# A Review of Multimodal Sentiment Analysis Based on Deep Learning

**Anna wang,Baoshan Sun**

**Abstract—Sentiment analysis is an important research topic in the realization of human-computer interaction. It can play an important role in discovering social public opinion, analyzing stock markets, improving products and services, and assisting medical assistance. Early sentiment analysis mostly focused on emotion recognition in texts. With the rapid development of the Internet and the rise of short video social media platforms such as TikTokand Kwai, a single text can no longer satisfy people's emotional expression media. More and more people gradually begin to express their emotions and opinions through multi-modalities such as voice or video.Compared with the sentiment analysis that only focuses on single-modal information such as text, multi-modal sentiment analysis can make full use of the emotional cues between the single modalities to complement each other, thereby more accurately identifying the user's emotional tendency. This is the primary motivation behind our first of its kind,comprehensive literature review of the diverse field of affective computing.Furthermore, existing literature surveys lack a detailed discussion of state of the art in multimodal affect analysis frameworks, which this review aims to address.We starts from the various steps of multi-modal sentiment analysis and analyzes the technology and current situation of multi-modal sentiment analysis in detail.**

***Keywords*** **: Multimodal; Sentiment Analysis; Attention Mechanism; Long Short-Term Memory; Natural Language Processing;CNN**

## I. INTRODUCTION

In March 2020, We Are Social and Hootsuite released the latest research report on "Global Network Overview 2020". The report shows that globally, the total number of Internet users has exceeded 4.5 billion, and the number of registered social networking sites has exceeded 3.8 billion. Compared with previous years, netizens around the world spend more and more time online, spending an average of 6 hours and 43 minutes online every day. Some netizens spend nearly 40% of their daytime on social networks. The Internet plays an irreplaceable role in people's lives, and more and more people are gradually accustomed to using social networks to express their opinions and opinions on an event.

On social media like Weibo, Facebook, Twitter, hundreds of millions of data are generated every day. Most of these data appear in the form of text, pictures, and videos, which contain rich emotional information. The emotional information behind the research data has great application value. For example, the emotional information contained in shopping reviews can assist merchants to understand the market response and popularity of their products, and can also help merchants lock their target customers as soon as possible to achieve product targeting. Sexual push. For merchants, accurate sales are achieved, and for customers, the time consumed when shopping for products is reduced. The emotional information contained in hot news reviews can help the state and the government to grasp the public opinion orientation of netizens in a timely manner, analyze the scope of news dissemination, and guide the government to introduce relevant policies. It can also timely monitor rumors and false information on the Internet to create A civilized and orderly network environment.

At the same time, with the rise of short video platforms such as Douyin and Kuaishou, more and more people are gradually outputting their emotions and opinions through multi-modal methods such as videos, instead of just using text in social forums and blogs. Media posted [1]. For example, vlog, currently popular on social media, is a way to express emotional information in a multi-modal way. Bloggers will take videos of unpacking the products they get after shopping online to record their actual experience of the products. These videos usually include comparisons and evaluations of such products, and the emotional information contained in them helps other users to watch the video. Then make more rational consumer behavior. Therefore, analyzing the emotional tendency in videos is gaining continuous attention from the academic community, and has gradually become one of the hot research directions in the field of artificial intelligence [3].

Compared with sentiment analysis focusing only on monomodal information fields such as text, the advantage of combining multimodal information in video for sentiment analysis is that it can make up for the lack of monomodal information in the sequence segment. For monomodal emotion analysis, a major challenge is insufficient monomodal information and easy interference from the surrounding environment. For example, monomodal emotion recognition based on voice is easily disturbed by surrounding noise, and monomodal emotion recognition based on facial expressions is easy Affected by facial occlusion and image sensitivity, text-based monomodal emotion recognition is susceptible to interference from insufficient information. Multi-modal sentiment analysis just enables the emotional cues provided by each modality to complement each other [1]. The speech content of the speaker in the video can be regarded as a text modality, and the relevant emotional information can be extracted from the text opinions by using the words, phrases and the dependencies between them in the text. At the same time, the facial expression and voice tone of the speaker in the video can provide important emotional clues for the content of the speech, so as to more accurately

**This work was supported by the Tianjin Postgraduate Research and Innovation Project under Grant 2019YJSS030.**

identify the true emotional state of the speaker. Therefore, in order to improve the recognition accuracy, fusion of the multi-modal information combination of text, speech and image in the video can help create a better sentiment analysis model [2].

