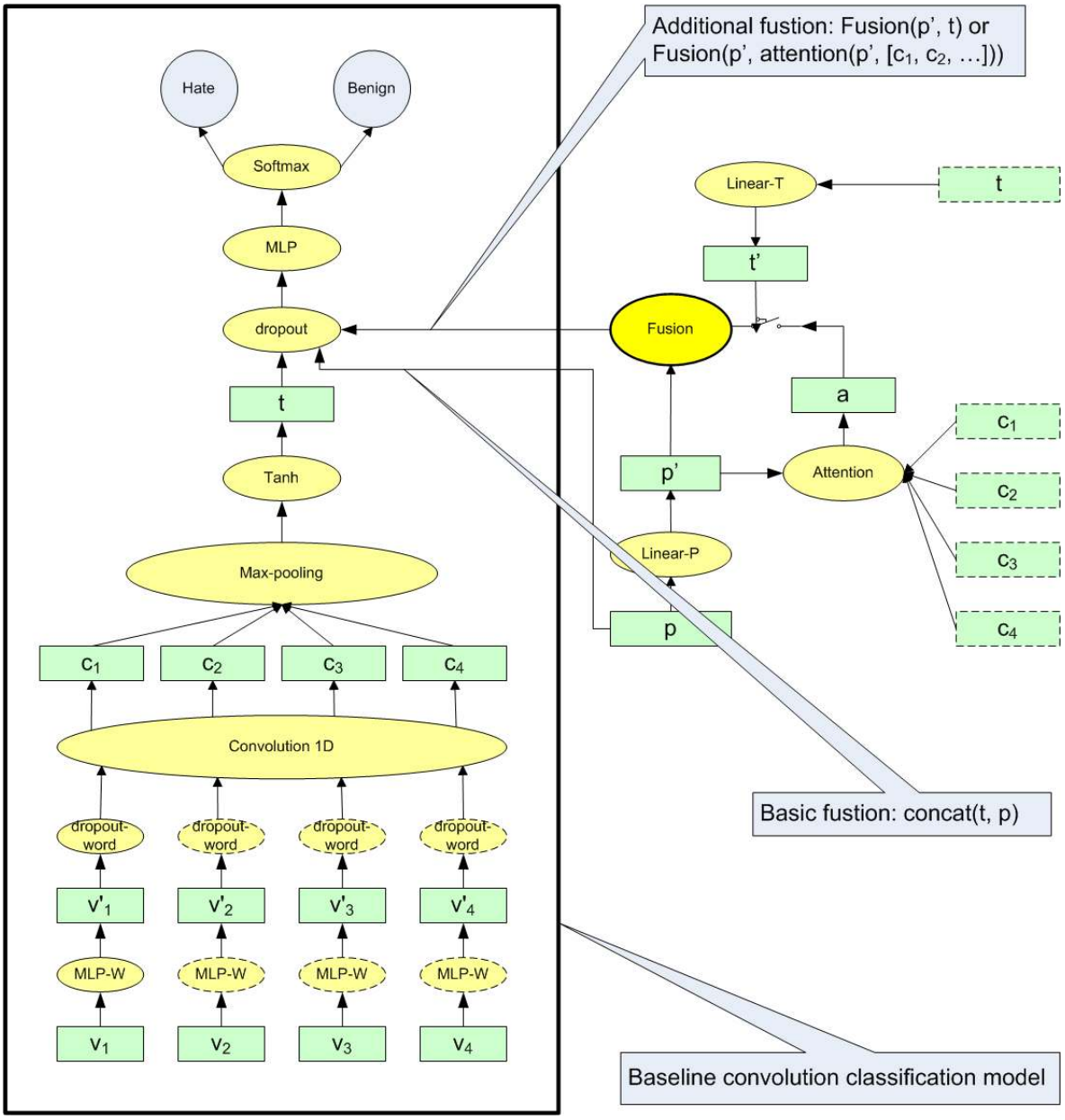


Figure 1: Model architecture of text and photo fusion: ellipses (in yellow) represent operations; and rectangles (in green) represent vectors. Shapes in dot lines are *clones* of their corresponding components.

