Label-Aware Attention for Fine-Grained Emotion Detection

Ethan Wu
yongyiw@andrew.cmu.edu

Yilin Wang
yilinwan@andrew.cmu.edu

Ziqi Liu
ziqil2@andrew.cmu.edu

Abstract

Recognizing underlying emotions from plain text enables human to build more intelligent systems, abridging the gap between us and machines. Working on by far the largest manually annotated dataset for emotion detection GoEmotions, we propose a label-aware attention mechanism, where label semantics are leveraged to integrate contextualized representations and inform classification decisions. Combined with the class-balanced loss, our method improves the macro F1 score by 0.06 compared to the baseline classifier, achieving state-of-the-art performance on this emotion detection benchmark. [1]

1 Introduction

Emotion detection is one of the fundamental tasks for more advanced applications of Natural Language Processing and Human-Computer Interaction, ranging from dialog systems to spam detection. This task aims to assign emotion labels that could be easily understood by humans to given input texts.

GoEmotions is a recent work that contains by far the largest manually annotated data of 58k English Reddit comments, labeled for 27 emotion categories or "neutral" Demszky et al. [2020]. The baseline model proposed by the authors follows the conventional classification regime, where a pretrained language model (in this case, cased BERT BASE) extracts the contextual representation of `[CLS]` token, which is then fed into a multilayer perceptron (MLP) to compute, for each emotion, the probability of existence in the given text. However, the fine-grained taxonomy in the GoEmotions dataset presents challenge for the baseline classifier to distinguish nuanced emotions, such as anger and annoyance, so the `[CLS]` token alone may not be able to carry enough information to yield desired performance across different labels.

In this work, we propose a label-aware attention mechanism where each emotion (classification label) maintains its own latent representation, and uses it to query and aggregate the contextual representation of all tokens from the input text. By computing attention distribution for each emotion independently, we hope the semantic information carried by the label would guide the classifier to focus on different aspects of the same text. Our method is reminiscent of AttentionXML (You et al. [2019]) and LEAM (Wang et al. [2018]) but is much easier to implement and train by avoiding the probabilistic label tree or the temporal convolution. Our model is trained under the class-balanced loss proposed by Cui et al. [2019], which helped alleviate the uneven distribution of emotion labels in the GoEmotions dataset.

The paper is organized as follows: Section 3 introduces the design of our proposed classifier. We then present experiment details Section 4 and show our method performs significantly better than the baseline. After that, we review related work in emotion detection in Section 2 and conclude our work with a discussion in Section 5.

[1] Data and code available at https://github.com/yongyi-wu/nlp-group.

2 Related Work

Emotion Detection Taxonomy Traditionally, researchers use two emotion categorization taxonomies: categorical and dimensional. On the categorical front, Ekman's model divides emotions into 6 fundamental categories (fear, disgust, surprise, joy, anger, sadness), while Plutchik's emotion wheels divide emotions into 8 fundamental types, with each being able to subdivide into subcategories. For dimensional approaches, emotion can be expressed continuously in two dimensions: valence (degree of positivity) and arousal (degree of intensity). GoEmotions dataset extends this traditional taxonomy to 27 subcategories (excluding neutral), allowing models to learn more fine-grained distinctions between emotions in the corpora.

LEAM This model Gaonkar et al. [2020] deploys a different way to compute the attention weights of the emotion labels towards the input text. To help highlight the emotional information contained in different portions of the input text, instead of using the bilinear key-query attention mechanism as in our model, this approach uses the cosine similarity between each label representations (initialized by the BERT word embeddings of the label word) and each input word representations. The resultant signals are processed by a convolution layer followed by a max pool layer to extract the most salient similarity signal. Finally, we linearly transform the extracted signal and softmax to get the attention weights between label emotions and input words.

Extreme-scale Multi-label Classification In extreme-scale multi-label classification, it is a common practice to introduce label semantics to help model distinguish nuance between different classes (Zhang et al. [2018], You et al. [2019], Chang et al. [2020]). AttentionXML builds a probabilistic label tree to perform classification in a divide-and-conquer fashion wherein a multi-label attention guides labels (or sets of labels) to focus on the most relevant part of a sentence.

Modeling Label-semantic for predicting emotion reactions Gaonkar et al. found that adding the pretrained label semantics embeddings as input and incorporating the learned correlation between emotions in the loss function improves the performance over the baseline BERT model Gaonkar et al. [2020]. Augmenting this method and applying semi-supervised techniques on the unlabeled portion of ROCStories gives the SOTA model with F1-score being 65.88, which is 4.92 points higher than the baseline BERT model.

