Data-efficient Alignment of Multimodal Sequences by Aligning Gradient Updates and Internal Feature Distributions: Supplementary Material

Jianan Wang[1], Boyang Li[2,3], Xiangyu Fan[1], Jing Lin[1], and Yanwei Fu[1,4]

[1]School of Data Science, Fudan University, China
[2]Nanyang Technological University, Singapore
[3]Alibaba-NTU Singapore Joint Research Institute
[4]MOE Frontiers Center for Brain Science, Fudan University, China

boyangli@outlook.com jinglin0224@163.com {jawang19, fanxy19, yanweifu}@fudan.edu.cn

A. Details on Experiment Setup

In this supplementary material, we give the experiment details in Section 4 in our main paper.

All the experiments share the same batch size 32 and the same learning rate scheduler, which halves the learning rate if training loss does not improve in the last ten epochs. The initial global learning rates for different experimental setups have been tuned for performance.

A.1. The "Our Pretraining" Baseline

This baseline employs Faster RCNN and BERT features, which were not used by [1]. With this baseline, we first pretrain NeuMATCH on the large LSMDC dataset and finetune it on YMS. For pretraining, the initial learning rate is set to 1×10^{-3} and weight decay to 1×10^{-6} and dropout to 0.1 for all fully-connected layers. During pretraining, we apply early stopping using the validation set, which stopped the training at Epoch 47. Finetuning on YMS uses the initial learning rate of 5×10^{-4} and dropout of 0.1 for all fully-connected layers. Weight decay is not used during finetuning. We use the Adam optimizer for both pretraining and finetuning.

A.2. The "LXMERT Features" Baseline

As another baseline, we extract features using the unimodal encoders from LXMERT. The full LXMERT model first feeds inputs through respective unimodal encoders, followed by a cross-modal encoder. As we do not have image-text correspondence before encoding, we simply use the pretrained unimodal encoders. We perform mean pooling over the RoI features extracted by the image encoder and over the word-level features from the text encoder.

In training this baseline, the initial learning rate is 3×10^{-4} with weight decay of 1×10^{-7} and dropout of 0.3 for all fully-connected layers. The Adam optimizer is used.

Feature	Optimizer	Normalization	Initial learning rate
Full	Adam	—	1×10^{-3}
Full	LARS	SBN	5×10^{-3}
Full	LARS	2×LN	4×10^{-3}
Full	LARS	4×LN	5×10^{-3}
RP	Adam	SBN	5×10^{-4}
RP	Adam	2×LN	1×10^{-2}
RP	Adam	4×LN	7×10^{-3}
RP	LARS	SBN	7×10^{-3}
RP	LARS	2×LN	5×10^{-3}
RP	LARS	4×LN	8×10^{-3}

Table 7: Initial global learning rates for ablation experiments. All models use a label smoothing regularization of 3%.

A.3. Ablation Experiments

Table 7 gives the experiment details from Section 4.4-4.6. All these experiments apply a label smoothing regularization to the distribution of ground truth label. The label smoothing hyperparameter $\epsilon = 0.03$. Weight decay and dropout are not used. In the setup of RP+Adam warm-up+SBN, we linearly increase the learning rate for 5 epochs from 5×10^{-5} to 5×10^{-4} without any label smoothing.

For completeness, we formally define label smoothing. With the hyperparameter ϵ, label smoothing modifies the ground-truth class probability to $(1 - \epsilon)$ and evenly distributes ϵ among the rest of the classes. We use the modified probability vector as the target in cross-entropy loss.

More formally, we can denote the ground-truth labels as $[y_1, \ldots, y_k, \ldots, y_K]$, $y_k \in \{0, 1\}$. When the true class is k, we set $y_k = 1$ and rest of the labels as 0. With label smoothing, the new labels are set to

$$y_k^{smooth} = (1 - \epsilon) y_k + \frac{\epsilon}{K}$$

We choose $\epsilon = 0.03$ for all the experiments in Table 7.

References

[1] Pelin Dogan, Boyang Li, Leonid Sigal, and Markus Gross. A neural multi-sequence alignment technique (NeuMATCH). In *CVPR*, 2018.