(Oquab et al., 2014), where we pre-train the network with a large amount of out-of-domain photos, and then fix the second last layer and below. The hash offers a generic representation for which we will then fine-tune with in-domain photos.

2.3 Basic fusion: concatenation

The most straightforward way of integrating text with photo features is to concatenate t and p, as illustrated in Figure 1, where t is the text representation vector after max-pooling and *tanh* activation function, and p is the 256-dimensional photo hash as mentioned before. The concatenated vector is followed by dropout, MLP and softmax operations for the final hate speech classification. Note that with this basic concatenation, the photo hash p would actually impact the text representation t through back-propagating the loss down to the word embeddings MLP.

2.4 Additional fusion

On top of the basic concatenation, we have also explored other fusion techniques: gated summation and bilinear transformation.

- **Gated summation** Miyamoto and Cho (2016) propose a gated summation approach

to integrate word and character information. We adopt their approach and apply it to text and photo fusion, as illustrated in Equation (1). We first apply linear transformations to t and p so that they have the same dimension $|t'| = |p'|$. We then calculate a gate G as a sigmoid (σ) function on p', where u_p (a weighed vector) and B_p (a bias scalar) are parameters to be learned. We then use the gate value G to weigh the summation of t' and p' to create the fusion vector f. We use the vector $concat(t, p, f)$ for the target hate speech classification.

$$\begin{aligned} t' &= W_t \cdot t + b_t \\ p' &= W_p \cdot p + b_p \\ G &= \sigma(u_p^T \cdot p' + B_p) \\ f &= G * t' + (1 - G) * p' \end{aligned} \tag{1}$$

The gated summation approach is later further extended in Lu et al. (2018), referred to as *visual modulation gate*, to dynamically control the combination of visual and textual signals, as illustrated in Equation (2).

$$\begin{aligned} \beta_t &= \sigma(W_t \cdot t' + U_t \cdot p' + b_t) \\ \beta_p &= \sigma(W_p \cdot t' + U_p \cdot p' + b_p) \\ m &= \tanh(W_m * t' + U_m * p' + b_m) \\ f &= \beta_t * t' + \beta_p * m \end{aligned} \tag{2}$$

In this paper, we will refer to Miyamoto and Cho (2016)'s formula as *simple-gated fusion* and Lu et al. (2018)'s formula as *symmetric-gated fusion*.

- **Bilinear transformation** is a filter to integrate the information of two vectors into one vector. Mathematically we have $bilinear(t', p', dim) = t'^T \cdot M \cdot p' + b$, where dim is a hyper-parameter indicating the expected dimension of the output vector, M is a weight matrix of dimension (dim, $|t'|$, $|p'|$), and b is a bias vector of dimension dim. Again we concatenate t, p, and $bilinear(t', p', dim)$ for hate speech classification.

2.5 Attention mechanism

Attention mechanism was initially proposed in neural machine translation to dynamically adjust the focus on the source sentence (Bahdanau et al., 2014), but its application has been extended to many areas including multimodal fusion (Lu et al., 2018; Ghosal et al., 2018; Bagher Zadeh et al., 2018). The idea of attention is to use the information of a vector (called *query*) to weighted-sum a list of vectors (called $context$). Mathematically, it is implemented as Equation (3). The context vector is the 1D-convolution output $[c_1, c_2, ..., c_n]$ from text, while the query vector is the photo vector p'. W_a is a parameter to be learned.

$$\begin{aligned} s_i &= softmax(c_i^T \cdot W_a \cdot p') \quad i = 1, ..., n \\ a &= sum(s_i * c_i) \end{aligned} \tag{3}$$

- **Simple vs symmetric-gated fusion** Once we have the attention vector a, which is a weighted sum of the c_i vectors from text signal only, we will further apply fusion with the photo information g'. Again we can consider the fusion techniques described in Section 2.4. In this paper we experiment with both the simple- and symmetric-gated fusions, as bilinear is pretty expensive to run. We use the concatenation of t, g, and $gated\_fusion(a, g')$ for hate speech classification.

- **Sparsemax vs softmax** We also experiment with sparsemax (Martins and Astudillo, 2016), an alternative to softmax, in Equation (3) for calculating the attention vector a. Sparsemax is an activation function that outputs a vector of sparse probabilities where most of the values are zero, which could offer a more selective and compact attention focus.

