题目：Balanced Multimodal Learning via On-the-fly Gradient Modulation

作者：Xiaokang Peng1,†, Yake Wei1,†, Andong Deng2, Dong Wang3, Di Hu1,*

机构：Gaoling School of Artificial Intelligence, Renmin University of China

实验室主页：GeWu-Lab

动机

- 为所有模态设定一个统一的优化目标会让单模态表征**under-optimized**，这可能是主导模态（dominated modality）导致的，如刮风事件里的声音

工作

- 为了缓解第一个问题，作者提出了**on-the-fly gradient modulation**，通过监控每个模态对优化目标的贡献差异，自适应的控制每种模态的优化，对于under-optimized的模态提供更多帮助
- 作者额外引入了一个动态改变的**Gaussian noise**来避免因gradient modulation导致的generalization drop，因为gradient modulation会导致stochastic gradient noise降低，论文[1,2]中证明了stochastic gradient noise的强度与generalization ability成positive correlation。

Imbalance analysis: dominated modality

$$\frac{\partial L}{\partial f(x_i)_c} = \frac{e^{(W^a \cdot \varphi_i^a + W^v \cdot \varphi_i^v + b)_c}}{\sum_{k=1}^{M} e^{(W^a \cdot \varphi_i^a + W^v \cdot \varphi_i^v + b)_k}} - 1_{c=y_i}$$

On-the-fly Gradient Modulation(OGM)

[redacted]

首先对梯度下降GD进行定义：

$$\theta_{t+1}^{u}=\theta_{t}^{u}-\eta\nabla_{\theta^{u}}L\left(\theta_{t}^{u}\right) \qquad ()$$

实际中，我们使用SGD：

$$\theta_{t+1}^{u}=\theta_{t}^{u}-\eta\tilde{g}\left(\theta_{t}^{u}\right) \qquad ()$$

其中$\tilde{g}\left(\theta_{t}^{u}\right)=\frac{1}{m}\sum_{x\in B_{t}}\nabla_{\theta^{u}}\ell\left(x;\theta_{t}^{u}\right)$，其是$\nabla_{\theta^{u}}L\left(\theta_{t}^{u}\right)$的unbiased estimation，有：

$$\tilde{g}\left(\theta_{t}^{u}\right)\sim\mathcal{N}\left(\nabla_{\theta^{u}}L\left(\theta_{t}^{u}\right),\Sigma^{sgd}\left(\theta_{t}^{u}\right)\right)$$

$$\Sigma^{sgd}\left(\theta_{t}^{u}\right)\approx\frac{1}{m}\left[\frac{1}{N}\sum_{i=1}^{N}\nabla_{\theta^{u}}\ell\left(x_{i};\theta_{t}^{u}\right)\nabla_{\theta^{u}}\ell(x_{i};\theta_{t}^{u})^{\mathrm{T}}\right.$$
$$\left.-\nabla_{\theta^{u}}L\left(\theta_{t}^{u}\right)\nabla_{\theta^{u}}L(\theta_{t}^{u})^{\mathrm{T}}\right].$$

接下来为了自适应的控制每种模态的优化，我们要得到每个模态梯度的加权系数，从而调整梯度大小，同步各个模态，

首先得到discrepancy ratio ρ_{t}^{v}：

$$s_{i}^{a}=\sum_{k=1}^{M}1_{k=y_{i}}\cdot\operatorname{softmax}\left(W_{t}^{a}\cdot\varphi_{t}^{a}\left(\theta^{a},x_{i}^{a}\right)+\frac{b}{2}\right)_{k},$$
$$s_{i}^{v}=\sum_{k=1}^{M}1_{k=y_{i}}\cdot\operatorname{softmax}\left(W_{t}^{v}\cdot\varphi_{t}^{v}\left(\theta^{v},x_{i}^{v}\right)+\frac{b}{2}\right)_{k},$$
$$\rho_{t}^{v}=\frac{\sum_{i\in B_{t}}s_{i}^{v}}{\sum_{i\in B_{t}}s_{i}^{a}}.$$

