

神经网络语言模型的性能优化研究

On Optimization Perspective of Neural Language Model

北京航空航天大学 计算机学院 研究生开题答辩

答辩人：姜楠 SY1506330

导师：荣文戈副教授

2016 年12 月20 日

论文选题的背景与意义

- 神经网络在语言模型中应用广泛；
- 循环神经网络语言模型精确度最好，但是计算费时需要优化。
- 本课题调研各种优化方案，并结合各方面的已有成果，在保证模型精度不降低的情况下，使其速度尽可能的达到最快。
- 同时我们还考虑了使用当前流行的CUDA计算方案。

语言模型历史发展

- 1. N-gram:
 - Kneser-Ney n-gram
- 2. Neural Network Language Model:
 - neural probabilistic language model 2007
- 3. Recurrent Neural Network Language Model
 - rnnlm 2010
- 4. LSTM/GRU language Model
 - lstm_lm 2012
- 5. Character Level Language Model
 - lstm-char-cnn 2016

部分参考文献

- Chen S F, Goodman J. An empirical study of smoothing techniques for language modeling[J]. Computer Speech & Language, 1999, 13(4): 359-394.
- Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. journal of machine learning research, 2003, 3(Feb): 1137-1155.
- Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model[C]//Interspeech. 2010, 2: 3.
- Sundermeyer M, Schlüter R, Ney H. LSTM Neural Networks for Language Modeling[C]//Interspeech. 2012: 194-197.
- Kim Y, Jernite Y, Sontag D, et al. Character-aware neural language models[J]. arXiv preprint arXiv:1508.06615, 2015.

国内外研究现状及发展动态

- 运行方案
 - CPU or GPU
- Larger vocabulary optimizations
 - 1. spherical family (曲面函数族)
 - 2. sampling estimation.
 - Noise contrastive estimation
 - blackout
 - 3. structured output layer
 - Class-based hierarchical softmax
 - Tree-based hierarchical softmax
 - Adaptive softmax

GPU和CPU性能对比

Operations	Results
matrix multiplication; Convolution; large element-wise operations;	Good
Indexing; dimension-shuffling; Reshaping;	Moderate
Summation over rows/columns	Bad
Copying large data	Terrible

国内外研究现状及发展动态

- Computational expensive.
- 1. spherical family
- 2. sampling estimation.
 - Noise contrastive estimation
 - blackout
- 3. structured output layer
 - Class-based hierarchical softmax
 - Tree-based hierarchical softmax
 - Adaptive softmax



1. spherical family(球面函数族)

Taylor Expansion

➤ SOFTMAX.

$$\mathbf{o} \mapsto f_{soft}(\mathbf{o})_k = \frac{\exp(o_k)}{\sum_{i=1}^D \exp(o_i)}.$$

➤ SPHERICAL SOFTMAX.

$$\mathbf{o} \mapsto f_{sph_soft}(\mathbf{o})_k = \frac{o_k^2 + \epsilon}{\sum_{i=1}^D (o_i^2 + \epsilon)}.$$

➤ TAYLOR SOFTMAX.

$$\mathbf{o} \mapsto f_{tay_soft}(\mathbf{o})_k = \frac{1 + o_k + \frac{1}{2}o_k^2}{\sum_{i=1}^D (1 + o_i + \frac{1}{2}o_i^2)}.$$

2. 采样方案

Cheap to compute in training while it fails in testing time.



Noise contrastive estimation [2010-2012]

Importance sampling. [2008]

Blackout sampling.[2016]

2.1 Noise contrastive estimation

- True data distribution Vs Noises data distribution.
- Usually use unigram distribution to model the noise distribution.

$$p(Y = \text{true}|w) = \frac{p_d(w)}{p_w(w) + kp_n(w)}$$

$$\tilde{p}(D = 1; h) = \frac{\exp(h^\top v'_{w_0})}{\exp(h^\top v'_{w_0}) + k * q(w_0)}$$

$$\tilde{p}(D = 0; h) = \prod_{i=1}^k \frac{k * q(w_i)}{\exp(h^\top v'_{w_i}) + k * q(w_i)}$$

$$\mathcal{L}(\theta) = \log \tilde{p}(D = 1|h_\theta) + \log \tilde{p}(D = 0|h_\theta)$$

2.2 Blackout sampling

➤ softmax


$$p_{\theta}(w_i|s) = \frac{\exp(\langle \theta_i, s \rangle)}{\sum_{j=1}^V \exp(\langle \theta_j, s \rangle)} \quad \forall i \in \{1, \dots, V\}.$$

