Modality-specific Learning Rates for Effective Multimodal Additive Late-fusion

Yiqun Yao
Computer Science and Engineering
University of Michigan
yaoyq@umich.edu

Rada Mihalcea
Computer Science and Engineering
University of Michigan
mihalcea@umich.edu

Abstract

In multimodal machine learning, additive late-fusion is a straightforward approach to combine the feature representations from different modalities, in which the final prediction can be formulated as the sum of unimodal predictions. While it has been found that certain late-fusion models can achieve competitive performance with lower computational costs compared to complex multimodal interactive models, how to effectively search for a good late-fusion model is still an open question. Moreover, for different modalities, the best unimodal models may work under significantly different learning rates due to the nature of the modality and the computational flow of the model; thus, selecting a global learning rate for late-fusion models can result in a vanishing gradient for some modalities. To help address these issues, we propose a Modality-Specific Learning Rate (MSLR) method to effectively build late-fusion multimodal models from fine-tuned unimodal models. We investigate three different strategies to assign learning rates to different modalities. Our experiments show that MSLR outperforms global learning rates on multiple tasks and settings, and enables the models to effectively learn each modality.

1 Introduction

Multimodal machine learning aims to jointly understand and process the inputs from different modalities (e.g., language, audio, vision). This usually requires a model to have the ability to incorporate the feature representations from each modality into a joint representation (the "multimodal fusion" problem). There are two types of commonly-used multimodal fusion methods: late-fusion and multimodal interaction. Late-fusion methods rely on the representation vectors computed from unimodal encoders, which are then combined into a joint representation using operations such as addition, multiplication (Kim et al., 2016), bi-linear pooling (Fukui et al., 2016; Yu et al., 2017b), and so on. Multimodal interactive methods apply complex operations such as cross-modal attention (Yu et al., 2017a), modulation (Yao et al., 2018), and multi-head self-attention such as multimodal transformers (Tan and Bansal, 2019; Tsai et al., 2019).

Despite the intuition that multimodal interaction leverages the inter-dependency across different modalities, (Hessel and Lee, 2020) proposed that there is a method to simulate the outputs of an additive late-fusion model that has the closest possible performance to an *arbitrary* interactive model (but not how to find the specific structure). According to the experimental results in (Hessel and Lee, 2020), the accuracy of the closest additive models is competitive with the corresponding interactive models in some selected tasks. This indicates that: (1) Currently, some interactive models are not strong enough to catch the complex real-world inter-dependencies between modalities. Studying the upper-bound of late-fusion methods can help evaluate the limitations of interactive models. (2) The application of late-fusion models is still open to in-depth research because they have the potential of reducing the computational costs while maintaining some effectiveness.

An additive late-fusion method with two modalities M, N and inputs m, n can be fomulated as follows:

$$f(m,n) = f_M(m) + f_N(n). \quad (1)$$

We assume that such a well-performing $f(m,n)$ can be built up with the most effective unimodal structures for f_M and f_N, i.e., a transformer (Vaswani et al., 2017) for the textual modality and convolution neural networks (CNN) (Ren et al., 2015) for the visual modality. While training $f(m,n)$, the most common current practice is to select a global learning rate. However, the optimal unimodal learning rates of f_M and f_N can be significantly different. For example, with an Adam optimizer (Kingma and Ba, 2014), the best learn-

Findings of the Association for Computational Linguistics: ACL 2022, pages 1824 - 1834

ing rate for the transformer is usually around 2e-5, while the best learning rate for Multi-Layer Perceptrons (MLP) can be up to 1e-3. While combining the two structures into a late-fusion model with a global learning rate, i.e., 3e-4, the transformer part turns out to be nearly frozen in the training procedure (see the "Conductance Analysis" subsections in the Experimental Results section).

To address this issue, we propose the Modality-Specific Learning Rate (MSLR) method, which uses different learning rates for different modalities while training an additive late-fusion model. We explore different model structures, tasks, and learning rate assignment strategies to analyse the impact of MSLR on the gradient effectiveness, predicative behaviors, and evaluation results.

Our contributions are as follows. Firstly, we propose MSLR as an effective strategy to train an additive late-fusion model for multimodal tasks; secondly, we analyse the predicative behavior and layer conductance to prove the necessity of using MSLR instead of global learning rates in some conditions; finally, experiments on three different tasks: MuSE Stress Detection (Jaiswal et al., 2019, 2020), MELD Sentiment Analysis (Poria et al., 2019), and MM-IMDb Movie Genre Classification (Ovalle et al., 2017) indicate that MSLR outperforms global learning rates with certain assignment strategies.

