

以看圖寫詩為例的創意文本生成

Data dev- Data Scientist

Johnson Wu

Agenda

- ▶ 什麼是基於視覺的文本生成
- ▶ 什麼是人工智能寫詩
- ▶ Case study of 看圖寫詩--原理、訓練與評價
- ▶ 成果展示

基於視覺輸入的文本生成分類 (Bernardi et al. 2016)

- ▶ 1. 基於視覺輸入的生成
 - ▶ 2. 基於視覺輸入的索引
 - ▶ 3. 基於多模態輸入的索引
-
- ▶ 監督式學習：需要圖像與文字的對應等資料



"man in black shirt is playing guitar."



"construction worker in orange safety vest is working on road."



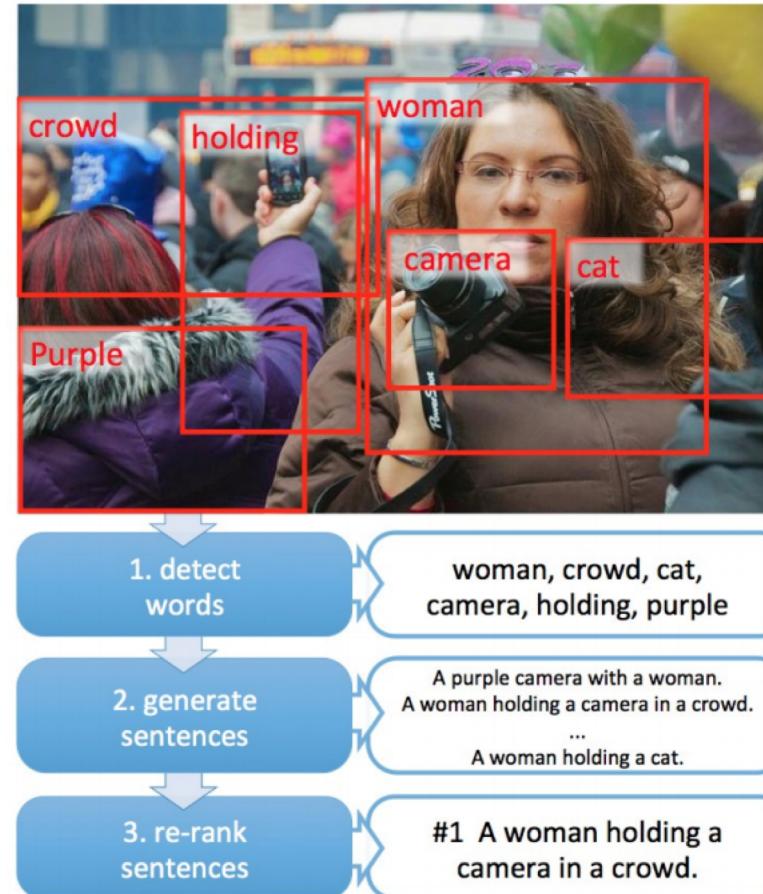
"two young girls are playing with lego toy."

Google
images

1. Patterson, G., Xu, C., Su, H., and Hays, J. The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision* 108(1-2):59-81. 2014
2. Devlin, J., Cheng, H., Fang, H., Gupta, S., Deng, L., He, X., Zweig, G., and Mitchell, M. Language models for image captioning: The quirks and what works. *arXiv preprint arXiv:1505.01809*. 2015
3. Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., and Ng, A. Y. Grounded compositional semantics for finding and describing images with sentences. *TACL* 2:207-218. 2014
4. Soto, A. J., Kiros, R., Keselj, V., and Milios, E. E. Machine learning meets visualization for extracting insights from text data. *AI Matters* 2(2):15-17. 2015
5. Karpathy, A., and Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3128-3137. 2015
6. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2625-2634. 2015
7. Schwarz, K., Berg, T. L., and Lensch, H. P. Autoillustrating poems and songs with style. In *Asian Conference on Computer Vision*, 87-103. 2016

From Captions to Image Concepts and Back (Fang et el, 2014)

