

# 学界 | FRATERNAL DROPOUT

2017-11-17 机器海岸线

选自 arXiv

作者: Konrad Żolna , Devansh Arpit , Dendi Suhubdy & Yoshua Bengio 等

机器海岸线编译

参与: 方建勇

## FRATERNAL DROPOUT

Konrad Żolna<sup>1,\*</sup>, Devansh Arpit<sup>2</sup>, Dendi Suhubdy<sup>2</sup> & Yoshua Bengio<sup>2,3</sup>

<sup>1</sup>Jagiellonian University

<sup>2</sup>MILA, Université de Montréal

<sup>3</sup>CIFAR Senior Fellow

论文链接: <https://arxiv.org/pdf/1711.00066>

**摘要:** 递归神经网络 (RNN) 是用于语言建模和顺序预测的神经网络中重要的一类架构。然而, 与前馈神经网络相比, RNN 优化是困难的。文献中提出了许多技术来解决这个问题。在本文中, 我们提出了一个简单的技术, 称为 FRATERNAL DROPOUT, 利用 DROPOUT 来实现这一目标。具体来说, 我们建议训练两个相同的副本 RNN (共享参数) 与不同的 DROPOUT 掩码, 同时最小化他们 (pre-softmax) 预测之间的差异。通过这种方式, 我们的规则化鼓励 RNN 的表示与 DROPOUT 掩码是不相关的, 因此是鲁棒的。我们证明了我们的正则化项是由期望 - 线性 DROPOUT 目标所限定的, 这个目标已经被证明是为了解决由于训练和推理 DROPOUT 阶段之间的差异而导致的差距。我们评估我们的模型, 并在两个基准数据集 Penn Treebank 和 Wikitext-2 上实现序列建模任务中的最新结果。我们还表明, 我们的方法导致图像字幕 (Microsoft COCO) 和半监督 (CIFAR-10) 任务的性能提高很大。

Model	Parameters	Validation	Test
Zaremba et al. (2014) - LSTM (medium)	10M	86.2	82.7
Zaremba et al. (2014) - LSTM (large)	24M	82.2	78.4
Gal & Ghahramani (2016) - Variational LSTM (medium)	20M	81.9	79.7
Gal & Ghahramani (2016) - Variational LSTM (large)	66M	77.9	75.2
Inan et al. (2016) - Variational LSTM	51M	71.1	68.5
Inan et al. (2016) - Variational RHN	24M	68.1	66.0
Zilly et al. (2016) - Variational RHN	23M	67.9	65.4
Melis et al. (2017) - 5-layer RHN	24M	64.8	62.2
Melis et al. (2017) - 4-layer skip connection LSTM	24M	60.9	58.3
Merity et al. (2017a) - AWD-LSTM 3-layer	24M	60.0	57.3
<i>fraternal dropout</i> + AWD-LSTM 3-layer	24M	<b>58.9</b>	<b>56.8</b>

表 1: Penn Treebank 词级语言建模任务的复杂度。

Model	Parameters	Validation	Test
Merity et al. (2016) - Variational LSTM + Zoneout	20M	108.7	100.9
Merity et al. (2016) - Variational LSTM	20M	101.7	96.3
Inan et al. (2016) - Variational LSTM	28M	91.5	87.0
Melis et al. (2017) - 5-layer RHN	24M	78.1	75.6
Melis et al. (2017) - 1-layer LSTM	24M	69.3	65.9
Melis et al. (2017) - 2-layer skip connection LSTM	24M	69.1	65.9
Merity et al. (2017a) - AWD-LSTM 3-layer	34M	68.6	65.8
<i>fraternal dropout</i> + AWD-LSTM 3-layer	34M	<b>66.8</b>	<b>64.1</b>

表 2: 对 WikiText-2 字级语言建模任务的复杂度。

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Show and Tell Xu et al. (2015)	66.6	46.1	32.9	24.6
Baseline	68.8	50.8	36.1	25.6
<i>Fraternal dropout</i> , $\kappa = 0.015$	<b>69.3</b>	51.4	36.6	26.1
<i>Fraternal dropout</i> , $\kappa = 0.005$	<b>69.3</b>	<b>51.5</b>	<b>36.9</b>	<b>26.3</b>

表 3: Microsoft COCO 图像字幕任务的 BLEU 分数，使用 FRATERNAL DROPOUT 是模型之间的唯一区别，剩下的超参数是一样的。

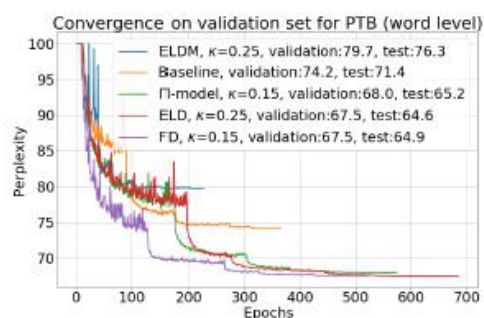
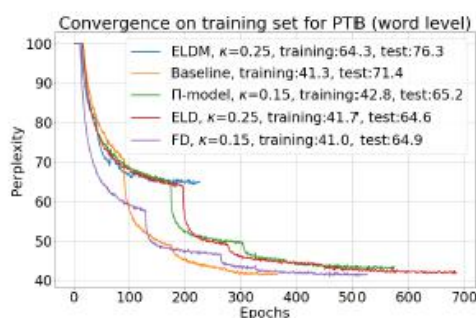