2 Related Work

2.1 Multimodal Classification

We focus on multimodal classification tasks which have broad applications in real life. In multimodal classification, the logits of each class predicted by each unimodal sub-part of the joint late-fusion model can be directly summed up and converted into an output distribution. Examples of commoly-studied multimodal classification tasks include sentiment analysis (Zadeh et al., 2016; Yao et al., 2020; Poria et al., 2019), emotion recognition (Busso et al., 2008; Jaiswal et al., 2020; Zadeh et al., 2018), and other real-world applications such as disaster classification (Tian et al., 2018) and movie genre classification (Ovalle et al., 2017).

The inputs of multimodal classification models are usually videos, which contain visual image frames, audio utterances and textual transcripts. These modalities are typically processed by different models based on the nature of each modality. For example, visual features are extracted by pre-trained Convolutional Neural Networks (CNNs) (Simonyan and Zisserman, 2014; Szegedy et al., 2017), spectral or temporal acoustic features are extracted using tools such as OpenSmile (Eyben et al., 2010) and Covarep (Degottex et al., 2014), textual features are usually achieved by pre-trained word embeddings (Peters et al., 2018) and Transformers (Devlin et al., 2019). An effective model should be able to incorporate these features with different numerical properties and natural distributions.

2.2 Multimodal Fusion

There are two mainstream methods to encode and combine multimodal features. The first approach is late-fusion, in which the features from different models are first encoded separately by unimodal encoders, and the single-vector representation is then combined into a joint representation and fed into the final classifier (Kim et al., 2016; Fukui et al., 2016; Yu et al., 2017b). The advantages of late-fusion is that the model is relatively light-weighted and interpretable, and the sub-parts processing each modality can be well-monitored. However, the low-level alignments across the modalities, such as the correspondence between a textual word and a visual object, can not be detected while computing the unimodal feature vectors. On the other hand, the multimodal interaction methods enable the encoders to interact with each other via cross-modal attention mechanisms (Yu et al., 2017a; Tan and Bansal, 2019; Tsai et al., 2019).

Although it is intuitive that the interaction methods can have better capability, Hessel and Lee (2020) showed that the prediction of any interactive model can be simulated by a corresponding late-fusion model, making it possible to reduce the computational costs without severely hurting the performances.

2.3 Modality Specific Learning

2.3.1 Modality-Specific Early Stopping.

A closely related work to ours is called Modality-Specific Early Stopping (MSES) (Fujimori et al., 2019). They stated the issue in multimodal learning as "overfitting in some modalities," and attributed it to "the convergence rate and generalization performance differ among modalities," which is similar to our claims and observations. However, they did not explore the cause of this overfitting, and proposed to solve the problem by applying early stopping for the modalities that have appeared to be

converged regarding the validation performances. Their method does not actually assign different step-sizes for different modalities and still chooses a global learning rate instead. In contrast, we investigate the layer conductance of the model and observe that the overfitting in certain modalities is because the global learning rate is beyond the numerical range where the model structure for that modality can work regularly. While one modality receives a vanishing gradient, the unimodal performance no longer improves and appears to overfit. Thus, we directly modify the initial learning rates according to the knowledge on learning rates achieved from unimodal fine-tuning. Our method is able to delay the overfitting to some extent, instead of simply choosing the best saved parameters for the overfit modalities and stopping training.

2.3.2 Gradient Blending

Another related work is Gradient Blending (Wang et al., 2020), which also states the difficulty of joint training as overfitting. Unlike MSES (Fujimori et al., 2019), they directly modifies the gradient descent process by substituting the total loss with a weighted sum of multiple unimodal loss, and the weight is computed based on a "overfitting-to-generalization ratio" (OGRs) that describes the overfitting conditions for each modality. However, the computation of OGRs relies on training each unimodal model for the first several epochs, while the initial learning rate for each modality is still chosen globally and does not guarantee the training behavior of these initial steps. As a result, if a model does not receive gradient at all when the training starts (which is possible in some of our experiments), the initial OGRs can be ill-formed, limiting the usage of Gradient Blending.

Besides, the tasks and situations they deal with are different from ours: in most of their cases, the joint training underperforms unimodal training, but in our tasks, a joint training with global learning rate can already outperform the unimodal results, and our method can bring further improvement. Also, the performance of Gradient Blending on the textual modality is not explored, while our method works well with both textual-visual and textual-audio data, as shown in our experiments.