- ▶ 2014 MSCOCO rank #1



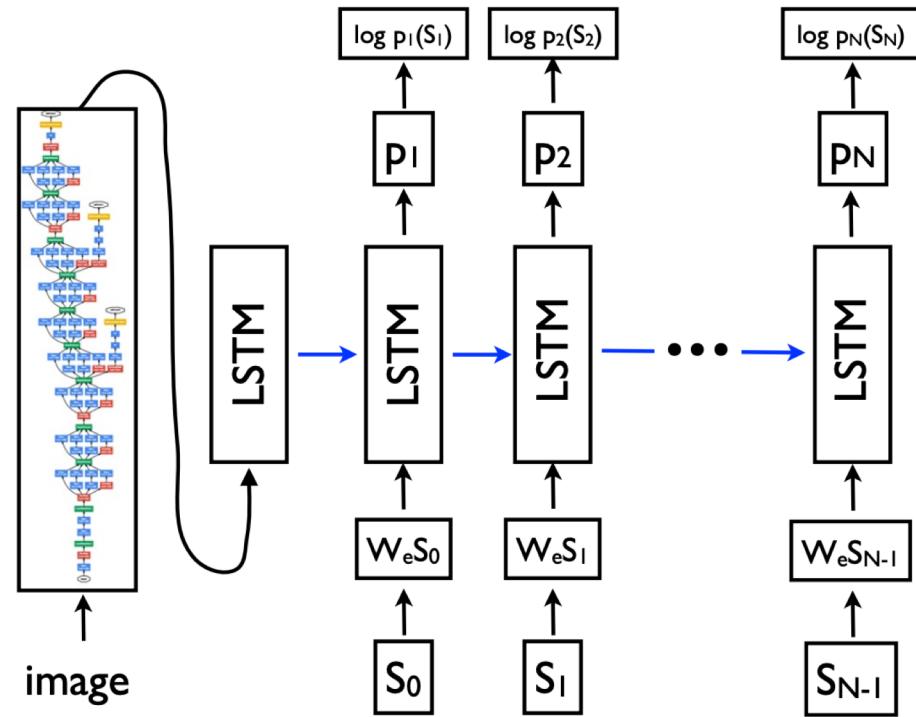
Classic Image caption generator (Vinyals et al, 2014.)

Also 2014 MSCOCO rank #1

$$x_{-I} = CNN(X)$$

$$x_t = W_e S^T, \quad t \in (0, \dots, N-1)$$

$$p_{t+1} = LSTM(x_t), \quad for t \in (0, \dots, N-1)$$



Attention added (Xu et el, 2015.)

$$\begin{pmatrix} \mathbf{i}_t \\ \mathbf{f}_t \\ \mathbf{o}_t \\ \mathbf{g}_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} T_{D+m+n, n} \begin{pmatrix} \mathbf{E} \mathbf{y}_{t-1} \\ \mathbf{h}_{t-1} \\ \hat{\mathbf{z}}_t \end{pmatrix}$$

$$e_{ti} = f_{\text{att}}(\mathbf{a}_i, \mathbf{h}_{t-1})$$

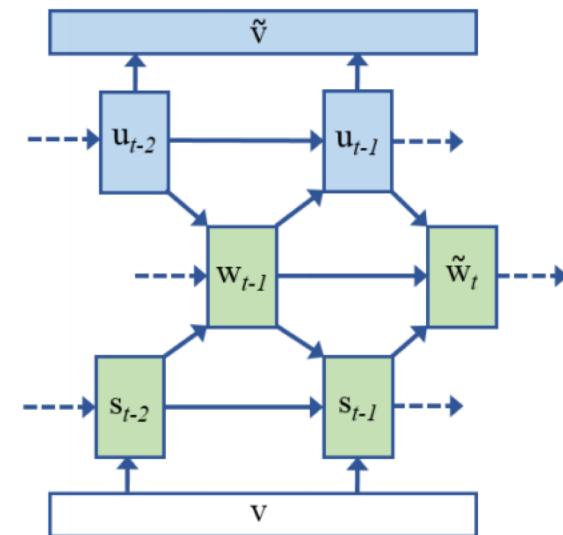
$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^L \exp(e_{tk})}$$

$$\hat{\mathbf{z}}_t = \phi(\{\mathbf{a}_i\}, \{\alpha_i\})$$

- ▶ Attention weight is calculated from decoder hidden state \mathbf{h} and encoder features .
- ▶ the context vector \mathbf{z} is calculated from the attention.
- ▶ Output and new hidden state is generated from the context vector \mathbf{z} and hidden state.

Input higher level of image features (Chen et el, 2015.)