然后计算系数k_{t}^{u}：

$$k_{t}^{u}=\begin{cases}1-\tanh\left(\alpha\cdot\rho_{t}^{u}\right) & \rho_{t}^{u}>1\\ 1 & \text{others}\end{cases}$$

带入SGD更新公式()：

$$\theta_{t+1}^{u}=\theta_{t}^{u}-\eta\cdot k_{t}^{u}\tilde{g}\left(\theta_{t}^{u}\right) \qquad ()$$

Generalization Enhancement(GE)

- 噪声确实减少

  考虑噪声的SGD过程：

$$\theta_{t+1}^{u}=\theta_{t}^{u}-\eta \nabla_{\theta^{u}} L\left(\theta_{t}^{u}\right)+\eta \xi_{t}, \xi_{t} \sim \mathcal{N}\left(0, \Sigma^{sgd}\left(\theta_{t}^{u}\right)\right)$$

  添加系数后：

$$\begin{gathered}\theta_{t+1}^{u}=\theta_{t}^{u}-\eta \nabla_{\theta^{u}} L^{\prime}\left(\theta_{t}^{u}\right)+\eta \xi_{t}^{\prime} \\ \xi_{t}^{\prime} \sim \mathcal{N}\left(0,\left(k_{t}^{u}\right)^{2} \cdot \Sigma^{sgd}\left(\theta_{t}^{u}\right)\right)\end{gathered}$$

  其中$\eta \nabla_{\theta^{u}} L^{\prime}\left(\theta_{t}^{u}\right)=k_{t}^{u} \cdot \eta \nabla_{\theta^{u}} L\left(\theta_{t}^{u}\right)$，而$k_{t}^{u} \in(0,1]$，所以$\xi_{t}^{\prime}<\xi_{t}$，所以需要增强

- 增加高斯噪声$h\left(\theta_{t}^{u}\right) \sim \mathcal{N}\left(0, \Sigma^{sgd}\left(\theta_{t}^{u}\right)\right)$

$$\begin{aligned}\theta_{t+1}^{u} & =\theta_{t}^{u}-\eta\left(k_{t}^{u} \tilde{g}\left(\theta_{t}^{u}\right)+h\left(\theta_{t}^{u}\right)\right) \\ & =\theta_{t}^{u}-\eta \nabla_{\theta^{u}} L^{\prime}\left(\theta_{t}^{u}\right)+\eta \xi_{t}^{\prime}+\eta \epsilon_{t} \\ & =\theta_{t}^{u}-\eta \nabla_{\theta^{u}} L^{\prime}\left(\theta_{t}^{u}\right)+\eta \xi_{t}^{\prime \prime}\end{aligned} \tag{0}$$

  其中$\epsilon_{t} \sim \mathcal{N}\left(0, \Sigma^{sgd}\left(\theta_{t}^{u}\right)\right)$，$\xi_{t}^{\prime \prime} \sim \mathcal{N}\left(0,\left(k_{t}^{u}\right)^{2}+1) \Sigma^{sgd}\left(\theta_{t}^{u}\right)\right)$，SGD噪声得以加强

算法过程

```
\begin{algorithm}
\caption{Multimodal learning with OGM-GE strategy}
\label{alg:1}
\begin{algorithmic}
\Require Training dataset $\mathcal{D}=\{(x^a_{i},x^v_{i}), y_{i}\}_{i=1,2...N}$, iteration number $T$, hyper-parameter $\alpha$, initialized modal-specific parameters $\theta^{u}$, $u \in \{a,v\}$.
\For{$t=0,\cdots,T-1$}
    \State Sample a fresh mini-batch $B_{t}$ from $\mathcal{D}$;
    \State Feed-forward the batched data $B_{t}$ to the model;
    \State Calculate $\rho^{u}$ using Equation~\ref{8} and~\ref{calcu_ratio};
    \State Calculate $k_{t}^{u}$ using Equation~\ref{k};
    \State Calculate gradient $\tilde{g}(\theta_{t}^{u})$ using back-propagation;
    \State Sample $h(\theta_{t}^{u})$ based on covariance of gradient $\tilde{g}(\theta_{t}^{u})$;
    \State Update using $\theta^{u}_{t+1} =\theta_{t}^{u} - \eta(k_{t}^{u}\tilde{g}(\theta_{t}^{u})+ h(\theta_{t}^{u}))$.
\EndFor
\end{algorithmic}
\end{algorithm}
```