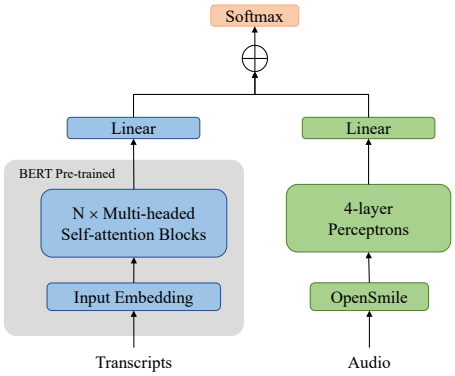

Figure 1: Late-fusion architecture for MuSE stress detection.

3 Modality-Specific Learning Rates

3.1 Learning Rates

The best learning rate for a model depends both on its structure and the optimization algorithm. The models structure further depends significantly on the modality of inputs, i.e., a transformer is effective for the textual modality, CNN for local image parts, and MLP is enough for a single hand-crafted feature vector. As a result, the best range for learning rates can be largely different across modalities.

For different optimizers, the default learning rate range also has large variation from less than 1e-3 (Adam-like) to 1.0 (Adadelta, (Zeiler, 2012)).

We propose to use modality specific learning rates, and include different *learning rate assigment strategies* to keep the models that work for each single modality still work in multimodal training, as described in the following three subsections. To focus on analysing the influence of modality, we use an AdamW optimizer (Loshchilov and Hutter, 2017) for all of our models. In this setting, the term "learning rate" stands for the step size α. Step size is a hyper-parameter independent of the cumulated first moment m_t and second moment v_t in each step of gradient descent. Please refer to (Kingma and Ba, 2014; Loshchilov and Hutter, 2017) for more details. In our strategies, we either choose a fixed α value for each modality or adjust α dynamically based on unimodal performance, which is still independent of the first and second moments.

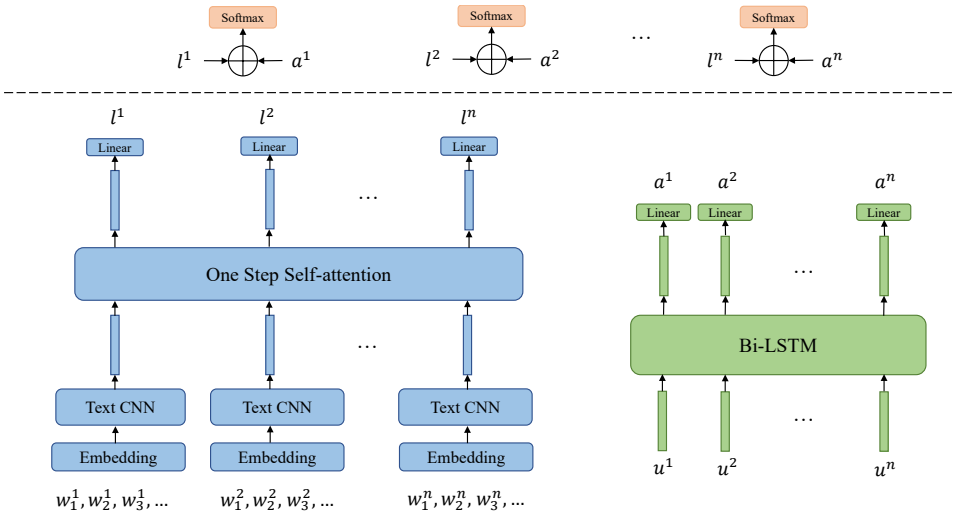

Figure 2: Late-fusion architecture for MELD sentiment analysis.

Table 1: Overlap and Confusion matrix for MSLR-Keep and Joint-global, compared to Audio-only.

Metrics	Overlap	1-1	0-0	1-0	0-1
Audio-only vs. Joint-global	0.86	0.46	0.39	0.09	0.05
Audio-only vs. Keep-ep20	0.81	0.47	0.34	0.09	0.11
Audio-only vs. Keep-ep100	**0.65**	**0.39**	**0.26**	**0.16**	**0.18**
Text-only vs. Joint-global	0.62	0.37	0.25	0.24	0.14
Text-only vs. Keep-ep20	0.70	0.44	0.25	0.17	0.13
Text-only vs. Keep-ep100	**0.73**	**0.46**	**0.27**	**0.15**	**0.11**
Text-only vs. Audio-only	0.62	0.40	0.23	0.22	0.16
Joint-global vs. Keep-ep20	0.84	0.47	0.38	0.11	0.04
Joint-global vs. Keep-ep100	0.62	0.37	0.25	0.24	0.14

3.2 The "Keep" Strategy

The most straight-forward MSLR strategy is keeping the best fine-tuned unimodal learning rate for different modalities while training the late-fusion model. This strategy is expected to ensure that each unimodal sub-part still has effective gradients.

