小組論文口稿

3.

本文提出一種新的 LSTM 訓練的新方法，將門閥的輸出值朝向 0 或 1。
雖然限制了模型能力， 但沒有性能下降，由於其更好的 generalization 能力(機器學習演算法對新鮮樣本的適應能力)得到比較好的比較結果。
可以更易於 LSTM 單元壓縮，本文結果優於為壓縮的模型

4.

解決傳統 RNN 的依賴性和梯度消失問題，提出了 LSTM 網絡，它引入了門閥函數來控制循環單元中的訊息流。
輸入門閥功能用於找到要被吸收到隱藏上下文中的相關訊息，以及用於預測和決策的輸出門功能。
為了便於優化，在實際中實現時，通常使用 element-wise sigmoid function 模擬門閥的功能，使其輸出是介於 0 和 1 之間。

5.

LSTM 通常比傳統 RNN 執行得更好。然而，深入研究單元時，會發現 gate 的價值沒有那麼有意義。
在圖 1 中，分佈在 forget gates 和 input gates 不清晰大多數值處於中間狀態（大約 0.5），這意味著大多數門閥值在 LSTM 中是不明確的。
其他作者也表明 LSTM 的大多數細胞坐標很難找到特別的意義。

6.

所以本篇將門閥的值推向它們的範圍邊界（0,1），將邊界推向 0 與 1 有下列優勢首先，它很好地符合門的發展的最初目的：到通過“打開”或“關閉”獲取訊息或跳過反復計算期間的門，反映更多準確而清晰的語言和結構信息。
第二，類似於圖像分類中的 BitNet（Courbariaux 等人，2016），通過推動激活函數進行二值化，我們可以學習一個可以進一步壓縮的模型(更有效率)。
第三，將 LSTM 訓練為二進制值門 能夠更好地生成學習模型。

7.

將門閥值的輸出推向了這樣的位置離散值是具有挑戰性。透過銳化 sigmoid fumction。但是，這相當於重新調整輸入而不能保證學習門的值接近 0 或者 1。
所以，在本文中利用 Gumbel-Softmax 來開發的方法
透過 Gumbel-Softmax 運用於門閥估計從參數給出的伯努利分佈中採樣的值，並使用標準的反向傳播方法訓練 LSTM 模型。我們稱之為學習模型 Gumbel-Gate LSTM（G2 -LSTM）。

8.
方法限制門輸出接近邊界，從而降低了表現力。 但是，沒有性能下降。此外，模型可以獲得更好與其他模型的可比性。
學習模型很容易進一步壓縮。將幾種模型壓縮算法應用於門閥中的參數，包括低精度預估和低排序預估，結果表明本文壓縮模型可以好比沒有壓縮模型更好本文調查了一組樣本並找到 gate 在學習的模型中，有意義且直觀解釋。可以自動顯示模型學習句子的界限。

9.
連續離散隨機變數在隨機計算圖中，並使用 Gumbel-Max 技巧，讓結果可以較接近原本分布。
透過使用 Gumbel-Softmax Estimator，我們可以生成樣本 y =（y1，...，yk）來近似分類分佈。此外，由於隨機性 q 獨立於 π（通常由一組參數定義），我們可以使用重新參數化技巧來優化使用標準反向傳播算法的模型參數。
Gumbel-Softmax 估計器已被多種應用所採用，例如變量自動編碼器（Jang 等人，2016），生成對抗網絡（Kusner & Hernández-Lobato，2016）和語言生成（Subramanian 等，2017）。據我們所知，這是第一部介紹 Gumbel-Softmax Estimator 運用在 LSTM 訓練上。
https://www.itread01.com/content/1545003007.html

10.
損失 f（·）的平坦最小值 x 對應於函數 f 的值在 x 的相對大的鄰域中緩慢變化的點。相反，sharp 的最小 x 使得函數 f 在 x 的小鄰域中快速變化。sharp 最小值處的損失函數的敏感性負面影響訓練模型對新數據的泛化能力。
通過使用小批量訓練顯示，學習模型更可能收斂到平坦區域。