➤ Cost function

$$J_{ml}^s(\theta) = \log p_{\theta}(w_i|s),$$

➤ gradient

$$\begin{aligned} \frac{\partial J_{ml}^s(\theta)}{\partial \theta} &= \frac{\partial}{\partial \theta} \langle \theta_i, s \rangle - \sum_{j=1}^V p_{\theta}(w_j|s) \frac{\partial}{\partial \theta} \langle \theta_j, s \rangle, \\ &= \frac{\partial}{\partial \theta} \langle \theta_i, s \rangle - \mathbb{E}_{p_{\theta}(w|s)} \left[\frac{\partial}{\partial \theta} \langle \theta_w, s \rangle \right]. \end{aligned}$$



部分参考文献

- Mnih A, Kavukcuoglu K. Learning word embeddings efficiently with noise-contrastive estimation[C]//Advances in Neural Information Processing Systems. 2013: 2265-2273.
- Gutmann M, Hyvärinen A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models[C]//AISTATS. 2010, 1(2): 6.
- Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems. 2013: 3111-3119.
- Mnih A, Teh Y W. A fast and simple algorithm for training neural probabilistic language models[J]. arXiv preprint arXiv:1206.6426, 2012.



课题研究内容与关键技术

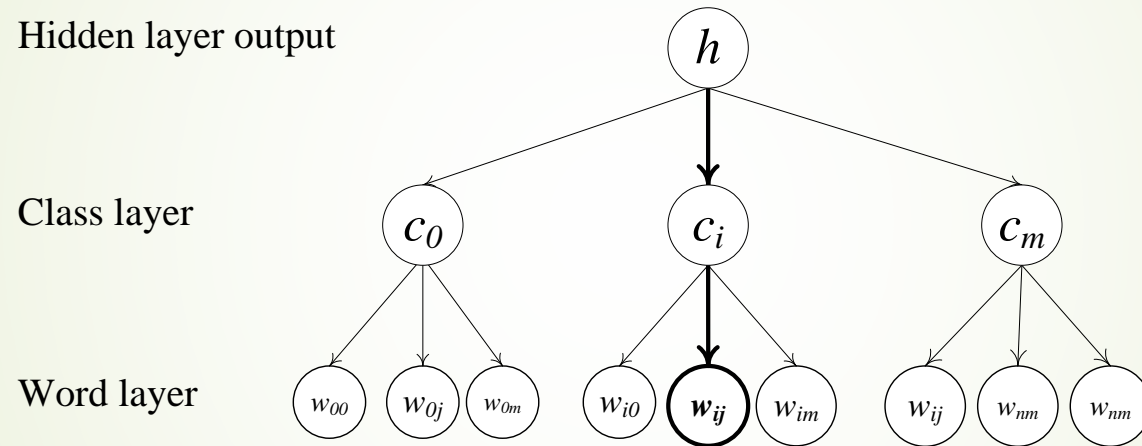
3. Structured output layer



课题研究内容与关键技术

- 1. class-based hierarchical softmax.
- 2. tree-based hierarchical softmax.
- 3. word clustering.