3.3 The "Smooth" Strategy

The "Smooth" strategy compromises different learning rates by shifting the learning rate for different modalities to be closer to the average learning rate of all modalities, resulting in smaller margins. This is supposed to lead to more stable training and yields better results when all the modalities work in relatively close learning rate ranges.

3.4 The "Dynamic" Strategy

Motivated by the dynamic sampling strategies (Guo et al., 2018; Gottumukkala et al., 2020; Yao et al., 2021) in multi-task learning, we leverage the validation set to measure how fast the model is learning each of its unimodal sub-parts. We start from the "Keep" strategy in the first epoch, and update the step-size for modality N after each epoch based on the performance of the unimodal prediction $f_N(n)$ on the validation set. Specifically, for epoch t and modality N, we update the step-size by:

$$\alpha_{t,N} = \alpha_{0,N} * r_{t,N}^{val}, \quad (2)$$

where $r_{t,N}^{val}$ is the ratio of the unimodal performance on the validation set in epoch t to the average performance of the previous 5$\sim$10 epochs, which is usually slightly larger or smaller than 1.0. We name this as the "Dynamic" strategy. The motivation for this strategy is that if the unimodal performance of a modality is significantly improved in an epoch, the learning rate for this modality should be increased to make full use of the current gradient direction; otherwise, if there is no significant difference with respect to previous epochs, we should maintain the current learning rate to keep it in the effective range for this modality.

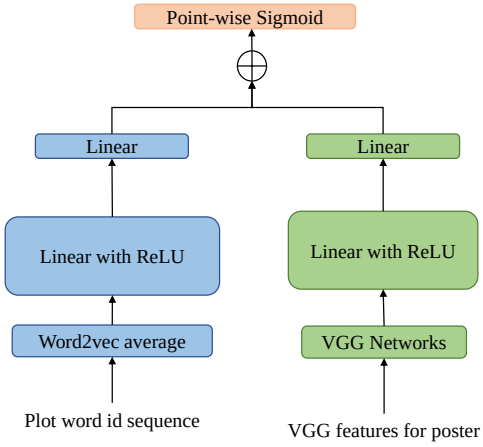

Figure 3: Late-fusion architecture for MM-IMDb Movie Genre Classification.

3.5 Computational Cost

A common concern of our methods might be the computational cost: all the MSLR strategies rely on searching for a best unimodal learning rate for each modality before the multimodal training starts. However, it is worth noticing that every model structure has its best learning rate range, which is sometimes unknown. Thus, it is necessary to do this search for newly-designed models and previously-unseen tasks. In other cases where the unimodal model structure and task is well-studied (i.e., BERT for textual classification), the best unimodal learning rate can also be directly determined based on one's experience.

In the worst case, existing methods train K times if there are K candidate learning rate values, while MSLR trains for additional K times for each modality involved, which grows only linearly with respect to the number of modalities. Besides, the unimodal models trained in these steps are not simply discarded: they can be used to make unimodal predictions while data from the other modalities are missing, which is often the case in real-world applications.

4 Tasks and Models

4.1 MuSE Stress Detection

Multimodal Stressed Emotion (MuSE) (Jaiswal et al., 2019, 2020) is a multimodal dataset for emotion recognition and stress detection, which is collected from student monologue sessions recorded before or after their final exams. The topic and content of each monologue is directed by random emotion-eliciting questions such as "tell me about an unhappy experience in your life." Monologue sentence clips are annotated with binary stress labels: "stressed" for monologues recorded right before final exams, and "non-stressed" for those after exams. For each sample, we make predictions using the audio utterance of a sentence in the monologue session, as well as its textual transcription. We use 1853, 200, and 273 samples for training, validation, and testing, respectively.

For the model structure, shown in Figure 1, we use a Transformer pre-trained with BERT (Devlin et al., 2019) as our textual encoder for the transcripts. For the audio inputs, we extract an 88-dimensional acoustic feature using OpenSmile (Eyben et al., 2010) with eGeMaps (Eyben et al., 2015) configuration for each sentence, and pass it through a 4-layer 256-dimensional MLP. The top-level 256-dimensional representations from both modalities are concatenated and projected into the output logits by a linear layer, which is equivalent to an additive late-fusion.