## II. RELATED TECHNOLOGY AND ANALYSIS

In this part, we mainly introduces the related technologies and methods of multi-modal sentiment analysis, expounds the basic principles of multi-modal sentiment analysis, briefly introduces the method of each modal feature extraction, and analyzes the fusion of commonly used multi-modal sentiment features in detail Methods. In addition, several commonly used techniques for multimodal sentiment analysis are introduced, including: RNN, CNN, and attention mechanism.

### *A. Basic Principles of Multimodal Sentiment Analysis*

In the multi-modal field, any sentiment analysis performed in a multi-modal scene can be called multi-modal sentiment analysis. The analysis of emotional information in the video generally focuses on text, audio, and visual. The text mainly refers to the content of the speaker's speech, and the text modal information is obtained by extracting the content of the speech in the entire video. The speech modal mainly studies the energy distribution of the frequency spectrum in each interval, and the difference of the frequency spectrum affects the difference of emotional expression. The image modality mainly studies the facial expressions of the characters in each frame of the video. This article will combine these three modalities to study their effects on emotional polarity.

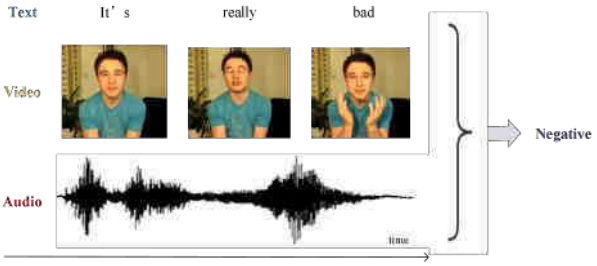


Fig.1 Example of multimodal sentiment classification

The sentiment analysis task is essentially a classification task, which outputs one or more corresponding sentiment labels for an input text. Assuming that sentiment is divided into C categories, the output of sentiment analysis is one of these C categories, which is usually represented by digital codes. And multimodal sentiment analysis is to analyze the multiple modalities contained in this type of problem. This article locks the research scene on the three modalities of text, audio and image in video comments. When using computer thinking to deal with classification problems, the most common method is to model all statistical information to estimate the probability distribution of the limited category. The final classification result is the one with the highest probability. Multi-modal sentiment analysis comprehensively considers the feature information from multiple modalities when classifying, and processes the fusion of these information to estimate the probability distribution of sentiment categories. An example of the principle of multi-modal emotion classification shown in Figure 1. The text modal of the video contains negative vocabulary, the waveform of the audio modal shows that the user's voice is low, and the image modal is angry and disappointed. expression. Based on the above information, there is a high probability that the user is negative when expressing emotions.

Sentiment analysis has many classification methods, which can be divided into coarse-grained and fine-grained based on the granularity. The object of coarse-grained sentiment analysis is generally the sentiment of the entire document or video segment, which is a simple division of the overall sentiment positive or negative, and is generally used to quickly obtain the user's overall attitude towards a product or event. Fine-grained sentiment analysis can divide emotions into finer dimensions, and the analysis unit is also smaller and more accurate. For example, when analyzing users' dining reviews, fine-grained sentiment analysis can derive different sentiments regarding the dining environment, taste of dishes, restaurant services. Usually used to evaluate the product in an all-round way and improve from the details.

### *B. Multi-modal Sentiment Feature Extraction Method*

Feature extraction is the first part of multi-modal sentiment analysis, and how fast the features extracted in this link directly determines the accuracy of subsequent model construction. This section analyzes the relevant technical routes and options for the extraction methods of multi-modal emotional features.

#### i. Facial emotion feature extraction

The shape of the facial features of a person will change to a certain extent according to the contraction of the facial muscles, such as the raising of the corners of the mouth and the frowning of the eyebrows, which form different facial expressions. Analyzing facial expressions uses image recognition technology. First, the input facial data is converted into digital information, and then the facial expression features are further extracted.

