题目：What Makes Training Multi-modal Classification Networks Hard?

作者：Weiyao Wang, Du Tran, Matt Feiszli

机构：Facebook AI

动机

- 相对于单模态模型，多模态模型接受更多的信息输入，应该比单模态部分强，但作者在实验中发现最好的单模态模型比多模态模型强

工作

- 作者将造成上述现象的原因分为两点：
  - 多模态模型参数量更多，更容易过拟合
  - 不同模态过拟合和泛化的速率不同，单个优化策略去训练是sub-optimal
- 作者提出**overfitting-to-generalization ratio (OGR)**去量化每个模态的过拟合现象，提出**Gradient-Blending (G-Blend)**，通过加权每个模态监督信号和joint head的监督信号，解决第二个问题，该方法在human action recognition，ego-centric action recognition， acoustic event detection上取得state-of-the-art的结果

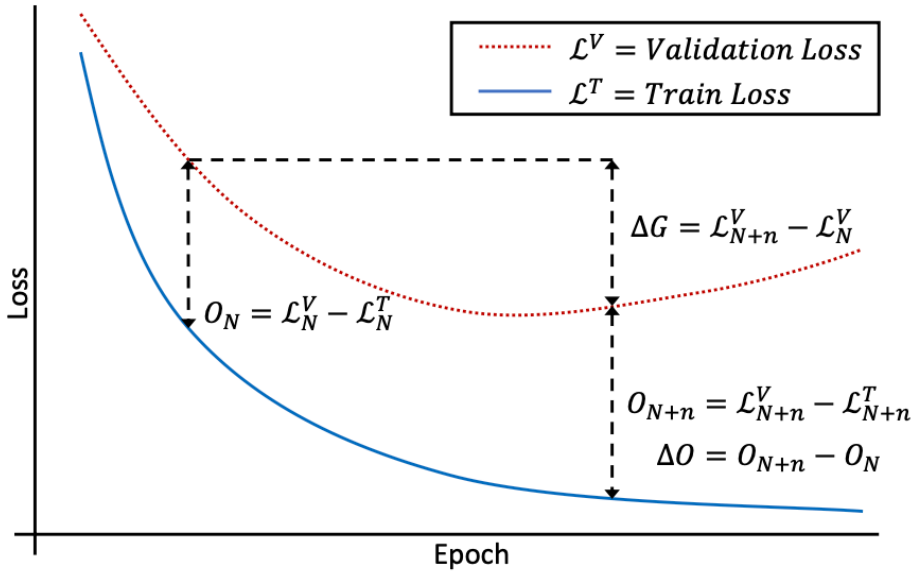
overfitting-to-generalization ratio (OGR)

$$OGR \equiv \left| \frac{\Delta O_{N,n}}{\Delta G_{N,n}} \right| = \left| \frac{O_{N+n} - O_N}{\mathcal{L}^*_N - \mathcal{L}^*_{N+n}} \right| \tag{1}$$

OGR即overfitting-to-generalization ratio，如(1)所示，它是一个量化过拟合的量。我们从下面的图进行解释：

首先是$\Delta O_{N,n}$这一项，该项等于$O_{N+n} - O_N$，表示在训练集与评估集loss的差值的变化量

然后是$\Delta G_{N,n}$这一项，该项等于$\mathcal{L}^*_N - \mathcal{L}^*_{N+n}$，$\mathcal{L}^*_N$表示了在hypothetical target distribution上真实的loss，我们认为近似等于验证集上的值。



由于两个CKPT的距离很小，其中$\hat{g}$是参数更新的梯度，我们用一范数对相关项近似：$\Delta G \approx \langle \nabla \mathcal{L}^*, \hat{g} \rangle$，$\Delta O \approx \langle \nabla \mathcal{L}^T - \nabla \mathcal{L}^*, \hat{g} \rangle$，则：

$$OGR^2 = \left(\frac{\langle \nabla \mathcal{L}^T - \nabla \mathcal{L}^*, \hat{g} \rangle}{\langle \nabla \mathcal{L}^*, \hat{g} \rangle} \right)^2 \tag{2}$$

Gradient-Blending (G-Blend)

Theory

$$w^* = \arg\min_w \mathbb{E} \left[\left(\frac{\langle \nabla \mathcal{L}^T - \nabla \mathcal{L}^*, \sum_k w_k v_k \rangle}{\langle \nabla \mathcal{L}^*, \sum_k w_k v_k \rangle} \right)^2 \right] \tag{3}$$