4.2 MELD Sentiment Analysis

The Multimodal Emotion Lines Dataset (MELD) (Poria et al., 2019) is an expansion of the Emotion Lines multi-party conversation dataset (Chen et al., 2018) and contains the audios and transcrips for the dialogues from the TV-series *Friends*, in which each sentence is annotated with emotion and sentiment labels. For the multimodal sentiment analysis task, there are three classes: positive, negative, and neutral, and two modalities: audio and textual. We use 1038, 114, and 280 dialogues for training, validation, and test, respectively.

For preprocessing, we follow (Poria et al., 2019) to apply feature selection on the 6373 dimensional acoustic features from OpenSmile, resulting in a 1422 dimensional dense audio representation for each sentence. We consider the dialogue as a sequence of sentences, regardless of the specific speaker. The maximum dialogue length is 33.

Our sentiment analysis model (Figure 2) contains a textual encoder and an audio encoder. The textual encoder has a word-level 2d Convolutional Neural Network (Zhang and Wallace, 2017) that outputs a 512-dimensional sentence embedding from the word embeddings. For the sentence embedding, we apply one step of masked self-

Table 2: Evaluation metrics for MuSE stress detection. "lr" stands for learning rate.

Model	Textual lr	Audio lr	Accuracy	Precision	Recall	F-score
Text-only	2e-5	-	0.69	0.77	0.74	0.75
Audio-only	-	5e-3	0.82	0.83	0.82	0.83
Joint-global	3e-4	3e-4	0.82	0.83	0.83	0.83
MSES (Fujimori et al., 2019)	3e-4	3e-4	0.80	0.79	**0.85**	0.82
MSLR: Keep	2e-5	5e-3	0.83	0.85	0.81	0.83
MSLR: Smooth	1e-4	1e-3	0.81	0.84	0.81	0.82
MSLR: Dynamic	-	-	**0.84**	**0.86**	0.83	**0.84**

Table 3: Evaluation metrics for MELD Sentiment Analysis.

F-score (%)	Textual lr	Audio lr	Neutral	Positive	Negative	Average
Text-only	1e-4	-	76.32	56.03	59.71	66.97
Audio-only	-	1e-3	64.40	12.94	42.38	47.10
Joint-global	5e-4	5e-4	76.58	53.97	57.32	65.92
MSES(Fujimori et al., 2019)	5e-4	5e-4	76.41	53.41	57.79	65.87
MSLR: Keep	1e-4	1e-3	75.61	55.40	59.31	66.37
MSLR: Smooth	2.5e-4	7.5e-4	76.44	**56.34**	**60.10**	**67.21**
MSLR: Dynamic	-	-	**77.14**	52.73	56.41	65.65

attention (Vaswani et al., 2017) on the sentence sequence in the same dialogue, resulting in a sequence of 512-dimensional textual hidden states. For the audio encoder, we use a bi-directional LSTM (Hochreiter and Schmidhuber, 1997) which takes the audio features for each utterance as input, and outputs 300 dimensional hidden states. For each time step (sentence), the output of self-attention layer and audio LSTM are concatenated and projected by a 512-dimensional linear layer to predict its sentiment class (additive late-fusion).

4.3 MM-IMDb Movie Genre Classification

The Multimodal IMDb (MM-IMDb) (Ovalle et al., 2017) dataset is built with 25,959 IMDb movies with their plots and posters; each movie is labeled with more than one genre, making it a multi-label classification task. There are two modalities: plot (textual) and poster (visual). We use a training/validation/test split of 15552/2608/7799 movies, respectively.

As for preprocessing, following related work (Ovalle et al., 2017) and (Fujimori et al., 2019), we use the VGG Neural Network (Simonyan and Zisserman, 2014) pre-trained on ImageNet (Deng et al., 2009) which produces 4096-dimensional visual features for the posters, and 300-dimensional Word2Vec [1] embeddings for the textual plots.

We implement the same model structure as described by (Fujimori et al., 2019), which is a linear layer with 2048 hidden states and ReLU activation, followed by a 512-dimensional linear layer as the classifier, for both modalities (Figure 3). There are 23 output neurons corresponding to the 23 genre classes. Each neuron has a sigmoid activation instead of softmax for multi-label classification. The motivation of using a Multi-layer Perceptrons (MLP) structure on both modality is to test the efficiency of our MSLR strategies while different modalities have similar computational flows, as well as to have a comparison with the related MSES method (Fujimori et al., 2019).