With the development of deep learning networks, methods based on deep feature extraction provide new ideas for image extraction. In-depth feature extraction can autonomously learn more essential sample features from the image, thereby making the sample more generalized. Compared with manual feature extraction methods, deep feature extraction can achieve autonomous learning and has better applicability to real business scenarios. In recent years, the research on image feature extraction has mostly focused on this field [4]. Shi et al. tried to use CNN to extract the depth features of the image, and achieved good results [5]; Siyue et al. proposed a CNN-based multi-layer image feature extraction network, which on the one hand extracts the local features of the image to highlight the expression Details, on the other hand, extract the global features from the image as a whole to highlight the high-level semantic features of expressions. Through the combination of two levels, more comprehensive expression features are extracted [6].

#### ii. Audio emotion feature extraction

In addition to facial features, the state of the user's voice is also a source of emotional expression. Researchers extract relevant emotional features from audio, mainly acoustic features and semantic features. Semantic information is to understand the overall meaning of speech according to the

meaning of the expressed utterance, and it can also be understood as a part of semantic understanding, but the semantic information is more suitable for the semantic understanding of the speaker. In reality, the extraction of the audio part of a video is not only speaker semantics, but also includes acoustic information. Acoustic information includes tone, intonation, and background sounds used to render emotions. Acoustic information is often considered to better reflect changes in emotions. In this article, the mode that conveys emotions through sound is called audio mode.

The extraction of audio features generally includes three types: prosodic features, spectral features and other non-linear features. In addition to Low-level Descriptors (LLD), such as pitch frequency, short-term energy, short-term zero-crossing rate, short-term autocorrelation coefficient, short-term average amplitude difference, spectrogram and short-term power spectral density . Spectral characteristics can be seen as a filter for the vocal tract, which reflects the vocal characteristics of the glottis, which is regarded as a manifestation of the correlation between the change of the vocal tract shape and the vocal movement. In addition, spectrum-based features include Mel frequency cepstrum coefficient linear prediction cepstrum coefficient and zero-crossing rate. In order to reduce the difficulty of feature extraction, the researchers developed OpenSmile open source software [7], which can simplify the audio feature extraction process and improve the extraction accuracy. The comprehensive feature extraction types make it widely used in feature extraction.

iii. Text emotion feature extraction

In the early days of text feature extraction, researchers used bag of words (BOW) and TD-IDF to represent text features. Because the above methods cannot reflect the word order and semantics of sentences, and even the matrix sparse problem caused by too large dimensions brings great difficulties to training, Mikolov et al. proposed the concept of distributed word vectors to map each word in the text To a fixed-dimensional vector, used to represent the word embedding of this word, called word2vec[8]. The word vector consists of two training models, namely continuous bag-of-words model (CBOW) and Skip-gram model. The principle of the former is to predict the current word according to the context of a word, and the principle of the latter is to predict the current word according to the current word. For words that appear in the context, the intermediate matrix obtained by training is the word vector matrix. Although word2vec can improve part of the downstream tasks of natural language processing, its static word vector properties make it unable to deal with polysemy problems well. In order to solve this problem, Peters et al. proposed the ELMo model, which uses a forward and reverse LSTM coding, which cannot learn the embedding representation of words, and can also obtain different levels of lexical features at different layers [9] . With the rise of pre-training models, the Google team proposed a self-encoding language model BERT [10] in 2018, which uses a bidirectional Transformer as a text feature extractor, and uses Masked LM and NSP methods to capture words and sentences respectively Representation of levels. This model has made huge improvements in eleven text-related tasks, and has become another innovation in the field of natural language processing.

### C. *Multi-modal Sentiment Feature Fusion Strategy*

Multi-modal fusion is to use the correlation between information to fuse information between different modalities. The fusion strategies proposed in the early research work [11] can be divided into three categories, namely feature layer fusion, fusion decision layer fusion and model layer fusion. These fusion methods are introduced separately below.

i. Feature layer fusion

Feature layer fusion [12] is also called early fusion, which extracts the feature information of each modal and connects them to form a total feature vector for emotion recognition. As shown in Figure 2, this method generally expresses the extracted text, audio and image features in a unified form during feature extraction, and then stitches them into a total vector and inputs it to the classifier for classification. This fusion strategy focuses on the interconnection between various modalities, and has been widely used in the field of multimodal sentiment analysis [13]. However, this strategy does not take into account the differences in the fusion of the various modes, and cannot accurately express the rich associations between the various modes. On the other hand, the feature layer fusion cannot handle the time synchronization of each mode well, and as the number of modes increases, the problem of too large dimensions and difficult to learn may appear.