3 Method

3.1 Overview

Our proposed classifier consists of two modules, a label-aware attention and a multilayer perceptron, building on top of contextualized representations extracted by a pretrained language model.

As illustrated in Figure 1, each of emotion labels, denoted by superscripts $(0), (1), \cdots, (C)$, computes its attention weights to all tokens from the last layer of the language model. For each label, the aggregated vector representation is then concatenated with the representation of `[CLS]` to form a label-aware sentence embedding. Finally, this embedding is fed into a feedforward network to produce the score for each label.

By allowing each label to attend to different components of a sentence, the label-aware attention enhances the expressiveness of the sentence embedding. Moreover, because the attention weights are computed independently for each label, similar to Vaswani et al. [2017], our method is embarrassingly parallelizable and easy to implement, capable of scaling to datasets with even more fine-grained annotation.

We also attempt to alleviate the class unbalance problem in the GoEmotions dataset, by replacing the cross-entropy loss with the class-balanced (CB) loss proposed by Cui et al. [2019]. The CB loss addresses the class unbalance problem by up-weighting the mistakes in less common classes.

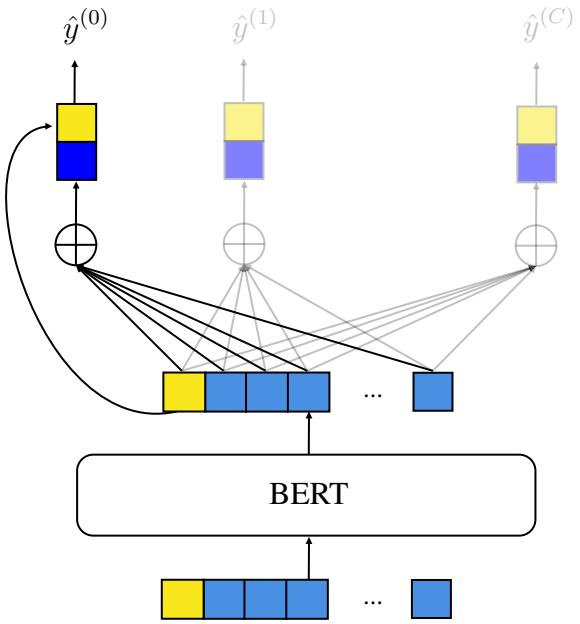

Figure 1: Visualization of the label-aware attention mechanism. The yellow block corresponds to the `[CLS]` token or its vector representation $\mathbf{h}_1$.

3.2 Label-Aware Attention

Given a tokenized text sequence $\mathbf{x}$ of length L, we first obtain its contextualized representation from a pretrained language model: $\mathbf{H} = [\mathbf{h}_1 \mathbf{h}_2 \cdots \mathbf{h}_L]^T \in \mathbb{R}^{L \times d}$, where $\mathbf{h}_i$ is a d-dimensional vector representation for i-th token.

For an emotion i, we maintain a trainable representation $\mathbf{e}^{(i)} \in \mathbb{R}^{d'}$ and use it to query $\mathbf{H}$, where the attention scores are computed by bilinear products via a trainable matrix $\mathbf{W} \in \mathbb{R}^{d' \times d}$ to help align vector space of $\{\mathbf{e}^{(i)}\}_{i=1}^{C}$ and $\{\mathbf{h}_j\}_{j=1}^{L}$. Meanwhile, We omit linear projections of queries, keys and values as in Vaswani et al. [2017]. A `softmax` layer is applied to the bilinear products to compute the attention weight distribution,

$$\alpha_j^{(i)} = \texttt{softmax}\Big(\mathbf{e}^{(i)T} \mathbf{W}_{\text{attn}} \mathbf{h}_j\Big) = \frac{\exp(\mathbf{e}^{(i)T} \mathbf{W}_{\text{attn}} \mathbf{h}_j)}{\sum_{k=1}^{L} \exp(\mathbf{e}^{(i)T} \mathbf{W}_{\text{attn}} \mathbf{h}_k)}$$

where $\alpha_j^{(i)}$ denotes the attention weight assigned by emotion i to the j-th token.