- **Deep vs shallow** Another implementation detail is whether to back-propagate the derivatives when we clone the vectors $c_1, c_2, ..., c_n$ for attention calculation. *Shallow clone*, which makes a copy of c_i but stops the back-propagation (during attention), has less impact on the convolutions and word-embeddings; while *deep clone*, passing the derivatives through to convolutions and word embeddings, has a bigger impact.

3 Experiments

3.1 Data

We sample from seven months of user-reported data on a social network platform, which users re-

	Positive	Negative	Total
Train & dev	320K	58K	378K
Test	42K	11K	53K

Table 1: Data set size

port as hate speech. Every piece of content contains some text and exactly one photo. These data are then reviewed by the platform according to the community standard[1]. Contents that are determined to violate the community standard receive a positive label while otherwise negative. We use the last month of the data as test set, while the first six months of data are randomly split with 90% as training set and 10% as development set for determining early stopping. Table 1 gives some rough stats of the data set size.

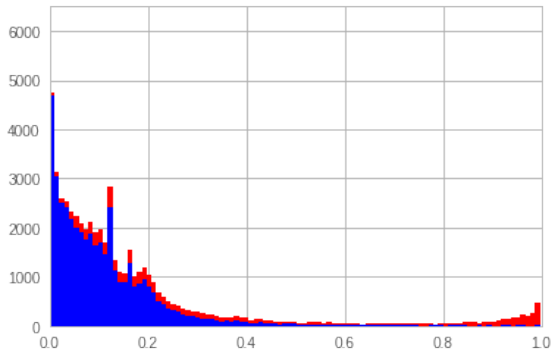
3.2 Hyper-parameters

In our experiments, the dimension of pre-trained word embeddings is 300. The new word embeddings after word-level MLP is also set at 300-dimension. Both word-level and classification-level dropout rates are set to 0.2. We use convolution windows [1, 3, 5] with 128 filters each. These parameters were tuned in pilot studies to optimize the baseline convolution text classification performance. The dimension of fusion vectors p', t', and a is set to be 128. We use ADAM optimizer with a learning rate of 0.001. We run 20 epochs for training and select the best model with development data.

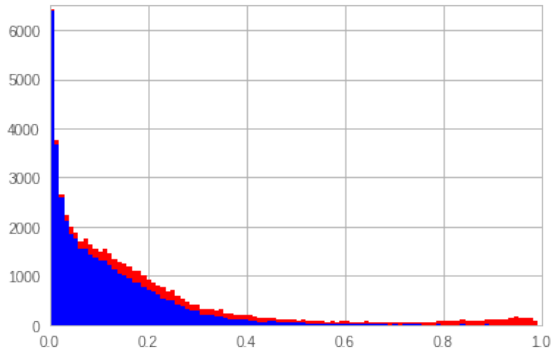
3.3 Results

A hate speech classifier can be used for many purposes, for example, to down-rank contents in newsfeed service, to proactively report contents for human reviews, to provide feedback for the creating users, or to provide warning message for consuming users. Generally a different decision threshold is needed for each scenario. Thus we use ROC-AUC as the performance metric in this paper, which measures the classifier's performance across all scoring points.

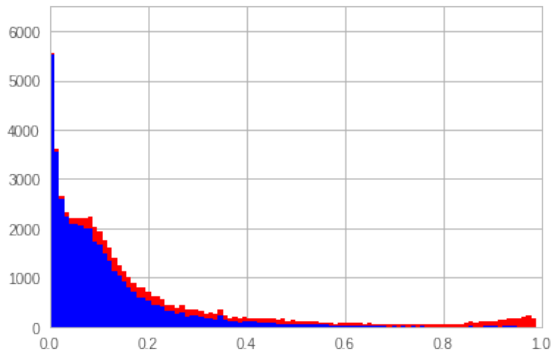
Results are shown in Table 2. The convolution text model gives us a baseline of 82.1. When concatenating the photo features p in the convolution training, we immediately get a nice boost to 84.0. We do not see a clear gain with additional fusion



(a) Convolutional text model



(b) Basic concatenation fusion model



(c) Attention model (symmetric, deep, sparsemax)

Figure 2: Score distribution histogram: blue for benign and red for hate speech. X axis is classifier score. Y axis is the count of items in the score segment.

using gated summation, either simple_gate or symmetric_gate. Bilinear transformation even brings the performance down. We speculate that there might be an overfitting issue with bilinear but we didn't investigate further as bilinear transformation runs very slow, about 8X to 10X slower than the other approaches.