图 1: 消融研究: 使用单层 LSTM (10M 参数) 对 PTB 字级模型进行训练 (左) 和验证 (右)。 这些曲线研究基线模型的学习动力学, 模型, 期望线性丢失 (ELD), 期望线性 dropout modification (ELDM) 和 fraternal dropout (FD, 我们的算法)。 我们发现 FD 的收敛速度比正规化者要快。

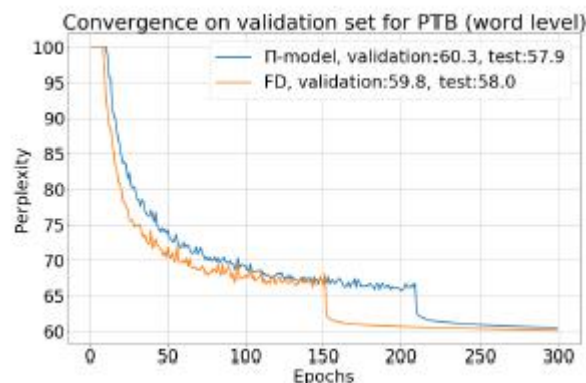


图 2: 对模型和 fraternal dropout. 的 PTB 词水平模型验证复杂度, 我们发现 FD 收敛速度更快。

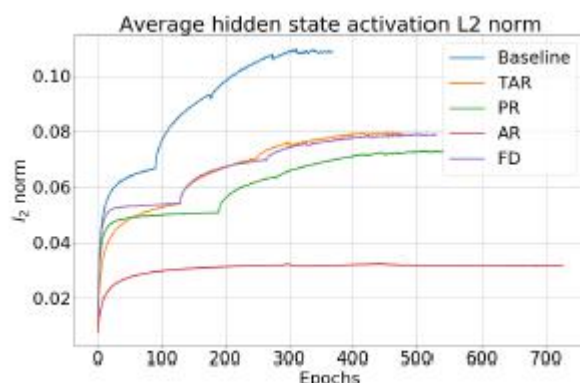


图 3: 当使用所描述的任何调节器时, 平均隐藏状态激活减少。

Model	Dropout rate	Unlabeled data used	Validation	Test
Ordinary	0.1	No	78.4 ( $\pm 0.25$ )	76.9 ( $\pm 0.31$ )
None	0.0	No	78.8 ( $\pm 0.59$ )	77.1 ( $\pm 0.3$ )
<i>Fraternal</i> ( $\kappa = 0.05$ )	0.05	No	79.3 ( $\pm 0.38$ )	77.6 ( $\pm 0.35$ )
Ordinary + $\Pi$ -model	0.1	Yes	80.2 ( $\pm 0.33$ )	78.5 ( $\pm 0.46$ )
<i>Fraternal</i> ( $\kappa = 0.15$ )	0.1	Yes	<b>80.5 (<math>\pm 0.18</math>)</b>	<b>79.1 (<math>\pm 0.37</math>)</b>

表 4: 基于 ResNet-56 的模型改变 (半监督) CIFAR-10 数据集的精度。我们发现我们的算法与  $\Pi$ -model 同等地执行。 当不使用未标记的数据时, 普通的丢失会损害性能, 而 fraternal dropout 提供稍好的结果。 这意味着我们的方法可能是有益的, 当我们缺乏数据, 并需要使用额外的正则化方法。

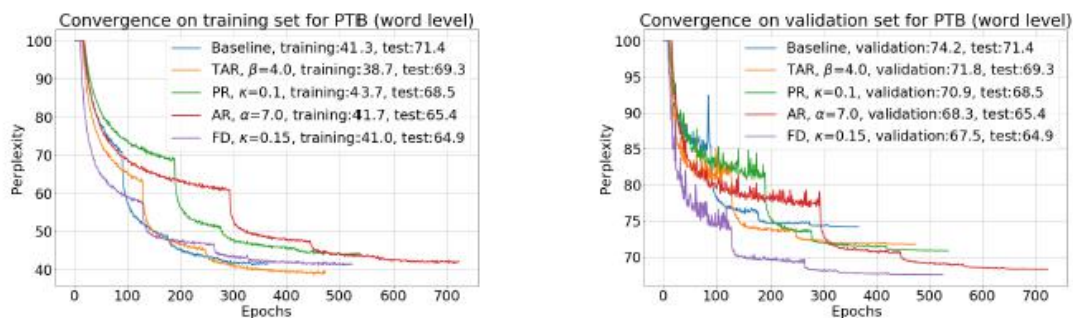


图4：消融研究：使用单层LSTM（10M 参数）对PTB 字级模型进行训练（左）和验证（右）。这些曲线研究基线模型，时间活动正则化（TAR），预测正则化（PR），活动正则化（AR）和fraternal dropout（FD，我们的算法）的学习动态。我们发现FD比正规化器收敛速度更快，概括性更好。

Dropout	Fine-tuning	PTB		WT2	
		Validation	Test	Validation	Test
Ordinary	None	60.7	58.8	69.1	66.0
Ordinary	One	60.0	57.3	68.6	65.8
<i>Fraternal</i>	None	59.8	58.0	68.3	65.3
<i>Fraternal</i>	One	58.9	56.8	<b>66.8</b>	<b>64.1</b>
<i>Fraternal</i>	Two	<b>58.5</b>	<b>56.2</b>	—	—

表5：AWD-LSTM 3 层模型的微调的重要性。Penn Treebank 和 WikiText-2 语言建模任务的复杂度。。

本文为机器海岸线编译，转载请联系 [fangjianyong@zuaa.zju.edu.cn](mailto:fangjianyong@zuaa.zju.edu.cn) 获得授权。

✂-----