Finally, we aggregate the contextualized representations using weights computed above to obtain label-aware sentence representations (as indicated by dark blue boxes in Figure 1):

$$\tilde{\mathbf{h}}^{(i)} = \sum_{k=1}^{L} \alpha_k^{(i)} \mathbf{h}_k.$$

3.3 Classification Head

While leveraging label semantics to pool last-layer representations from the pretrained language model, we do not waste the language model's intrinsic ability to extract sentence-level representation. Therefore, for an emotion i, we concatenate the label-aware sentence embedding $\tilde{\mathbf{h}}^{(i)}$ with the typical embedding from `[CLS]` token $\mathbf{h}_1$. The concatenated vector is passed into a MLP, which outputs the scalar probability that emotion i exists in the input text:

$$\hat{y}^{(i)} = \sigma\Big(\text{FFN}_i([\tilde{\mathbf{h}}^{(i)} \| \mathbf{h}_1])\Big).$$

We tie weights of classification head FFN_i for all emotion i to encourage the attention module to focus on different sets of contextualized embeddings in $\mathbf{H}$. This also prevents the classifier size grows linearly with respect to the number of labels. However, it is possible to inject stronger inductive bias and design specialized head for each label. We leave this idea for future works.

3.4 Class-Balanced Loss based on ENS

In previous error analysis, we found that the baseline model often fails to distinguish low-resourced emotions from high-resourced emotions. This is an issue as the distribution of emotions in the goemotions dataset is unbalanced. To address this issue, we used the class-balanced loss introduced by Cui et al. [2019]. The effective number of samples (ENS) is defined as $E_c = (1 - \beta^{n_c})/(1 - \beta)$, where β is a hyper-parameter (usually between 0.9 and 1) and n_c is number of samples with emotion c. The class-balanced (CB) loss $\mathcal{L}_{CB}(p, y)$ weights the original loss by the ENS as follows:

$$\mathcal{L}_{CB}(p, y) = \frac{1}{E_c}\mathcal{L}(p, y) = \frac{1 - \beta}{1 - \beta^{n_c}}\mathcal{L}(p, y)$$

where $\mathcal{L}(p, y)$ is the cross entropy loss.

4 Experiments

4.1 Dataset

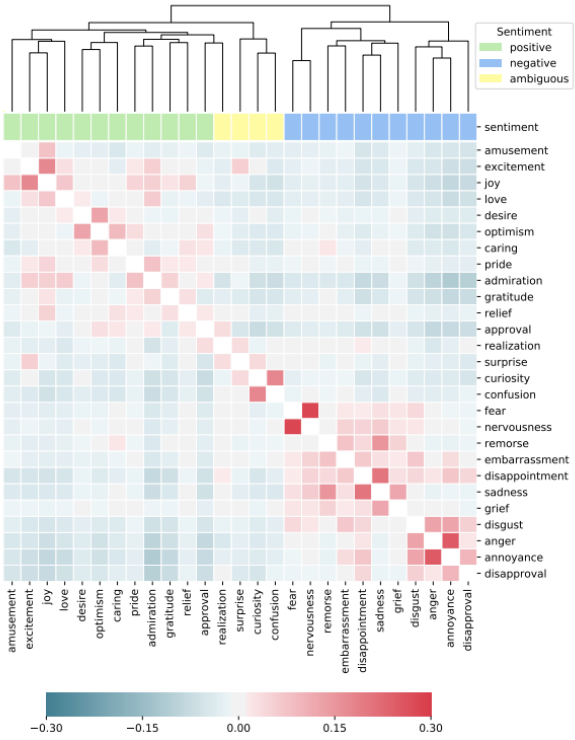

Figure 2: Annotation results of the GoEmotions dataset. The heatmap shows the correlation between ratings for each emotion. The dendrogram represents the a hierarchical clustering of the ratings Demszky et al. [2020].

GoEmotions is a large dataset that consists of 58K English Reddit comments, each of which is labeled with one or more of 28 emotions (including "neutral"). While GoEmotions characterizes a multi-label emotion classification task, most (83%) of the samples have a single label.

Despite some degree of inter-rater agreement, the heatmap from Figure 2 shows, some nuanced emotions, such as "disgust", "anger" ,and "annoyance" (bottom right corner), present even some

annotation challenges. Such a fine-grained emotion taxonomy requires more a careful design of classifiers.