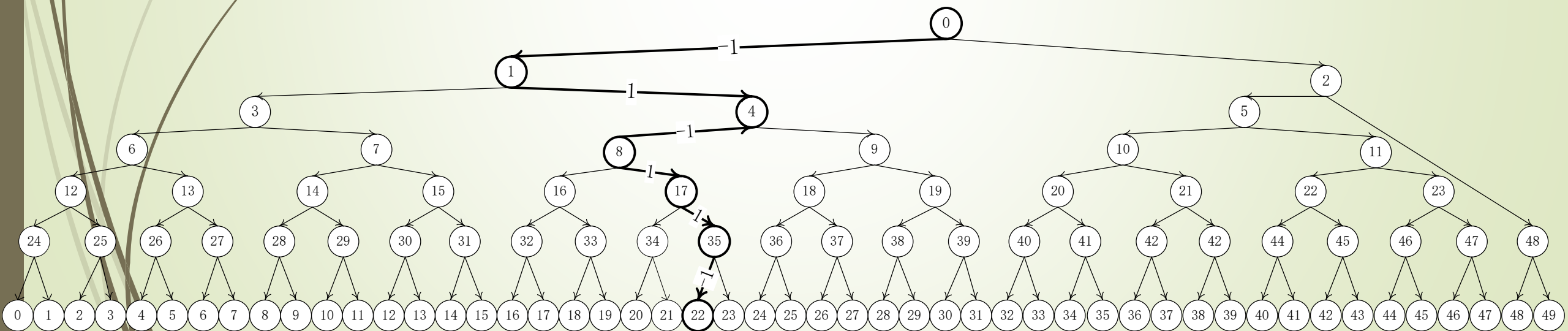
3.1 Class-based hierarchical softmax



$$p(w_{ij}|h) = p(c_i|h) \cdot p(w_{ij}|c_i)$$

3.2 Tree-based Hierarchical softmax

➡ 加速比: $v/\log(V)$

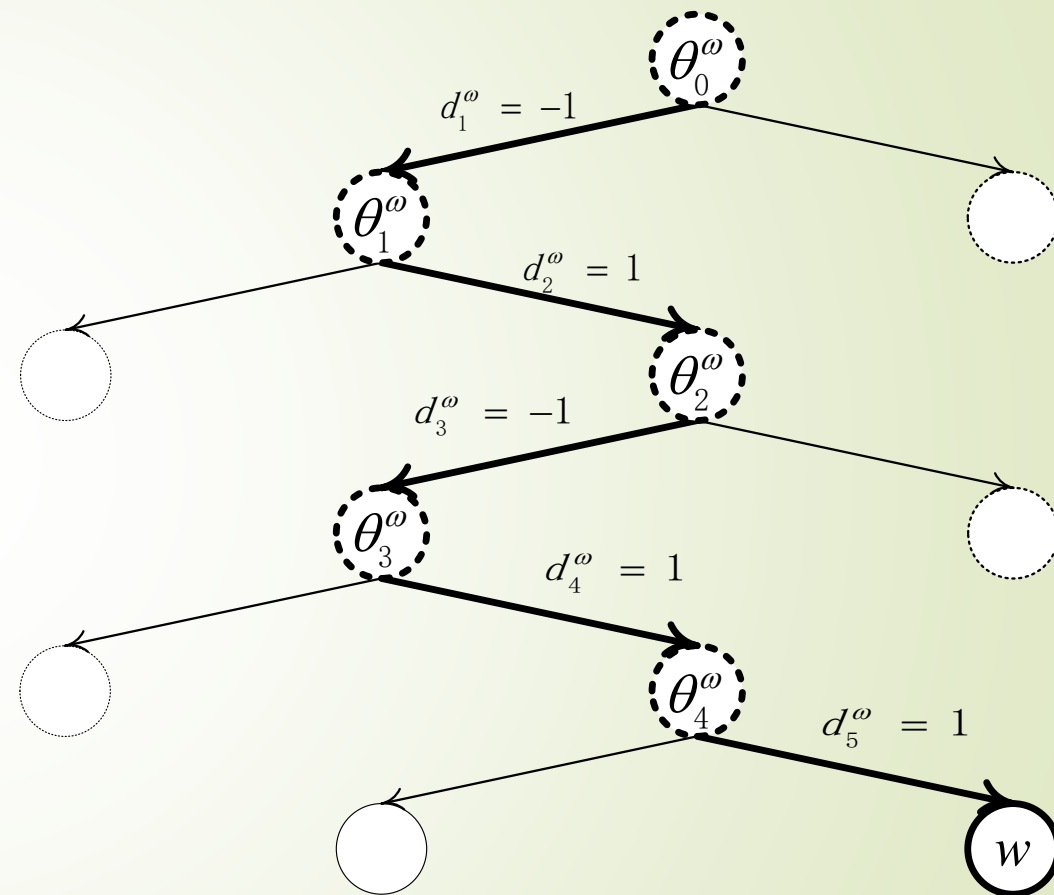


模型数学化

$$p(d_i^w = \pm 1 | \theta_{i-1}^w, h) = \sigma(d_i^w \theta_{i-1}^w h)$$

$$p(w|h) = \prod_{i=1}^l p(d_i^w | \theta_{i-1}^w, h) = \prod_{i=1}^l \sigma(d_i^w \theta_{i-1}^w h)$$

$$\mathcal{L}(\theta|h) = \sum_{i=1}^l \log(1 + \exp(-d_i^w \theta_{i-1}^w h))$$





构建树的策略

- 均匀划分单词类别：
 - Frequency binning.
 - Word-net binning.
 - word-embedding clustering.
- 非均匀划分单词类别：
 - Brown clustering



评价指标

- 混杂度(Perplexity).
- 计算复杂度
- 训练参数

一些现有的代码库

Theano

- https://github.com/jiangnanHugo/language_modeling
- https://github.com/pascal20100/factored_output_layer

C++

- <https://github.com/IntelLabs/rnnlm>

CUDA

- https://github.com/isi-nlp/Zoph_RNN

Tensorflow

- https://github.com/tensorflow/models/tree/master/lm_1b



开发环境设置

- Linux: 操作系统
- R: 主要用于数据收集和图表处理
- Python2.7: 使用的开发语言和开发环境
- Theano: 主要的建模语言
- 其他: bash script 和 cuda



时间安排

- 2016 年12 月 2017 年1 月: 整理资料, 学习研究语言模型的领域知识;
- 2017 年2 月 2017 年4 月: 研究学习深度学习模型的知识, 特别是循环神经网络的建模过程;
- 2017 年5 月2017 年7 月: 调研并实现解决大词表问题的主要手段, 并实现基本代码框架;
- 2017 年8 月2017 年10 月: 实验验证与完善;
- 2017 年11 月2018 年3 月: 资料整理和论文撰写.



Thanks !