5 Experimental Results

5.1 General Settings

For all our experiments with the "Dynamic" strategy, we compute the ratio r with respect to the previous 5 epochs. All the MSES methods used for comparison are based on our implementation. The best unimodal and global learning rates for each task, as well as all the other hyperparameters, are found by a linear search based on the metrics on the validation sets. All our experiments are implemented with Pytorch[2] and ran on 1 GeForce RTX

[1] https://code.google.com/archive/p/word2vec/

[2] https://pytorch.org/

Table 4: Evaluation metrics for MM-IMDb Movie Genre Classification.

F-score	Textual lr	Audio lr	Micro	Macro	Weighted	Sample
Text-only	1e-2	-	0.582	0.470	0.562	0.577
Visual-only	-	1e-4	0.419	0.243	0.377	0.409
Joint-global	1e-3	1e-3	0.588	0.441	0.562	0.578
MSES(Fujimori et al., 2019)	5e-4	5e-4	0.579	0.486	0.567	0.571
MSLR: Keep	1e-2	1e-4	0.587	0.443	0.557	**0.582**
MSLR: Smooth	3e-3	3e-4	0.579	0.448	0.566	0.570
MSLR: Dynamic	-	-	**0.592**	**0.518**	**0.587**	0.581

2080 super GPU and Intel i7 9700k processor.

5.2 MuSE Stress Detection

For the MuSE stress detection task and late-fusion structure with a Transformer + MLP structure, we use a batch size of 32. A learning rate of 2e-5 works the best for the textual modality, while 5e-3 works best for the audio modality. The late-fusion model works the best with a global learning rate of 3e-4. We name these models "Text-only", "Audio-only", and "Joint-global", respectively.

5.2.1 Conductance Analysis

Layer Conductance (Sundararajan et al., 2017; Shrikumar et al., 2018) evaluates the importance of each neuron to the final prediction. It is worth noticing that the conductance value itself is not directly related to the training gradients with respect to this specific neuron. However, we compute the average Layer Conductance of all the neurons in the textual/visual/audio representations, and further averaged over all the samples in the dataset. The result stands for the importance of each single modality as a whole. If the Layer Conductance of a modality is close to 0, it is reasonable to claim that this modality is not effectively trained at all and has vanishing gradients in the training procedure.

We analyse the Layer Conductance for the outputs of the textual and acoustic encoder, separately, using the Captum (Kokhlikyan et al., 2020) package. The layer conductance result for MuSE Stress Detection is averaged among all the 256 neurons of the linear layer for each modality and shown in Table 5.

We observe in Table 5 that with a joint-global learning rate (3e-4), the textual Transformer works beyond its comfort zone (around 2e-5) and has vanished gradients (conductance close to 0). This indicates that the model's multimodal performance is limited because it can not effectively learn the textual modality while using a global learning rate. In contrast, we observe that using the MSLR "Keep" strategy solves this issue.

Table 5: Layer conductance for different models on the textual and audio modality for the MuSE Stress Detection task.

Modality	Textual	Audio
Text-only	0.002	-
Audio-only	-	0.25
Joint-global	1e-8	0.01
MSLR: Keep - epoch 20	0.005	0.014
MSLR: Keep - epoch 100	0.007	0.015

5.2.2 Prediction Similarity

Another approach of exploring how different are the learned models with MSLR and global learning rates is to directly analyse the predictions on the test set. If the language encoder has vanished gradients, the multimodal predicative behavior should be close to the unimodal audio model. In Table 1, we show the overlap rate (the ratio of the two models making the same prediction for a sample) for different model pairs, as well as the full confusion matrix for the stressed (1) and non-stressed (0) labels. We choose the joint model at the 20-th epoch (Keep-ep20, when the training is on-going) and the 100-th epoch (Keep-ep100, when the training is converged) for comparison with the Audio-only and Text-only models. We highlight the joint model that is less similar to the audio model and more similar to the textual model, since going closer to the textual model indicates a valid gradient for the textual modality.

We observe that without MSLR, the joint-global model has 0.86 overlap with the Audio-only model and only 0.62 with the Text-only model. However, if MSLR is applied, as the training goes on (from epoch 20 to 100), MSLR gets away from the Audio-

only model and becomes closer to the Text-only model, which is consistent with Table 5 showing that the textual part is receiving gradients. Besides, after 100 epochs, MSLR results in a very different model from all the joint and unimodal models.