Fusion using attention mechanism turn out to work pretty well. Generally, we see that deep cloning tends to perform better than shallow cloning, suggesting the benefit of deeper engagement of text and photo information. We see that sparsemax tends to perform better than softmax,

[1] https://www.facebook.com/communitystandards/hate_speech

Inputs	Additional Fusion Mode	Attention Mode		ROC-AUC
		max	clone	
t				82.1
t, g				84.0
t, g, fusion(t', g')	simple_gated			83.9
t, g, fusion(t', g')	symmetric_gated			84.1
t, g, fusion(t', g')	bilinear			82.7
t, g, fusion(attention(t', g'), g')	simple_gated	softmax	shallow	84.0
t, g, fusion(attention(t', g'), g')	simple_gated	softmax	deep	84.6
t, g, fusion(attention(t', g'), g')	simple_gated	sparsemax	shallow	84.3
t, g, fusion(attention(t', g'), g')	simple_gated	sparsemax	deep	84.6
t, g, fusion(attention(t', g'), g')	symmetric_gated	softmax	shallow	84.1
t, g, fusion(attention(t', g'), g')	symmetric_gated	softmax	deep	84.7
t, g, fusion(attention(t', g'), g')	symmetric_gated	sparsemax	shallow	84.3
t, g, fusion(attention(t', g'), g')	symmetric_gated	sparsemax	deep	84.8

Table 2: Experimental Results

suggesting the benefit of sparse weights on the summation of convolution outputs, which gives a higher focus on the important segments and totally ignores the trivial segments. We also see that symmetric_gate tends to perform better than simple_gate, suggesting the benefit of weighing the gated summation using both text and photo information (over using the photo channel only). Finally using the attention fusion with deep cloning, sparsemax, and symmetric_gate gives us a performance of 84.8, another nice improvement over basic concatenation, which is statistically significant at the 99% confidence level. In practice, we have found that improvement of 0.5 AUC would generally lead to observed production quality.

3.4 Discussion

Figure 2 shows the score distributions for three models: the baseline convolutional text model, the basic concatenation fusion model, and the attention fusion model with symmetric-gate, deep clone, and sparsemax. The baseline model has a spike at the score of about 0.13, which involves a significant false negative. Error analysis reveals that this is the section where posts contain none but OOV words.[2] Thus the text model extracts no useful signals but only uses the prior distribution which classifies all those posts as benign. With the concatenation of photo signals, the model can then learn to classify a piece of content as hate speech if there is a similar photo previously labelled as hate speech in the training data, which helps to improve recall.

We have also found cases where the photo signals help to improve precision as well. We found that when users have their posts deleted by the platform they sometimes make a screen shot (which is a photo) of the deleted post, and post it with some texts complaining or appealing about the community standard. The majority of these reposts are still hate speech, with a few exceptions where the original posts were deleted by mistakes. When training with text signals only, the model is overfitted towards text and it thus treats all the posts that complain or appeal the community standards as hate speech. With the integration of photo signals, the model actually learns that a piece of text complaining about community standard policy with a benign photo does not necessarily create hate speech, and so is able to avoid fitting all posts of policy complaining to hate speech.

The improvement of additional attention fusion over basic concatenation is a bit subtle. We observe that when both the text and the photo alone do not constitute a strong signal for hate speech, the basic concatenation model tends to classify the post as benign, although together they might create an impression of hate speech. With the additional attention fusion, the model would be able to highlight on some key phrases in the text to

[2]Texts on social network platforms are very noisy – there are typos, misspellings, long digits, foreign languages, and other online specials such as hashtags that we do not have in our limited vocabulary. A character model such as (Zhang et al., 2015) and (Bojanowski et al., 2017) should help to alleviate such problems though.

correctly recall some posts of hate speech. For example, with the text "If you look at the photo, I do think that they are disgusting parasites" and a photo of people, the attention model would be able to focus on the word "parasites" and catches it as hate speech. Sparsemax shines especially for longer texts. This is also shown in Figure 2 as the attention model is able to push more hate speech posts (in red) to the right hand side.