作者假设下面约束成立：

$$\left\langle \nabla \mathcal{L}^*, \sum_k w_k v_k \right\rangle = 1 \tag{4}$$

那么：

$$w^* = \arg\min_w \mathbb{E} \left[\left(\left\langle \nabla \mathcal{L}^T - \nabla \mathcal{L}^*, \sum_k w_k v_k \right\rangle \right)^2 \right] \tag{5}$$

然后计算期望：

$$
\begin{aligned}
&\mathbb{E}\left[\left(\left\langle \nabla\mathcal{L}^{\mathcal{T}}-\nabla\mathcal{L}^{*},\sum_{k}w_{k}v_{k}\right\rangle\right)^{2}\right]\\
&=\mathbb{E}\left[\left(\sum_{k}w_{k}\left\langle \nabla\mathcal{L}^{\mathcal{T}}-\nabla\mathcal{L}^{*},v_{k}\right\rangle\right)^{2}\right]\\
&=\mathbb{E}\left[\sum_{k,j}w_{k}w_{j}\left\langle \nabla\mathcal{L}^{\mathcal{T}}-\nabla\mathcal{L}^{*},v_{k}\right\rangle\left\langle \nabla\mathcal{L}^{\mathcal{T}}-\nabla\mathcal{L}^{*},v_{j}\right\rangle\right]\\
&=\sum_{k,j}w_{k}w_{j}\mathbb{E}\left[\left\langle \nabla\mathcal{L}^{\mathcal{T}}-\nabla\mathcal{L}^{*},v_{k}\right\rangle\left\langle \nabla\mathcal{L}^{\mathcal{T}}-\nabla\mathcal{L}^{*},v_{j}\right\rangle\right]\\
&=\sum_{k}w_{k}^{2}\sigma_{k}^{2}
\end{aligned}
\tag{6}
$$

其中$\sigma_{k}^{2}=\mathbb{E}\left[\left\langle \nabla\mathcal{L}^{\mathcal{T}}-\nabla\mathcal{L}^{*},v_{k}\right\rangle^{2}\right]$，在目标期望(6)上应用拉格朗日乘数法（约束(4)），则有：

$$
L=\sum_{k}w_{k}^{2}\sigma_{k}^{2}-\lambda\left(\sum_{k}w_{k}\left\langle \nabla\mathcal{L}^{*},v_{k}\right\rangle-1\right)
\tag{7}
$$

求微分：

$$
\frac{\partial L}{\partial w_{k}}=2w_{k}\sigma_{k}^{2}-\lambda\left\langle \nabla\mathcal{L}^{*},v_{k}\right\rangle=0
\tag{8}
$$

则：

$$
w_{k}^{*}=\lambda\frac{\left\langle \nabla\mathcal{L}^{*},v_{k}\right\rangle}{2\sigma_{k}^{2}}
\tag{9}
$$

根据约束(4)：

$$
1=\sum_{k}w_{k}\left\langle \nabla\mathcal{L}^{*},v_{k}\right\rangle=\sum_{k}\lambda\frac{\left\langle \nabla\mathcal{L}^{*},v_{k}\right\rangle^{2}}{2\sigma_{k}^{2}}
\tag{10}
$$

则：

$$
\lambda=\frac{1}{\sum_{k}\frac{\left\langle \nabla\mathcal{L}^{*},v_{k}\right\rangle^{2}}{2\sigma_{k}^{2}}}
\tag{11}
$$

设$Z=1/\lambda$，那么：

$$
w_{k}^{*}=\frac{1}{Z}\frac{\left\langle \nabla\mathcal{L}^{*},v_{k}\right\rangle}{2\sigma_{k}^{2}}
\tag{12}
$$

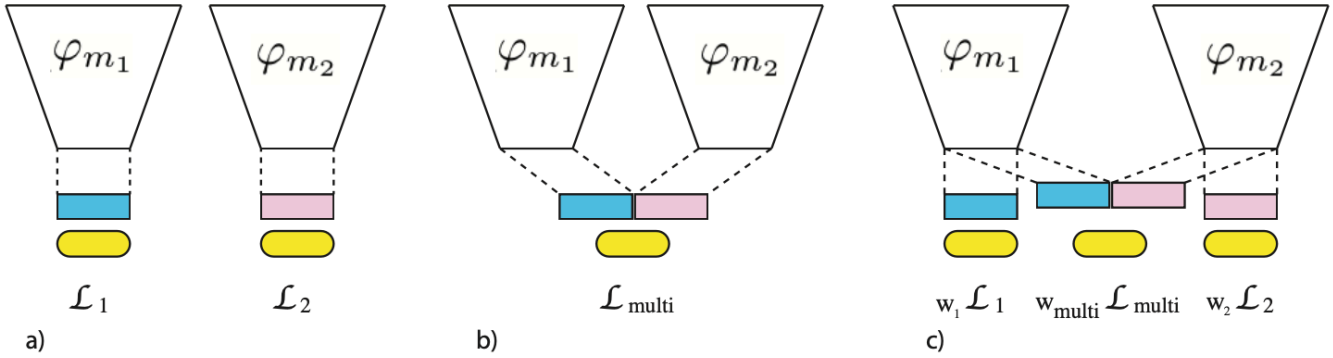
其中，$\sigma_{k}^{2}\equiv\mathbb{E}\left[\left\langle \nabla\mathcal{L}^{\mathcal{T}}-\nabla\mathcal{L}^{*},v_{k}\right\rangle^{2}\right]$ and $Z=\sum_{k}\frac{\left\langle \nabla\mathcal{L}^{*},v_{k}\right\rangle}{2\sigma_{k}^{2}}$