- ▶ Reconstruct visual representation
- ▶ Stronger connection between images and texts



AI嘗試寫小說

- ▶ 星新一賞入圍
- ▶ AI 小說生成 “電腦寫小說的一天”
- ▶ 仍有大量成份為人工編寫而成

最初の 1 バイトを書き込んだ。

0

その後ろに、もう 6 バイト書き込んだ。

0, 1, 1

もう、止まらない。

0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, 610, 987, 1597, 2584, 4181, 6765, 10946, 17711, 28657, 46368, 75025, 121393, 196418, 317811, 514229, 832040, 1346269, 2178309, 3524578, 5702887, 9227465, 14930352, 24157817, 39088169, 63245986, 102334155, 165580141, 267914296, 433494437, 701408733, 1134903170, 1836311903, 2971215073, 4807526976, 7778742049, 12586269025, ...

私は、夢中になって書き続けた。

その日は、雲が低く垂れ込めた、どんよりとした日だった。

自動詩歌生成？

- ▶ 生成
- ▶ 生成後的評價
- ▶ 如何結合評價來輔助生成

自動詩歌生成

最早：1950 Stochastische Texte 隨機的字詞組合

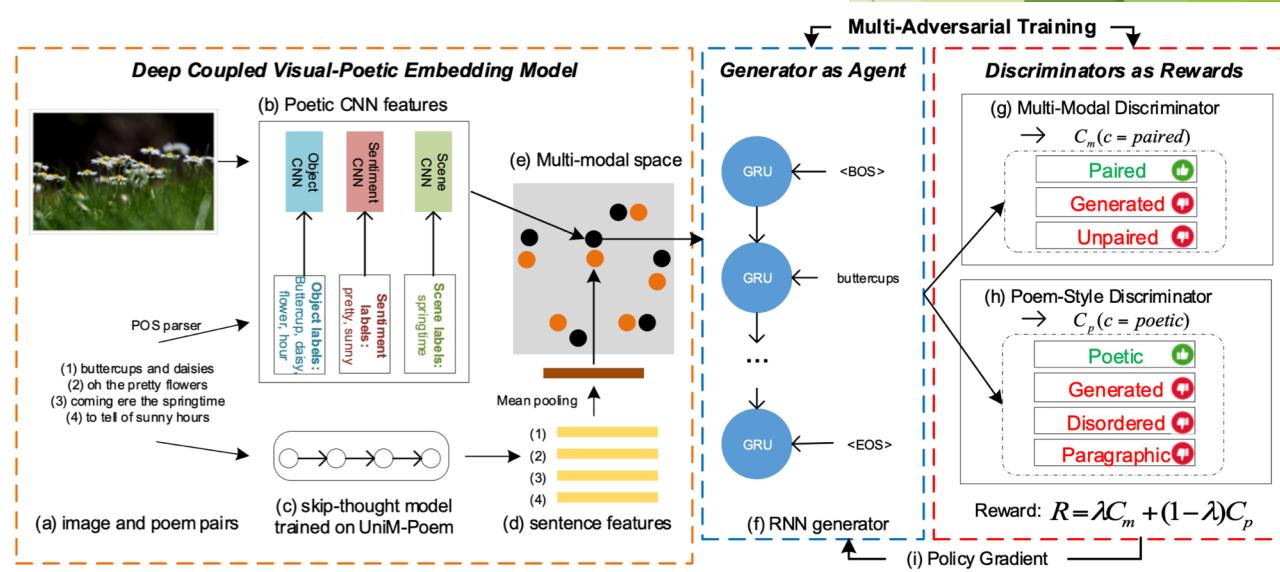
主要的兩種分類：

1. 基於模板填入適當的字句
2. 基於統計來重組新的字句

$$P(W_t | W_{t-1:1}) = P(w_1) * P(w_2 | w_1) * P(w_3 | w_1w_2) \dots$$

相似度的計算？計算概率的模型？

在深度學習基礎上，更加結合圖片與文字的關聯性
(Liu et al. 2018)



自動詩歌生成的評價

- ▶ 詩歌生成的ground truth?
- ▶ 正確性、流暢度
- ▶ 評價對象、主觀性導致的不一致
- ▶ 大量的評價需求
- ▶ Perplexity, BLEU score 在自動訓練/評價模型上的廣泛運用以及缺失