5.2.3 Evaluation Metrics

The evaluation metrics we use for the MuSE Stress Detection task include the total accuracy and the precision, recall and F-score for the "stressed" label (Table 2). We observe that the "Keep" strategy achieves competitive scores with the best global learning rate model while the model's predicative behavior is very different as shown by the previous subsection. Additionally, the "Dynamic" strategy significantly outperforms both the global learning rate and the Multimodal Early Stopping (MSES) method ($p < 0.05$, t-test). We believe that starting from "Keep" enables the model to learn both modalities with valid gradients, and the "Dynamic" strategy helps adjust the learning rate according to the validation performance of the unimodal models, which brings further improvements.

5.3 MELD Sentiment Analysis

For the MELD Sentiment Analysis dataset, we use a batch size of 10; the best learning rate for Text-only, Audio-only and Joint-global is 1e-4, 1e-3 and 5e-4, respectively. For the "Smooth" strategy, we use a learning rate of 2.5e-4 for textual modality and 7.5e-4 for audio.

5.3.1 Conductance Analysis

We apply Layer Conductance analysis on the 512 neurons of the top linear layer for each modality, as we did in the MuSE Stress Detection. The results are in Table 6. In this case, since the gap between the suitable learning rate for the two modalities is smaller than the MuSE task, we observe non-zero layer conductance for both modalities for the global learning rate method. The MSLR method, on the other hand, still achieves higher value of conductance as the training goes on.

5.3.2 Evaluation Metrics

Following (Poria et al., 2019), the MELD Sentiment Analysis task is evaluated with the F-scores for each class and their weighted average (Table 3).

We observe that the "Smooth" strategy works slightly better than the "Keep" strategy in this case. This is potentially because the smaller learning rate gap makes 5e-4 an acceptable learning rate for both modalities with valid gradient flows. The "Keep" strategy maintains the large gap, which makes the training less stable compared to the "Smooth" strategy which can be considered as a reconcile with the global learning rate. The "Smooth" strategy also outperforms the "Dynamic" strategy since the latter starts from the same initial learning rates with a large gap as in the "Keep" strategy.

Table 6: Layer conductance for different models on the textual and audio modality for the MELD Sentiment Analysis task.

Modality	Textual	Audio
Text-only	0.011	-
Audio-only	-	0.024
Joint-global	0.011	0.006
MSLR: Keep - epoch 20	0.034	0.027
MSLR: Keep - epoch 100	0.041	0.033

5.4 MM-IMDb Movie Genre Classification

For the MM-IMDB dataset, we use a batch size of 128. We name the unimodal model using only the plot the "Text-only" model, and the model using only the poster the "Visual-only" model. The best fine-tuned learning rates for Text-only, Visual-only and Joint-global models are 1e-2, 1e-4, and 1e-3, respectively. It is worth noticing that although we have similar MLP structures for both modalities, the best learning rates can still have a 100-time gap between the two modalities. This is perhaps because of the numerical properties of the features from different modalities, as well as the pre-processing methods. For the "Smooth" strategy, we use a learning rate of 3e-3 for the textual modality and 3e-4 for the visual modality.

5.4.1 Conductance Analysis

We apply the same Layer Conductance analysis as the other two datasets on the 512 hidden units of the top-level linear layer for each modality. The results are in Table 7.

We observe that the textual representation has relatively low average conductance compared to the visual one when the model converges with a global learning rate. The MSLR strategy helps alleviate this issue and makes the training more efficient.

Based on the gradient analysis on all the three tasks, we conclude that choosing an initial learning rate according to unimodal results is a simple and effective approach to help with the vanishing gradient problem in certain cases.

Table 7: Layer conductance for different models on the textual and audio modality for the IMDb Movie Genre classification task.

Modality	Textual	Audio
Text-only	0.010	-
Visual-only	-	0.007
Joint-global	0.002	0.019
MSLR: Keep - epoch 20	0.006	0.007
MSLR: Keep - epoch 100	0.011	0.031

5.4.2 Evaluation Metrics

Following (Ovalle et al., 2017), the performance of genre classification is evaluated by F-scores computed by four different averaging algorithms: micro, macro, weighted, and samples. The results are shown in Table 4. We reach the same conclusion as in the MuSE Stress Detection task: when the best learning rates are extremely different, the "Keep" and "Dynamic" strategies work better than "Smooth" and all the other baselines.

6 Lessons Learned

In this work, we proposed modality-specific learning rates (MSLR) for training multimodal late-fusion models built up with unimodal encoders. To summarize, we have the following findings:

Firstly, we showed that learning multimodal late-fusion models can be difficult if the best learning rate for each modality is significantly different. A global learning rate may not work for all the modalities according to our Layer Conductance analysis for the representations from different modalities.

