Improving Multimodal Learning via Imbalanced Learning

Dataset	CREMA-D	AVE	KS	MOSI
Method	Acc	Acc	Acc	Acc
Concatenation	58.83	66.15	64.97	76.92
Summation	62.32	67.42	64.15	76.83
Film	56.92	59.91	57.83	76.99
Gated	57.79	65.82	63.72	77.19
Concatenation†	75.26	71.90	74.39	79.94
Summation†	**75.06**	**72.62**	74.49	**80.14**
Film†	65.49	68.81	72.26	79.20
Gated†	75.14	70.22	**75.31**	79.27

Table 1. Performance on CREMA-D, AVE, KS, and MOSI datasets with various fusion methods. † indicates ARL is applied.

Comparison on conventional fusion methods. In this experiment, we apply the ARL strategy to four vanilla fusion methods: Concatenation Summation, Film, and Gated. Among these, Summation is the type of late fusion method that fuses information at the logit level. The other three are the intermediate fusion methods that fuse information at the representation level.

As shown in Table 1, the accuracy of the visual-only model on the CREMA-D dataset is better than all of the vanilla fusion methods, which indicates that the learning of the multimodal model is insufficient. After combining with ARL, the performance of all the vanilla fusion methods consistently gains considerable improvement on all datasets, verifying the effectiveness and flexibility of our method. In particular, for Gated fusion, ARL achieves +18.14%, +9.8%, +11.59%, and +2.08% accuracy improvement on the CRAME-D, AVE, KS, and CEFA datasets, respectively, showing its superiority.

Adaptation to other optimizers. To explore the effectiveness of ARL in combination with other optimizers beyond SGD, we apply it to widely used optimizers AdaGrad and Adam. As shown in Table 2, we empirically show that ARL can work with different optimizers well and gain much performance improvement. The results show that our method can be well adapted to different optimizers, achieving consistent performance improvement.

Dataset	CREMA-D	AVE	KS
Optimizer	Acc	Acc	Acc
SGD	58.83	66.15	64.97
SGD†	**75.26**	**71.90**	**74.39**
Adam	62.15	65.16	60.28
Adam†	**69.28**	**72.71**	**73.51**
AdaGrad	58.62	66.54	57.88
AdaGrad†	**67.51**	**70.64**	**68.81**

Table 2. Ablation experiments of optimizers on different datasets. † indicates ARL is applied.

The ablation explanation of temperature coefficient T. The experiment results on the CRAME-D dataset are shown in Table 3. We can see that as T increases, the performance first increases and then decreases, and the optimal performance is achieved when $T = 8$. This is because too large T may change the modality contribution ratio from less than the variance ratio to greater than the variance ratio, resulting in sub-optimal results.

T	1	4	8	12
Acc	73.25	74.12	**75.26**	74.65

Table 3. The ablation explanation of temperature coefficient T

The ablation explanation of γ. The experiment results on the CRAME-D dataset are shown in Table 4. We can see that as γ increases, the performance first increases and then decreases, and the optimal performance is achieved when $\gamma = 4$. This is because two large γ may suppress the optimization of multimodal loss and result in sub-optimal results.

γ	1	2	4	8
Acc	73.35	75.23	76.68	75.49

Table 4. The ablation explanation of hype-parameter γ

0.1. Derivation for $Bias(f(x), y)^2$

According to $w_0 + w_1 = 1$, we can rewrite $Bias(f(x), y)^2$ as follows,

$$\begin{aligned}
Bias(f(x), y)^2 &= E\left[w_0(s^{m_0} - y) + w_1(s^{m_1} - y)\right]^2 \\
&= w_0^2 E(s^{m_0} - y)^2 + w_1^2 E(s^{m_1} - y)^2 + 2w_0 w_1 E(s^{m_0} - y)E(s^{m_1} - y) \\
&= w_0^2 Bias(s^{m_0}, y)^2 + w_1^2 Bias(s^{m_1}, y)^2 + +2w_0 w_1 Bias(s^{m_0}, y) Bias(s^{m_1}, y) \\
&= (w_0 Bias(s^{m_0}, y) + w_1 Bias(s^{m_1}, y))^2
\end{aligned} \tag{1}$$

Since $w_0 > 0$,$w_1 > 0$,$Bias(s^{m_0}, y) > 0$, and $Bias(s^{m_1}, y) > 0$, minimizing the $Bias(f(x), y)^2$ is equivalent minimizing the following target,

$$w_0 Bias(s^{m_0}, y) + w_1 Bias(s^{m_1}, y) \tag{2}$$

Since $w_0 + w_1 = 1$, we can use the Lagrange multiplier method to obtain the solution as follows,

$$\begin{cases}
w_0 = \dfrac{Bias(s^{m_1}, y)}{Bias(s^{m_1}, y) - Bias(s^{m_0}, y)} & (3) \\
w_1 = \dfrac{-Bias(s^{m_0}, y)}{Bias(s^{m_1}, y) - Bias(s^{m_0}, y)} & (4)
\end{cases}$$

0.2. Derivation for $Var(f(x))$

According to $w_0 + w_1 = 1$, we can rewrite $Var(f(x))$ as follows,

$$\begin{aligned}
Var(f(x)) &= \left[(w_0 s^{m_0} + w_1 s^{m_1})^2\right] - E\left[w_0 f(x) + w_1 f(x)\right] \\
&= (w_0^2 E\left[(s^{m_0})^2\right] + w_1^2 E\left[(s^{m_1})^2\right] + 2w_0 w_1 E\left[(s^{m_0})\right] E\left[(s^{m_1})\right]) \\
&- (w_0^2 E[s^{m_0}]^2 + w_1^2 E[s^{m_1}]^2 + 2w_0 w_1 E\left[(s^{m_0})\right] E\left[(s^{m_1})\right]) \\
&= (w_0^2 E\left[(s^{m_0})^2\right] + w_1^2 E\left[(s^{m_1})^2\right] - w_0^2 E[s^{m_0}]^2 + w_1^2 E[s^{m_1}] \\
&= w_0^2 (E\left[(s^{m_0})^2\right] - E[s^{m_0}]^2) + w_1^2 ({}^2 E\left[(s^{m_1})^2\right] - E[s^{m_1}]^2) \\
&= w_0^2 Var(s^{m_0}) + w_1^2 Var(s^{m_1})
\end{aligned} \tag{5}$$

Since $w_0 + w_1 = 1$, we can use the Lagrange multiplier method to obtain the solution as follows,

$$\begin{cases}
w_0 = \dfrac{Var(s^{m_1})}{Var(s^{m_1}) + Var(s^{m_1})} & (6) \\
w_1 = \dfrac{Var(s^{m_0})}{Var(s^{m_1}) + Var(s^{m_1})} & (7)
\end{cases}$$

References