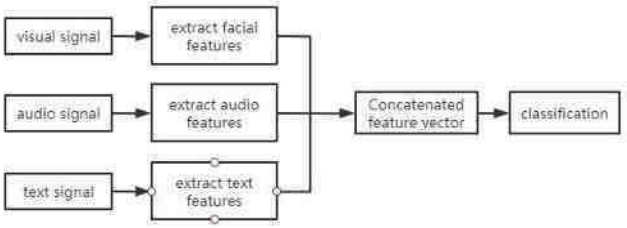


Fig.2 Feature layer fusion

ii. Decision level fusion

Decision layer fusion [14] is also called post fusion. It constructs its own classifiers for each mode, and then fusions the results of each classifier according to the decision rules to obtain the final emotion detection result. As shown in Figure 3, after extracting each modal feature, for different modalities, a classifier suitable for each modal will be used to recognize single-modal features. The final multi-modal emotion detection results need to be based on each The modal detection results are determined by certain decision-making strategies. Decision strategies such as Kalman filter and weighted evaluation are often used in this fusion. Other commonly used decision rules also include averaging, weighting, maximum, minimum, and product.

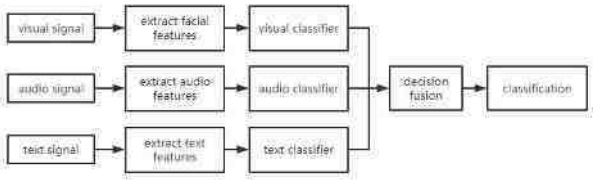


Fig.3 Decision level fusion

Compared with the feature layer fusion, the decision layer

fusion does not need to consider the initial state of each mode. The first classification operation makes the final results obtained by the model consistent, so it is easier to operate. The decision-making layer fusion fully considers the differences in modal fusion, so that each modalities such as speech, text, and image can choose the most suitable classifier for classification, but it ignores the relationship between emotional features and models separately. It loses the mutual information between modalities, and the effect of multi-modal emotion recognition is not ideal. On the other hand, the method of learning the classifiers separately also increases the computational difficulty, making the learning process lengthy and time-consuming.

iii. Mixed layer fusion

Hybrid layer fusion [15] combines the two fusion strategies of feature layer fusion and decision layer fusion, and has the advantages of both fusion strategies. Literature [15] proposed a method based on hybrid multimodal fusion. As shown in Figure 2-4, the model consists of two processing methods, one is to extract the MFCC features of the audio from the audio signal, and then use the speech recognition module to obtain the text features of the signal, and finally SVM makes classification prediction on the language features. The other method first extracts audio and image signals using feature layer fusion, then extracts the features of their respective modalities, and finally sends these two features into the two-way LSTM to obtain the classification results. Finally, the processing results of the two methods are predicted by the decision-making layer fusion method, and the final multi-modal emotion recognition result is obtained.

*D. Techniques of Multi-modal Sentiment Analysis*

With the development of deep learning, methods such as CNN, RNN, and attention mechanism are often used for multimodal sentiment analysis. These techniques are introduced below.

i. CNN

Convolutional Neural Networks (CNN) [16] is a type of feedforward neural network that includes convolution calculations in deep learning. Its essence is similar to the multilayer perceptron of artificial neural networks. It is often used to analyze visual images. CNN has two characteristics: local connection and weight sharing. Through these two methods, not only the number of network parameters is reduced, making it easier to train and optimize, but also the complexity of the network is reduced and overfitting is avoided.