实验

- ## 与传统方法对比

  作者提出的OGM-GE确实能够提高传统方法的性能

- ## 与其他modulation策略对比

Dataset	CREMA-D	KS
Method	Acc	Acc
Concatenation	51.7	59.8
Modality-Drop [9] (audio)	54.4	60.3
Modality-Drop [9] (visual)	53.3	61.3
Grad-Blending [39]	56.8	62.2
OGM	59.0	61.1
OGM-GE	**61.9**	**62.3**

- ## 将OGM-GE插入现有方法

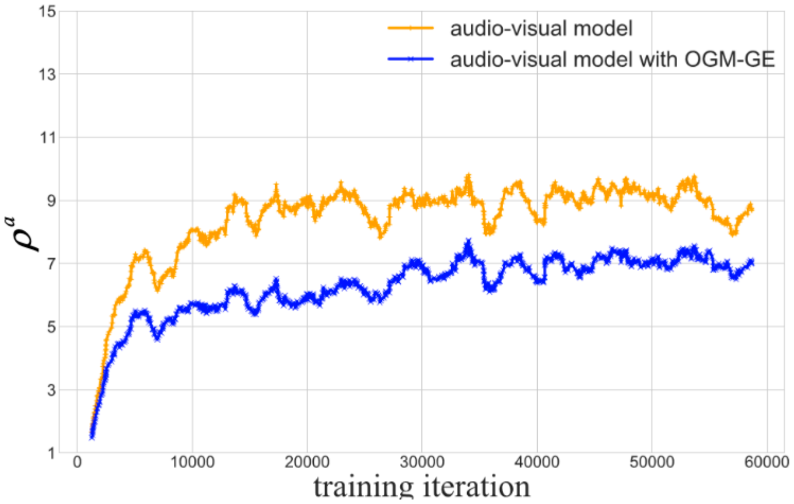
- ## 分类任务以外的应用

  作者在audio-visual event localization任务上进行尝试：

Audio-visual Event Localization		
w/ or w/o OGM-GE	w/o	w/
AVGA [36]	72.0	**72.8**
PSP [46]	76.2	**76.9**

- ## 消融

  - more balanced



- optimizer

Dataset	CREMA-D	KS	VGGSound
Method	Acc	Acc	Acc
SGD	51.7	59.8	49.1
SGD†	**61.9**	**63.1**	**50.6**
Adam	49.7	57.4	47.3
Adam†	**54.6**	**58.9**	**48.2**

- different noise intensities

Settings	CREMA-D	VGGSound
(b=64, lr=1e-4)	50.4	48.3
(b=64, lr=5e-4)	51.0	48.7
(b=64, lr=1e-3)	51.8	49.1
(b= 64, lr=1e-3)	51.8	49.1
(b=128, lr=1e-3)	50.2	48.8
(b=256, lr=1e-3)	48.6	47.7
(b= 64, lr=1e-3) w/ GE	60.2	50.3

问题

1. CREMA-D数据集里val集合的划分比例在论文是0.1，但所提供的代码仓库里只划分了train/test
2. CREMA-D： 0.716（bs=32训练）>0.641（Repo提供的ckpt）>0.619（论文），论文中解释了这个现象，更小的bs会导致更大的高斯噪声，但这个差距实在是太大了

限制

- the uni-modal performance in multimodal model still do not surpass the best uni-modal model.
- solely leveraging optimization-oriented method could not thoroughly solve the imbalance problem, more advanced fusion strategy or network architectures are needed
- 作者只尝试了两个模态，depth, optical flow, language也会有用

[1] Jastrzębski, Stanisław, et al. "Three factors influencing minima in sgd." arXiv preprint arXiv:1711.04623 (2017).

[2] Zhou, Mo, et al. "Toward understanding the importance of noise in training neural networks." International Conference on Machine Learning. PMLR, 2019.