看圖寫詩

- ▶ 為什麼要寫詩
- ▶ 寫詩的原理
- ▶ 開發過程



圖源：<https://poem.msxiaobing.com/>

寫詩的起源

做一個有個性的AI

寫散文還是寫詩？

倪匡小說？

如何寫詩 - 例子

- ▶ 像每一座城市愧對鄉村
我才有一個美好的完成
每個失眠的夜晚
我是一個花言巧語的人
隱匿在靈魂最迷失的火
- ▶ 繞出城市的邊緣
美好的
在風裡
最輕微的觸動



► 在花里表演
揚起明亮的羽毛
靜靜地翕動著
所有生命的期限



A I 寫詩與徐志摩

- ▶ 她嫁了人間許多的顏色
那繁星閃爍的幾天蒼色
那滿心的紅日
看萬里天使在世界
我就像夢
看那星閃爍的幾顆星
西山上的太陽
青蛙兒正在遠遠的淺水
她嫁了人間許多的顏色
- ▶ “她嫁了”來自徐志摩翻譯的一首詩
“渦堤孩新婚歌”（節選）：
小溪兒笑呷呷的跳入了河，
鬧嚷嚷的合唱一曲新婚歌，
“開門，水晶的龍官，
渦堤孩已經成功，
她嫁了一個美麗的丈夫，

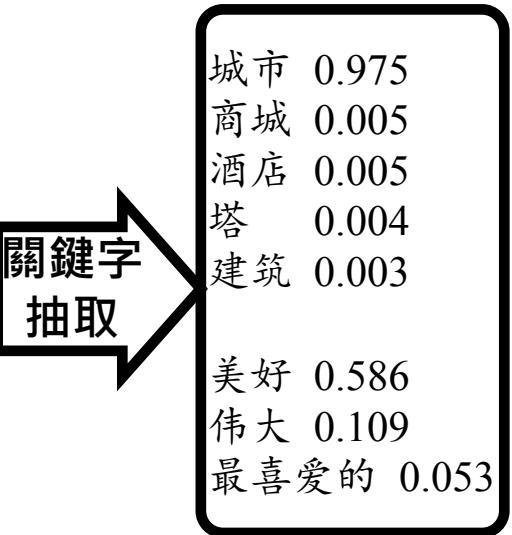
Case Study: 如何學看圖寫詩?

- ▶ 訓練模型(黑盒子): 透過不斷讀取(閱讀)大量的詩詞來學習詩的生成方法
- ▶ 透過519位詩人 1920~1980年代約90000句現代詩的訓練達成 (最初版)
- ▶ 生成: 透過圖片得到寫詩的起點，再透過之前閱讀大量詩詞的訓練(模型黑盒子)來生成針對圖片的詩句。

嘗試以人類的思考方式觀察看圖寫詩

- ▶ 從一個主題開始（圖片）
- ▶ 從主題推敲靈感（關鍵字）
- ▶ 從靈感開始一句詩句
- ▶ 利用以往的閱讀經驗來完成字詞的組合，同時保持句意的連貫性
- ▶ 用自動評價的方式反覆琢磨、修改詩句