The common convolutional neural network structure generally has a five-layer structure: the first layer is the data input layer (Input layer), and the processing of this layer is mainly to preprocess the original image data, including de-averaging, normalization, Operations such as dimensionality reduction; the second layer is the convolution calculation layer (CONV layer), this layer is the most important layer of the convolutional neural network, which is essentially the point multiplication and summation of two matrices, one of which is The input data matrix, the other matrix is the convolution kernel, and the sum result can be regarded as the local features extracted from the input image; the third layer is the ReLU activation layer (ReLU layer), which uses the output of the convolution layer Non-linear mapping obtains the activated matrix; the fourth layer of pooling layer (Pooling layer) sandwiched between the continuous convolutional layer, used to compress the amount of data and parameters, reduce over-fitting, pooling layer method There are maximum pooling (Max pooling) and average pooling (Average pooling).; the last layer is a fully connected layer (FC layer), this layer uses the softmax function, and the output result is the image feature extracted by CNN.

ii. RNN

Recurrent Neural Network (RNN) [17] is a recurrent neural network with recurrent units. This network has the function of short-term memory and is generally used to process text or speech and other tasks that take time as a sequence input. Each cyclic unit in the RNN is connected in a chained manner. During the cycle, the cyclic unit can receive its own information input from the network, as well as the output information of other units, thus forming a loop. The structure diagram of RNN deployment is shown in Figure 2-5.

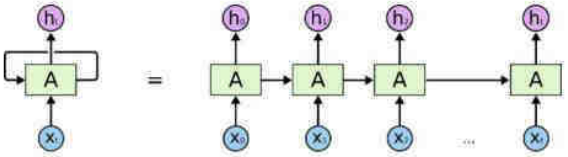


Fig.4 The structure diagram of RNN

RNN uses backpropagation to learn the parameters that need to be updated in the network. For a long sequence in the input network, the error is passed forward in the reverse order, and the gradient disappears or the gradient explodes after multiple time steps. In order to cope with the phenomenon of gradient disappearance and gradient explosion that may occur in RNN, a gating mechanism is introduced in the recurrent unit. By controlling the accumulation of information inside the RNN, it can not lose long-distance information during back propagation, and Can choose to forget some unimportant information to prevent information overload. The networks improved in this way include long and short-term memory networks (LSTM) and gated recurrent units (GRU).

iii. Attention mechanism

The attention mechanism was inspired by the human attention mechanism and was first applied in the field of visual images. Researchers found that when people observe an image, they cannot notice every position in the image, nor can they remember the pixels at every position in the picture, but pay attention to specific parts according to their own preferences or needs. . Moreover, if people find the part they want to watch in some part of a picture, they will usually pay attention to several similar areas when the same picture appears in the future.

The attention mechanism that is really popular in the academic world is a paper published by the Google DeepMind team in 2014 [18], in which the attention mechanism is used on the RNN model to classify images. Subsequently, in the paper [19], Bahdanau et al. used the attention mechanism to train translation and alignment

together on machine translation tasks. Their work applied the attention mechanism to the field of natural language processing for the first time, and then, the introduction of the attention mechanism in RNN began to extend to a variety of natural language processing tasks.

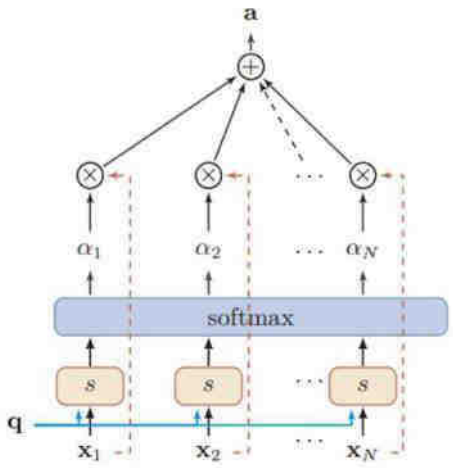


Fig.5 The process of soft attention mechanism

According to the research on attention mechanism in recent years, it can be divided into soft attention mechanism, hard attention mechanism, local attention mechanism, global attention mechanism and self-attention.

Compared with the hard attention mechanism only pays attention to the information in a certain position, the soft attention mechanism can pay attention to all the input information. When selecting information, it does not select only one from all the input information, but calculates The input information is expected under the attention distribution, and then the output result is sent to the neural network.