11.
此部分為 RNN 與 LSTM 公式
LSTM 單元需要比參數更多的參數，大參數的百分比用於計算門 （sigmoid）功能。

12.
如果我們可以推動 gate 的輸出到 sigmoid 函數的飽和區域（即朝向 0 或 1），關於門中參數的損耗函數將是平坦的：如果門中的參數擾動，由於 sigmoid 形運算符，門的輸出變化很小（見圖 2）
然後損失會變小，這代表損失的平坦區域。第一， 因為這樣的模型對於小的參數變化是穩健的，它是對不同的模型壓縮方法是穩健的，例如，低精度壓縮或低排序壓縮。

第二， 正如相關作者所討論的那樣，最小平坦區域更有可能更好的結果，從而朝向二進制值門可以帶來更好的測試性能。

13.
透過較小的 tempture 參數來銳化 Sigmoid 功能，通常，在大範圍內使用具有大學習率的初始點將損害優化過程，並且顯然不能保證輸出在訓練之後接近邊界。

14.
我們利用最近開發的 GumbelSoftmax 技巧。這種技巧在近似離散分佈中是有效的，並且是在隨機計算圖中學習離散隨機變量的廣泛使用的方法之一。我們首先提出關於伯努利分佈的這種技巧的近似能力的命題，這將在我們提出的算法中使用。

15.
透過這種方式，可以優化二進制值門。只將 input gate 的 output 的值與 forget gate 的值推向二元值，output gate 通常需要精細訊息用於決策這使二進制值理不理想。為了證明這一點，進行了類似的實驗並觀察了性能，將 input gates 與 forget gates 的輸出推向 0 與 1，。
在前向傳遞中，我們首先獨立採樣值 U 在每個時間，然後更新 G2 -LSTMs 使用 Eqn。
在向後傳球中，因為 G 是連續且可微分的關於參數和損失是連續的相對於 G 而言，我們可以使用任何標準基於梯度的方法來更新模型參數。

17.
此任務是訓練 LSTM 模型以正確預測以前單詞為條件的下一個單詞。通過預測困惑來評估模型：較小的困惑，更好的預測。
此部分說明語言模型參考相關作者設計的部分。使用三層的 LSTM 模型與 ASGD 優化，500 Epoch training 與 500 Epoch finetune，將 tempture 設定在 0.9，加入 neural cache model 在模型改善困惑度

18.
此部分說明翻譯資料集
德文至中文部分使用 IWSLT'14 資料，預處理使用 byte pair encoding 方法。
英文至德文 WMT'14 資料，也使用 byte pair encoding 方法處理資料。
兩者 資料均去除 64 子詞單元，在德文至英文部分使用兩階層 encoder-decoder 架構，將 word embedding 與 hidden state 設定至 256

19.
第一個算法（我們稱之為 Baseline），我們刪除了 Gumble-Softmax 技巧並使用標準訓練模型優化方法。
第二種算法（我們稱之為 Sharpened Sigmoid），我們使用尖銳的 Sharpened Sigmoid 如 3.2 節所述，設定 τ= 0.2 並檢查 這種技巧是否能帶來更好的表現。

20.
Plexity 值越低較佳
BLEU 較高較佳
在 language model 我們的模型優於其他作者
在 machine translation 部份，唯一不同在 G2-LSTM 為訓練演算法，雖然它們採用相同的模型結構，G2 的效果更好 -LSTM 證明了我們提出的訓練方法的有效性。這表明了這一點 將門的輸出限制為二進制值 根本不會帶來性能下降。相反， 表現甚至更好。我們得出結論，這樣的好處可能來自更好的泛化能力。
我們還列出了以往文學作品的表現， 可以採用不同的模型架構或設置。 對於語言建模，與表中列出的先前工作相比，我們獲得了更好的性能結果。對於英語→德語翻譯，我們的結果是 比 GNMT 更差（Wu et al。，2016）因為他們使用堆疊 八層模型，而我們只使用三層模型。