4 Conclusion

Interactions among users on social network platforms enable constructive and insightful conversations and civic participation; however, verbal abuse such as hate speech could also happen and lead to degraded user experience or even worse consequence. As users' interactions on today's social networks involve multiple modalities, in this paper we take the challenge of automatically identifying hate speech with deep multimodal technologies, expanding on previous research that mostly focuses on the text signal alone. We explore a number of fusion approaches to integrate text and photo signals, including concatenation, bilinear, gated summation, and attention fusion. We find that simply concatenating the text and photo embeddings immediately leads to a boost in performance, while additional attention fusion with symmetric gate, deep clone, and sparsemax brings further improvement. Our future work includes investigating fusion with multiple photos, and fusion with more modalities (such as audio and video).

References

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1924, Vancouver, Canada. Association for Computational Linguistics.

Jean-Benoit Delbrouck and Stéphane Dupont. 2017. An empirical study on the effectiveness of images in multimodal neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 910–919. Association for Computational Linguistics.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Golnoosh Farnadi, Jie Tang, Martine De Cock, and Marie-Francine Moens. 2018. User profiling through deep multimodal fusion. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pages 171–179, New York, NY, USA. ACM.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. *CoRR*, abs/1705.03122.

Deepanway Ghosal, Md Shad Akhtar, Dushyant Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Contextual inter-modal attention for multi-modal sentiment analysis. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3454–3466. Association for Computational Linguistics.

Yunchao Gong, Svetlana Lazebnik, Albert Gordo, and Florent Perronnin. 2013. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(12):2916–2929.

Isuru Gunasekara and Isar Nejadgholi. 2018. A review of standard text classification practices for multi-label toxicity identification of online content. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 21–25. Association for Computational Linguistics.

K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751. Association for Computational Linguistics.

Rohan Kshirsagar, Tyrus Cukuvac, Kathy McKeown, and Susan McGregor. 2018. Predictive embeddings for hate speech detection on twitter. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 26–32. Association for Computational Linguistics.

Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2017. Multi-modal summarization for asynchronous collection of text, image, audio and video. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1092–1102. Association for Computational Linguistics.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2247–2256. Association for Computational Linguistics.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1990–1999. Association for Computational Linguistics.

Rijul Magu and Jiebo Luo. 2018. Determining code words in euphemistic hate speech using word embedding networks. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 93–100. Association for Computational Linguistics.

André F. T. Martins and Ramón F. Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 1614–1623. JMLR.org.

Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2018. Neural character-based composition models for abuse detection. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 1–10. Association for Computational Linguistics.

Yasumasa Miyamoto and Kyunghyun Cho. 2016. Gated word-character recurrent language model. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1992–1997. Association for Computational Linguistics.

Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. 2014. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. 2018. A neural network catalyzer for multi-dimensional similarity search. *CoRR*, abs/1806.03198.

Magnus Sahlgren, Tim Isbister, and Fredrik Olsson. 2018. Learning representations for detecting abusive language. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 115–123. Association for Computational Linguistics.

Edmund Tong, Amir Zadeh, Cara Jones, and Louis-Philippe Morency. 2017. Combating human trafficking with multimodal deep models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1547–1556, Vancouver, Canada. Association for Computational Linguistics.

Prashanth Vijayaraghavan, Soroush Vosoughi, and Deb Roy. 2017. Twitter demographic classification using deep multi-modal multi-task learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 478–483, Vancouver, Canada. Association for Computational Linguistics.

Liwei Wang, Yin Li, and Svetlana Lazebnik. 2017. Learning two-branch neural networks for image-text matching tasks. *CoRR*, abs/1704.03470.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114. Association for Computational Linguistics.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *CoRR*, abs/1509.01626.

Haoti Zhong, Hao Li, Anna Squicciarini, Sarah Rajtmajer, Christopher Griffin, David Miller, and Cornelia Caragea. 2016. Content-driven detection of cyberbullying on the instagram social network. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 3952–3958. AAAI Press.