Practice: loss re-weighting

$$\mathcal{L}_{blend} = \sum_{i=1}^{k+1} w_i \mathcal{L}_i \tag{13}$$

为了计算OGR，有近似$\mathcal{L}^{\mathcal{V}} \approx \mathcal{L}^*$，根据$(12)$计算$w_i = \frac{1}{Z}\frac{G^i}{2O^{i^2}} = \frac{1}{Z}\frac{G_{N,n}}{2O_{N,n}{}^2}$

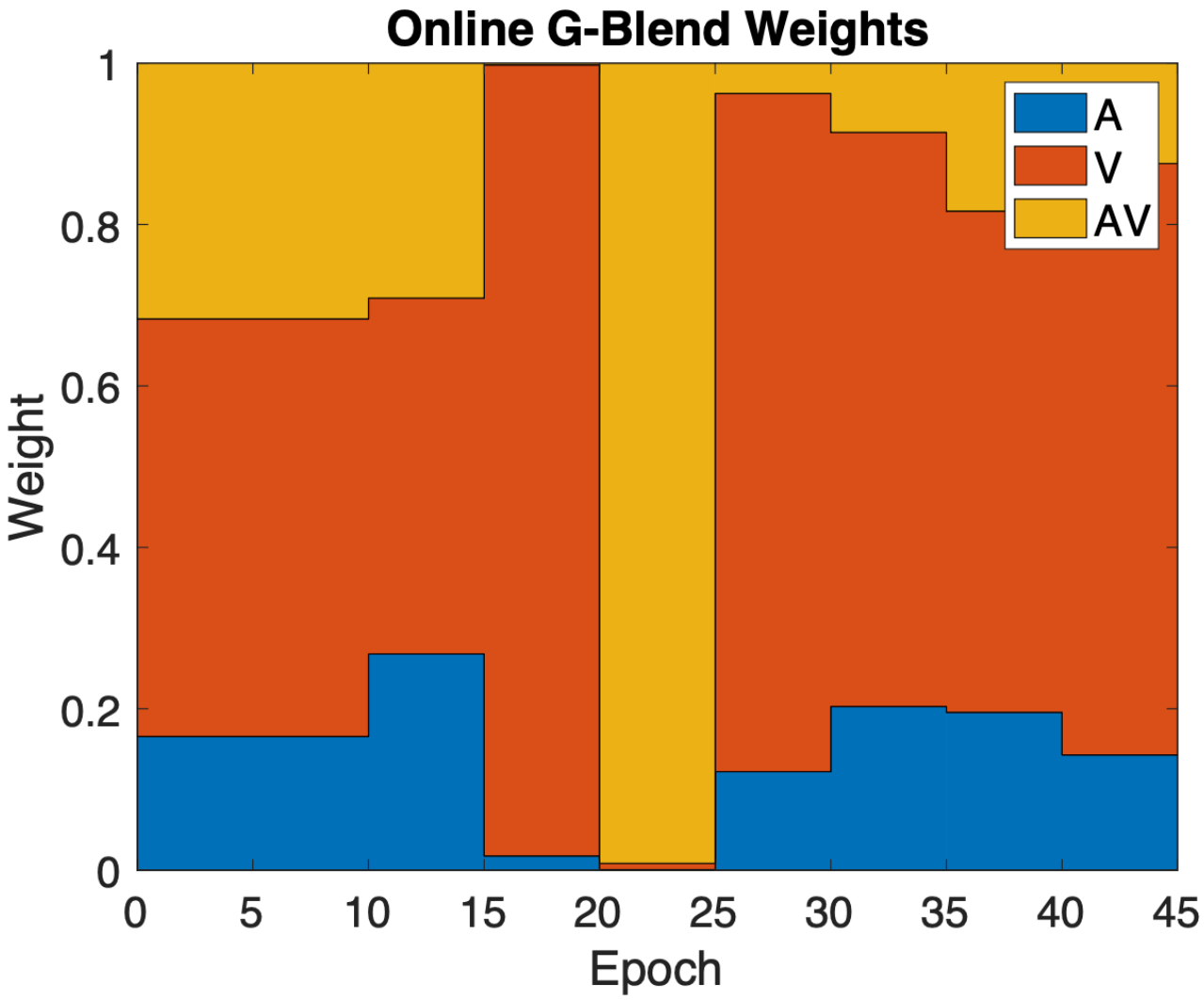
加权blending如下图所示：