4.2 Model and Parameter

To draw a fair comparison and demonstrate the benefit of label-aware attention, we use the same language model as Demszky et al. [2020], which is a pretrained cased BERT BASE. The vector representation for each label is initialized to be a 768-dimensional word embedding of the corresponding emotion. However, the label representations are finetuned separately from the word embedding. Finally, the feedforward network in the classification head is a single linear layer with a bias term.

When finetuning the pretrained BERT model, we use the same set of hyperparameters as Demszky et al. [2020]. Specifically, the batch size is 16 and the learning rate is 5e-5 with the slanted triangular scheduler. For the CB loss, we choose $\beta = 0.95$. The finetuning is run for 4 epochs with no parameter held frozen. We run 10 experiments for each model under different seeds, and we average the performance over 10 experiments.

4.3 Results

In Table 1, we see that our model, which is equipped with label-aware attention (LAA) mechanism and trained under class-balanced (CB) loss, outperforms the baseline model. The Marco-average F1 of our model is 0.06 higher than that of the baseline model. Moreover, our model shows much lower variance under different seeds, which is an indication of robust design.

Aside from our best model, we also run two experiments which respectively involve LAA and LEAM only, and the results are presented in Appendix Table 4. We see that our three models all outperform the baseline model, and the use of CB loss improved performance for the LAA model.

4.4 Error Analysis

As discussed in table 1, our methods outperform the baseline model. We see that our model performs better in almost all emotions compared to the baseline model. In the last report, we argued that the baseline model often fails to distinguish between “neutral" and other emotions (see Appendix, table 5). In this section, we will examine qualitatively how our best-performing model (LAA with CB loss) is better than the baseline model.

For the sake of simplicity, we use the following procedures in later error analysis. If the sample contains only has 1 true label, then it is considered correctly classified if the top 1 predicted label is the same as the true label. If the sample contains multiple true labels, it is considered correct if all its label is contained in the top 3 predicted label.

Table 2 present the top 8 circumstances where the text is misclassified by the baseline model but are correctly classified by our model. We see that our model effectively prevents “neutral" from being misclassified as other emotions. However, there is little improvement in preventing other emotions from being misclassified as “neutral".

Table 3 presents two representative sentences that are misclassifed by the baseline model but are correctly classified by our model. We see that the two sentences both employ rhetoric techniques that would make them liable to being misclassified for models that do not fully understand the semantics but instead rely on certain features of the sentence to make the classification (e.g., “?" in text (B)). The fact that they are correctly classified by our model suggests that the label-aware attention mechanism effectively facilitated natural language understanding (NLU) in our model.

5 Conclusion

In conclusion, we propose the Label-Aware Attention(LAA) mechanism, which incorporates label semantics in the model. To address the class unbalance problem, we adopt the class-balanced (CB) loss instead of the regular cross-entropy loss. Among all experiments we performed, the LAA model trained under CB loss performed the best, leading to a 0.06 increase in Marco-average F1 score over the baseline model. Error analysis suggests that the LAA mechanism facilitated NLU in our model.

Emotions	Baseline	LAA w/ CB Loss (ours)
Admiration	0.65	**0.67**
Amusement	0.80	**0.81**
Anger	0.47	**0.49**
Annoyance	0.34	**0.37**
Approval	0.36	**0.41**
Caring	0.39	**0.43**
Confusion	0.37	**0.42**
Curiosity	0.54	**0.56**
Desire	**0.49**	**0.49**
Disappointment	0.28	**0.33**
Disapproval	0.39	**0.41**
Disgust	0.45	**0.48**
Embarrassment	0.43	**0.45**
Excitement	0.34	**0.41**
Fear	0.60	**0.67**
Gratitude	0.86	**0.90**
Grief	0.00	**0.42**
Joy	0.51	**0.60**
Love	0.78	**0.79**
Nervousness	**0.34**	**0.34**
Neutral	**0.68**	0.67
Optimism	0.51	**0.56**
Pride	0.36	0.49
Realization	0.21	**0.26**
Relief	0.15	**0.36**
Remorse	0.66	**0.68**
Sadness	0.49	**0.53**
Surprise	0.50	**0.54**
Macro Avg.	0.46	**0.52**
Std	0.19	0.01

Table 1: F1 scores across all emotions in GoEmotions dataset, in comparison with baseline. Our method combines the Label-Aware Attention (LAA) with class-balanced (CB) cross-entropy loss.