Secondly, we tried solving this problem using MSLR. According to both the conductance analysis and the predicative performance with the "Keep" Strategy, we conclude that it helps prevent the vanishing gradient, and when the training converges, it results in a model that is different compard to the global learning rates.

Thirdly, we evaluated three different MSLR strategies on three different multimodal tasks with various model structures. We observed that MSLR generally achieves competitive or better scores on most of the commonly-used evaluation metrics as compared to baselines using a global learning rate or related modality-specific learning methods.

Specifically, the experimental results on the MELD Sentiment Analysis task indicated that when different modalities have close ranges of best learning rates, the model with a global learning rate is a strong baseline, while MSLR achieves competitive performance with the "Smooth" strategy performing the best. Otherwise, in the MuSE and MM-IMDb tasks where the learning rate gaps are large, the "Keep" and "Dynamic" strategies outperform the global learning rate model because they ensure a valid gradient on all the modalities.

A potential disadvantage of MSLR is the unstable training process, which can be the topic of future work. We also hope that our work inspires more research on new learning strategies for multimodal interactive models and generative tasks.

Acknowledgments

This research was partially supported by a grant from the Automotive Research Center (ARC) at the University of Michigan.

References

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335.

Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Lun-Wei Ku, et al. 2018. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379*.

G. Degottex, John Kane, Thomas Drugman, T. Raitio, and Stefan Scherer. 2014. Covarep — a collaborative voice analysis repository for speech technologies. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 960–964.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The geneva minimalistic acoustic parameter set (gemaps) for voice

research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.

Naotsuna Fujimori, Rei Endo, Yoshihiko Kawai, and Takahiro Mochizuki. 2019. Modality-specific learning rate control for multimodal classification. In *Asian Conference on Pattern Recognition*, pages 412–422. Springer.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.

Ananth Gottumukkala, Dheeru Dua, Sameer Singh, and Matt Gardner. 2020. Dynamic sampling strategies for multi-task reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 920–924.

Michelle Guo, Albert Haque, De-An Huang, Serena Yeung, and Li Fei-Fei. 2018. Dynamic task prioritization for multitask learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 270–287.

Jack Hessel and Lillian Lee. 2020. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 861–877.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Mimansa Jaiswal, Zakaria Aldeneh, Cristian-Paul Bara, Yuanhang Luo, Mihai Burzo, Rada Mihalcea, and Emily Mower Provost. 2019. Muse-ing on the impact of utterance ordering on crowdsourced emotion annotations. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7415–7419. IEEE.

Mimansa Jaiswal, Cristian-Paul Bara, Yuanhang Luo, Mihai Burzo, Rada Mihalcea, and Emily Mower Provost. 2020. Muse: a multimodal dataset of stressed emotion. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1499–1510.

Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2016. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. 2020. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

John Edison Arevalo Ovalle, Thamar Solorio, Manuel Montes-y-Gómez, and Fabio A. González. 2017. Gated multimodal units for information fusion. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pages 91–99.

Avanti Shrikumar, Jocelin Su, and Anshul Kundaje. 2018. Computationally efficient measures of internal neuron importance. *arXiv preprint arXiv:1807.09946*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR.

Christian Szegedy, S. Ioffe, V. Vanhoucke, and Alexander Amir Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5103–5114.

Haiman Tian, Yudong Tao, Samira Pouyanfar, Shu-Ching Chen, and M. Shyu. 2018. Multimodal deep representation learning for video classification. *World Wide Web*, 22:1325–1341.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

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

Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What makes training multi-modal classification networks hard? *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12692–12702.

Yiqun Yao, Michalis Papakostas, Mihai Burzo, Mohamed Abouelenien, and Rada Mihalcea. 2021. Muser: Multimodal stress detection using emotion recognition as an auxiliary task. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2714–2725.

Yiqun Yao, Verónica Pérez-Rosas, Mohamed Abouelenien, and Mihai Burzo. 2020. Morse: Multimodal sentiment analysis for real-life settings. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 387–396.

Yiqun Yao, Jiaming Xu, Feng Wang, and Bo Xu. 2018. Cascaded mutual modulation for visual reasoning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 975–980.

Dongfei Yu, Jianlong Fu, Tao Mei, and Yong Rui. 2017a. Multi-level attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4709–4717.

Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017b. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 1821–1830.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

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.

Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Ye Zhang and Byron C Wallace. 2017. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 253–263.