实验

权重变化

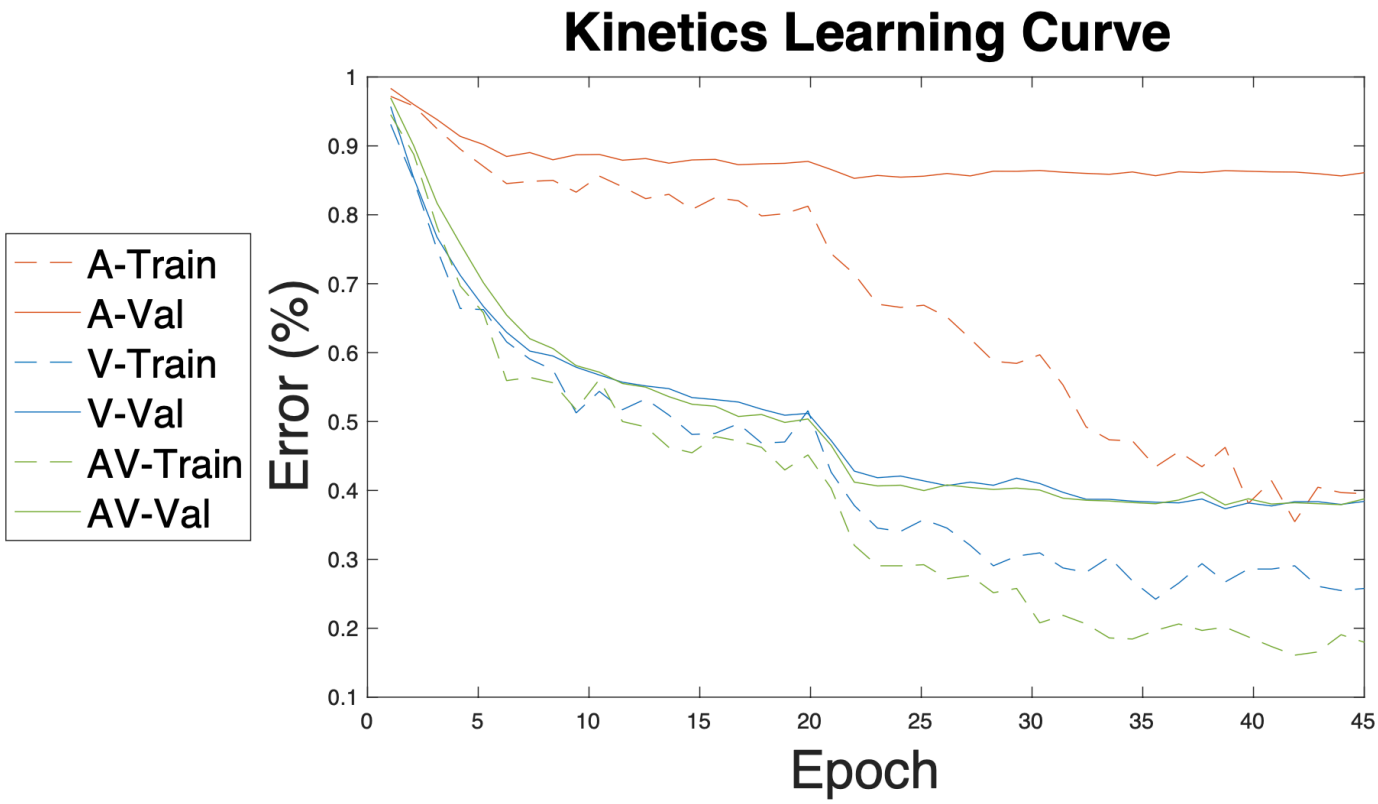
随着训练，权重变化：



- V始终占据一个比较大的权重，说明V的OGR小，即泛化的提升更大
- 15-20，只进行模态V的训练，20-25的突变，只学习joint
- 在不同的训练阶段，神经网络学习到的patterns是不同的

train-val loss变化

随着训练，各模态和joint head的train-val loss变化：



- 可以发现最后多模态的结果和单模态V在验证集上的性能差不多
- train-val之间有gap，说明overfitting
- 单模态A的效果很差，train-val的gap非常大