## III. Conclusion

With the rapid increase in the number of smart phones and the rise of social networks, the ways of information dissemination have become diverse and convenient, and people can express their opinions and ideas through the Internet without being restricted by space and time. With the rise of video media such as Tiktok, lots of users are gradually expressing their emotions and opinions through video, not just in the form of text.

Compared with sentiment analysis focusing on single-modal information such as text, the advantage of combining multi-modal information in video for sentiment analysis is that the emotional cues provided by each modal can complement each other. The speech content of the speaker in the video can be regarded as a text modality, and the relevant emotional information can be extracted from the text opinions by using the words, phrases and the dependencies between them in the text. At the same time, the facial expression and voice tone of the speaker in the video can provide important emotional clues for the content of the speech, so as to more accurately identify the true emotional state of the speaker.

In this paper,we analyze the commonly used algorithms in multimodal sentiment analysis, summarize the existing technical routes, and provide help for the further development of this field.

## Acknowledgment

This work was supported by the Tianjin Postgraduate Research and Innovation Project under Grant 2019YJSS030.

## References

[1] Poria S, Cambria E, Hazarika D, et al. Context-dependent sentiment analysis in user-generated videos[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017, 1: 873-883.

[2] Poria S , Cambria E , Bajpai R , et al. A review of affective computing: From unimodal analysis to multimodal fusion[J]. Information Fusion, 2017, 37:98-125.

[3] Wang A, Sun B, Jin R. An Improved Model of Multi-attention LSTM for Multimodal Sentiment Analysis[C]//Proceedings of the 2019 3rd International Conference on Computer Science and Artificial Intelligence. 2019: 144-149.

[4] Bengio, Yoshua, Courville, et al. Representation Learning: A Review and New Perspectives[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013, 35(8):1798-1828.

[5] Shi B , Bai X , Yao C . Script identification in the wild via discriminative convolutional neural network[M]. Elsevier Science Inc. 2016,52:448-458.

[6] Siyue, Xie, Haifeng, et al. Facial Expression Recognition Using Hierarchical Features With Deep Comprehensive Multipatches Aggregation Convolutional Neural Networks[J]. IEEE Transactions on Multimedia, 2019, 21(1):211-220.

[7] Eyben F, Wöllmer M, Schuller B. Opensmile: the munich versatile and fast open-source audio feature extractor[C]//Proceedings of the 18th ACM international conference on Multimedia. ACM, 2010: 1459-1462.

[8] Mikolov T , Chen K , Corrado G , et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer ence, 2013.

[9] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[J]. arXiv preprint arXiv:1802.05365, 2018.

[10] Devlin J , Chang M W , Lee K , et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.

[11] Zhalehpour S, Onder O, Akhtar Z, et al. BAUM-1: A spontaneous audio-visual face database of affective and mental states[J]. IEEE Transactions on Affective Computing, 2016, 8(3): 300-313.

[12] Poria S, Cambria E, Gelbukh A. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis[C]//Proceedings of the 2015 conference on empirical methods in natural language processing. 2015: 2539-2544.

[13] Wang Y, Guan L, Venetsanopoulos A N. Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition[J]. IEEE Transactions on Multimedia, 2012, 14(3): 597-607.

[14] Sahoo S, Routray A. Emotion recognition from audio-visual data using rule based decision level fusion[C]//2016 IEEE Students' Technology Symposium (TechSym). IEEE, 2016: 7-12.

[15] Wöllmer M, Weninger F, Knaup T, et al. Youtube movie reviews: Sentiment analysis in an audio-visual context[J]. IEEE Intelligent Systems, 2013, 28(3): 46-53.

[16] Kim Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1408.5882, 2014.

[17] Zaremba W, Sutskever I, Vinyals O. Recurrent neural network regularization[J]. arXiv preprint arXiv:1409.2329, 2014.

[18] Mnih V, Heess N, Graves A. Recurrent models of visual attention[C]//Advances in neural information processing systems. 2014: 2204-2212.

[19] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.

[20] Akhtar M S, Chauhan D S, Ghosal D, et al. Multi-task learning for multi-modal emotion recognition and sentiment analysis[J]. arXiv preprint arXiv:1905.05812, 2019.