一首詩生成的過程



關鍵字過濾
+ 關鍵字擴張

城市
美好
人最

遞迴生成
重新生成
通順評價
評價門檻是否通過

像每一座城市愧对乡村
我才有个美好的完成
每个失眠的夜晚我是个花言巧语的人
隐匿在灵魂最迷失的火

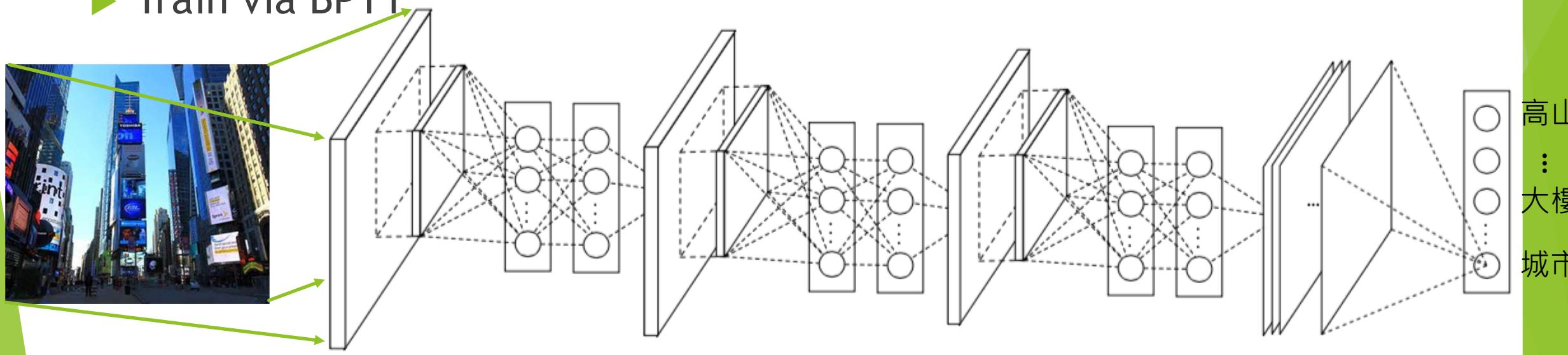
關鍵字的生成

詩的生成

1. 圖片辨識，並且擴增關鍵字庫。
2. 透過與原詩訓練集的契合度、挑選適合寫詩的字詞
3. 接著透過循環類神經網路語言模型以及後製自動評測來完成一句一句詩句。

生成第一步：辨識圖片

- ▶ 從圖片辨識出物體、意象
- ▶ 透過卷積神經網路來達成 (pretrained on Krizhevsky et al, 2012.)
- ▶ Input: image pixels, output: label Noun or Adj; two CNN models.
- ▶ Generate $P(C | I) = f(W_c * I)$, W_c : CNN's parameters, I : input image,
*: operations of convolution, pooling, activation.
- ▶ Train via BPTT



生成第二步 預處理圖片關鍵字

城市、酒店、
鄉城；美好、
偉大、最喜
愛的



原詩統計：

城市: 174 出處
酒店: 4出處
鄉城: 0出處
美好: 41 出處
偉大: 37出處
最喜愛的: 0出處

一個**城市**有一個人
這個**城市的風**把她吹得更**卷**
站在**城市的**樓頂
他身穿微服獨自一人出現在**城市**
彷彿他們才是**城市的**主人
...

[城市、美好]



城市、美好、人、最 就成為生詩的四詞組

挑選原詩出現頻率
高的名詞與形容詞
→比較好生詩

對照原詩中包含
關鍵詞的詩句

尋找其他的
名詞、形容詞、
副詞 by similarity
of word2vec

生成第三步:前後生成與自動評價

- ▶ 採用前後遞迴生成的方式 (recursive generation) · 用正向與反向生成模型 (charRNN based)從圖片關鍵字前後生成出接續的詞句
- ▶ Ex. 關鍵字: 城市

第一次前後生成: 出←城市→的

第二次前後生成: 繞←出城市的→邊

第三次前後生成:[句首]←繞出城市的邊→緣

第四次前後生成 X 繞出城市的邊緣→[句尾] 生成完成

如何決定一次生成接下去的字? --語言模型

生成第三步使用的語言模型

- ▶ 模型基於機率選擇—骰骰子
- ▶ 在給定了 “出城市的” ，模型學習過關於“的” 後面接續的搭配:
- ▶ 每一種搭配屬於 n-gram 的表達. E.g 4-gram:
- ▶ 由連續 n 個字組成的詞組: 4-gram:城市的x
- ▶ 最後詞組搭配的選擇是按照機率來挑選
- ▶ 城市的+ 邊 =城市的邊 選擇這個結果的機率有30%
- ▶ 每次都隨機挑選→同樣關鍵字也有可能有不同結果(with beam search)



如何訓練寫詩模型

- ▶ 語言模型: 紿定一些前述的字句，根據機率預測下一個字/詞/句子。
- ▶ 經典的語言模型: n-gram 語言模型訓練
訓練資料: 以 bi-gram 語言模型為例，每個詞組以連續2字表達

“已如空殼在慢慢飄出城市的邊緣變為晨霧，
我的敵手往城市上。”

第一句的訓練讀到 **城市、市的、的邊、邊緣**:

這幾組bigram，模型變化: 對於給定“城市”，預測“的”的機率提高了，而其他bigram與城市的配對機率則下調

第二句的訓練讀到了 **城市、市上**:

對於給定“城市” 預測“上”的機率提升了，其他組合的機率則會下調。

語言模型的訓練 --- 以原詩為例

- ▶ 在詩人們的原詩中，關於各種詞語的使用、詞語連接、句法結構。

藏匿於黑黝黝的叢林 邊遮掩掩

隱匿在衣領和眼睛後面

像隱匿在林中野貓的眼睛在閃爍

- ▶ 機器會學會“隱”+“匿”的機率大於“藏”+“匿”，

同樣的 $P(\text{隱匿在}) > P(\text{隱匿於})$

模型也學會這些字詞共同出現的規律

i.e [眼睛、閃爍], [匿、眼睛]

透過語言模型的概念來訓練機器，機器可以學會字詞的正確搭配



搭配輸入大量文本的詩句來達成

如何訓練寫詩模型(黑盒子)?

- ▶ N-gram的缺點:
 - ▶ 電腦的記憶體不能承受：只能計算n元語法詞組的機率，若n越來越大導致組合過多
 - ▶ 同義字詞的關係也沒辦法學習(詞向量關係)
 - ▶ 資料的分布稀疏也會造成學習上的困難
- ▶ AI的語言模型: 使用的是深度學習的類神經網路RNN based模型
- ▶ RNN 語言模型: 能夠紀錄並更新長期記憶，能使單句句意更連貫