True Label	Predicted Label	Occurrence
neutral	amusement	50
neutral	anger	36
neutral	approval	32
neutral	annoyance	23
neutral	curiosity	16
curiosity	anger	13
approval	anger	11
neutral	admiration	11

Table 2: Top 8 common circumstances where the text is misclassified by the baseline model but are correctly classified by our model

ID	Sentence	True Label	Predicted Label
(A)	Everyone knows the good [NAME] are the dirty little devil's	neutral	annoyance
(B)	[NAME]? never met one those.	neutral	curiosity

Table 3: Two representative mistakes of the baseline model, correctly classified by our model

In addition, our model shows much lower variance under different seeds compared to the baseline model, which indicates robustness in the design.

From previous error analyses, we argued that the baseline model often fails to distinguish “neutral" and other emotions. The current error analysis suggests that our proposed method effectively prevents “neutral" from being classified as other emotions. However, our model does not appear effective in preventing other emotions from being classified as “neutral". Future work should aim to improve the model’s ability to distinguish other emotions from “neutral". A generative adversarial network (GAN) might be used to achieve this objective. Alternatively, we could consider a two-stage model where the model first performs a binary classification task between “neutral" and non-neutral, and then assign labels to texts that are classified as non-neutral by the first-stage classifier.

References

Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit Dhillon. Taming pretrained transformers for extreme multi-label text classification, 2020.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples, 2019.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions, 2020.

Radhika Gaonkar, Heeyoung Kwon, Mohaddeseh Bastan, Niranjan Balasubramanian, and Nathanael Chambers. Modeling label semantics for predicting emotional reactions, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

Guoyin Wang, Chunyuan Li, Wenlin Wang, Yizhe Zhang, Dinghan Shen, Xinyuan Zhang, Ricardo Henao, and Lawrence Carin. Joint embedding of words and labels for text classification, 2018.

Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification, 2019.

Wenjie Zhang, Junchi Yan, Xiangfeng Wang, and Hongyuan Zha. Deep extreme multi-label learning, 2018.

A Additional Results

Emotions	Baseline	LAA w/CB Loss	LAA w/CE Loss	LEAM w/CE Loss
Admiration	0.65	0.67	**0.68**	0.67
Amusement	0.80	**0.81**	0.80	0.79
Anger	0.47	**0.49**	0.48	0.46
Annoyance	0.34	**0.37**	0.36	0.33
Approval	0.36	**0.41**	0.40	0.39
Caring	0.39	0.43	**0.44**	0.43
Confusion	0.37	**0.42**	**0.42**	**0.42**
Curiosity	0.54	**0.56**	0.55	0.53
Desire	0.49	0.49	0.48	**0.53**
Disappointment	0.28	**0.33**	**0.33**	0.29
Disapproval	0.39	**0.41**	**0.41**	0.40
Disgust	0.45	**0.48**	0.46	0.46
Embarrassment	0.43	**0.45**	**0.45**	0.44
Excitement	0.34	0.41	0.41	**0.43**
Fear	0.60	0.67	**0.68**	**0.68**
Gratitude	0.86	0.90	0.90	**0.91**
Grief	0.00	**0.42**	0.35	0.21
Joy	0.51	**0.60**	0.59	0.59
Love	0.78	0.79	**0.80**	0.79
Nervousness	**0.34**	**0.34**	**0.34**	**0.34**
Neutral	**0.68**	0.67	0.66	0.65
Optimism	0.51	**0.56**	0.55	0.54
Pride	0.36	0.49	**0.50**	0.45
Realization	0.21	**0.26**	0.25	0.23
Relief	0.15	0.36	0.35	**0.38**
Remorse	0.66	**0.68**	0.67	0.65
Sadness	0.49	0.53	0.53	**0.54**
Surprise	0.50	0.54	**0.55**	0.52
Macro Avg.	0.46	**0.52**	0.51	0.50
Std	0.19	0.01	0.01	0.01

Table 4: F1 scores across all emotions in GoEmotions dataset, in comparison with baseline. Column three to 5 are respectively F1 scores for Label-Aware Attention (LAA) with class-balanced (CB) loss; LAA with cross entropy (CE) loss; LEAM with CE loss.

Table 5: Top 7 common error for the baseline model

True Label	Predicted Label	Occurrence
neutral	approval	89
neutral	disapproval	72
neutral	curiosity	68
approval	neutral	67
neutral	annoyance	61
disapproval	neutral	56
annoyance	neutral	55