21.
接下來是敏感性分析，壓縮 input and forget gates 參數，所以模型可以變較小，並使用 round and clip 在 input and forget gates
測試了兩種低精度壓縮設置。在第一個設置（命名為 Round）中，我們使用 Eqn 舍入參數。 （17）。通過這種方式，我們減少了門中參數的支持集。在第二個設置（命名為 Round＆Clip）中，我們使用 Eqn 進一步將舍入值剪切為固定範圍。 （18）因此限制了不同值的數量。由於這兩個任務差別很大，我們為語言建模任務設置了舍入參數 r = 0.2 和剪輯參數 c = 0.4，並為神經機器翻譯設置了 c = 1.0 和 r = 0.5。因此，語言建模中的輸入門和遺忘門的參數只能取（0.0，±0.2，±0.4）和（0.0，±0.5，±1.0）的值用於機器翻譯。

22.
我們將輸入/忘記門的參數矩陣壓縮到較低排序的矩陣通過奇異值分解，可以減少模型大小導致更快的矩陣乘法。特定語言建模任務的隱藏狀態是 比神經機器翻譯更大的維度，我們設置 rank = 64/128 用於語言建模和 rank = 16/32 用於神經機器翻譯。
在語言模型上低精度壓縮都非常穩健
在機器翻譯上可以獲得與具有完整參數的模型大致相當的翻譯精度。所有結果表明，使用我們提出的方法訓練的模型對參數壓縮不太敏感。

23.
我們從訓練集中抽取了 10000 個句子對，並將它們輸入到學習模型中。我們得到了解碼器第一層中輸入/忘記門的輸出值向量。我們記錄了輸出向量中每個元素的值，並繪製了圖 1 和圖 3 中的分佈。
在 LSTM 中，門閥極值的分佈是相對均勻的並且具有沒有明確的集中。相反，輸入的值 G2 的大門 -LSTM 集中在靠近的地區 1，這表明我們學到的模型試圖保持最多來自輸入詞的訊息;忘記門閥集中在邊界區域，接近 0 的區域或接近 1 的區域。這一觀察表明我們的訓練算法符合我們的期望 並成功將大門推至 0/1。

24.
除了在一組採樣訓練數據上的門值的總體分佈之外，這裡我們提供了一個抽樣句子的案例研究。
計算了每個字的輸入和遺忘門函數的輸出向量的平均值。特別是，我們關注第一層中輸入/遺忘門函數的平均值，並檢查平均值是否合理。
G2-LSTM 不會降低輸入門功能的信息，因為所有字的平均值都相對較大。相反，LSTM 的輸入門的平均值有時很小（小於 0.5），即使對於像“錯誤”這樣的有意義的詞也是如此。
由於這些單詞未包含在 LSTM 中，因此無法對其進行有效編碼和解碼，從而導致錯誤的轉換結果。
其次，對於 G2-LSTM，忘記門的值小的大多數單詞是功能詞（例如，連詞和標點符號）或子句中的邊界。也就是說，我們的訓練算法確實確保了模型忘記有關句子邊界的信息，並用新輸入重置隱藏狀態。

25.
在本文中，利用開發的 Gumbel Softmax estimator 為 LSTM 設計了一種新的訓練算法。訓練算法可以將輸入和忘記門的值推到 0 或 1，從而產生強大的 LSTM 模型。語言建模和機器翻譯的實驗證明了所提出的訓練算法的有效性。
將在未來探索以下方向。首先，我們 將我們的算法應用於更深層次的模型（例如，8 +層） 並測試較大的數據集。其次，我們考慮過了 語言建模和機器翻譯的任務。我們 將研究更多的應用，如問答和文字摘要。