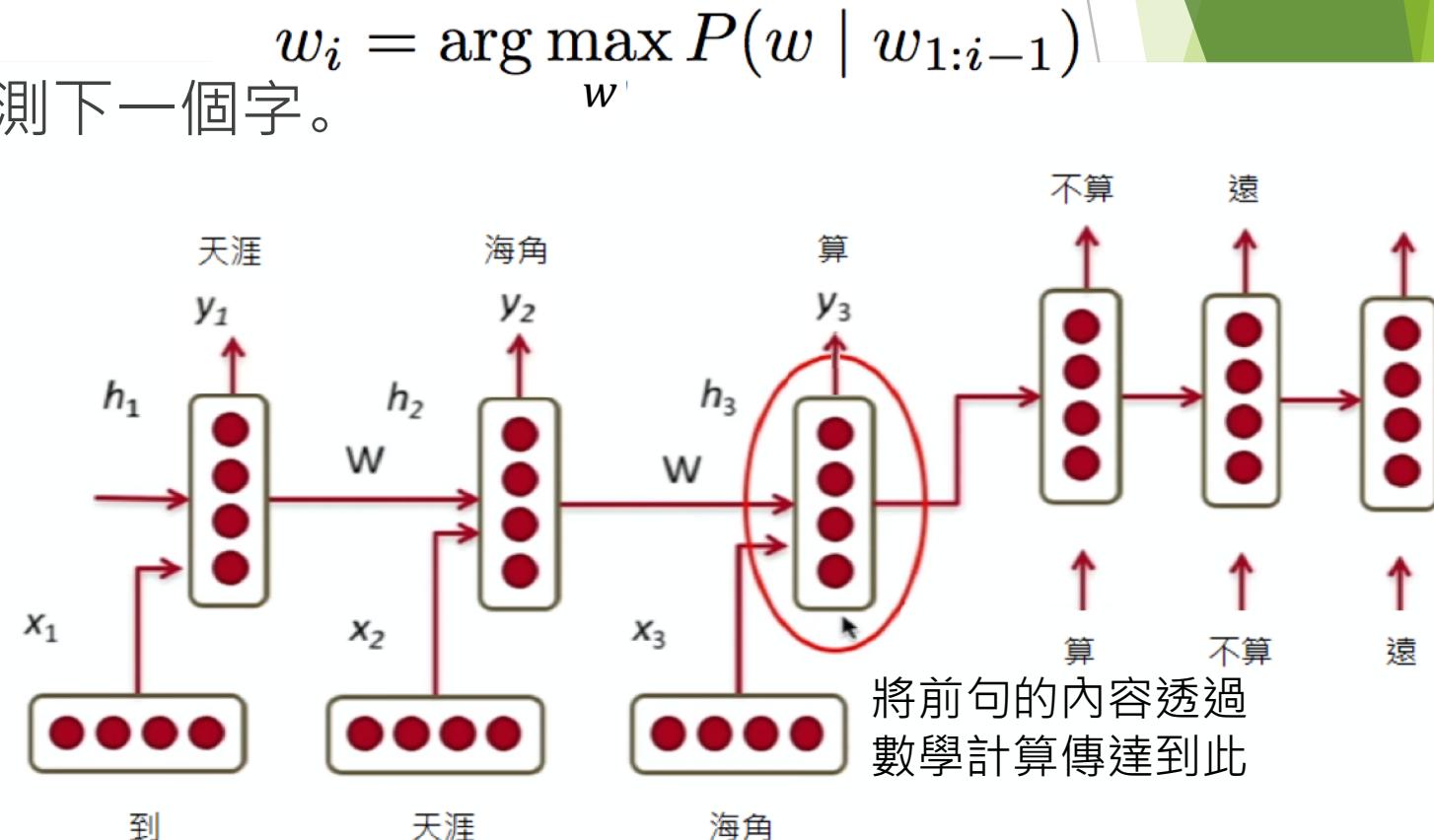
舊城市和不同國地

有如沙粒的城市

城市愧對鄉村

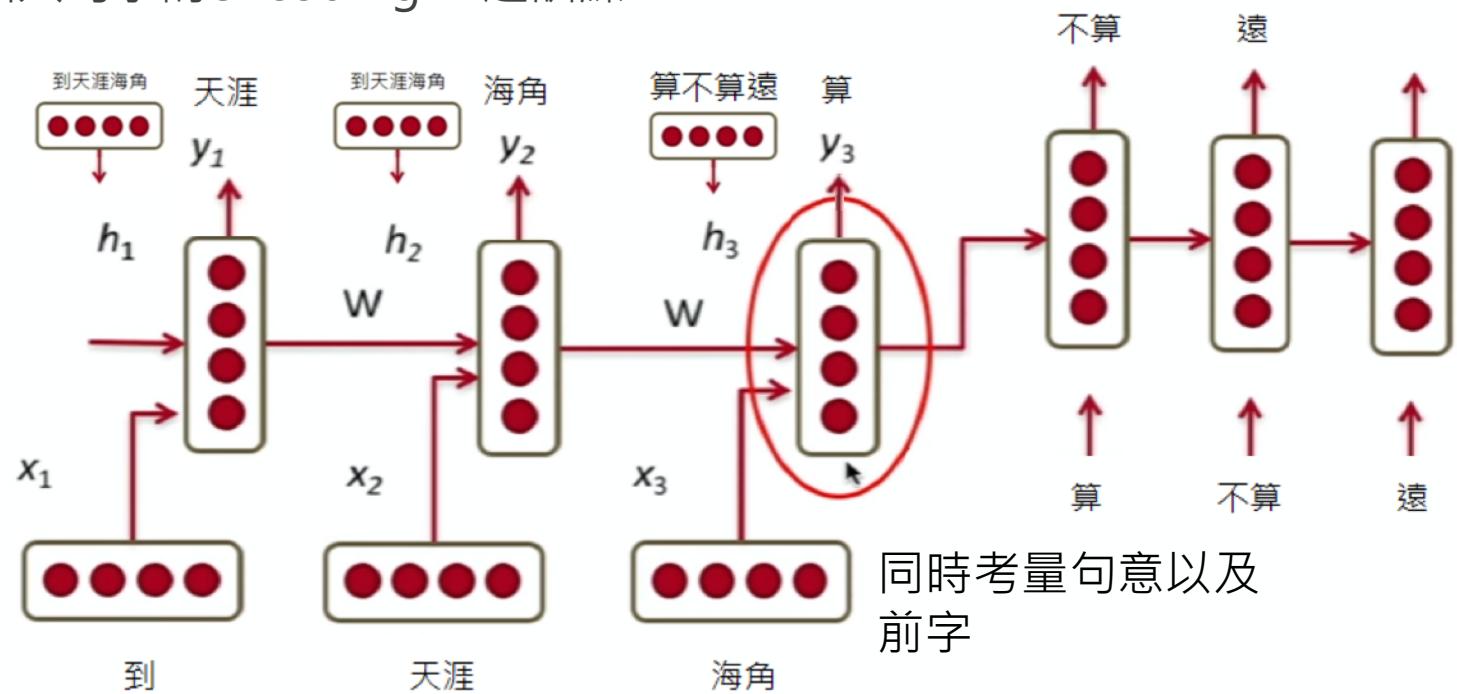
語言模型的訓練 (charRNN的訓練)

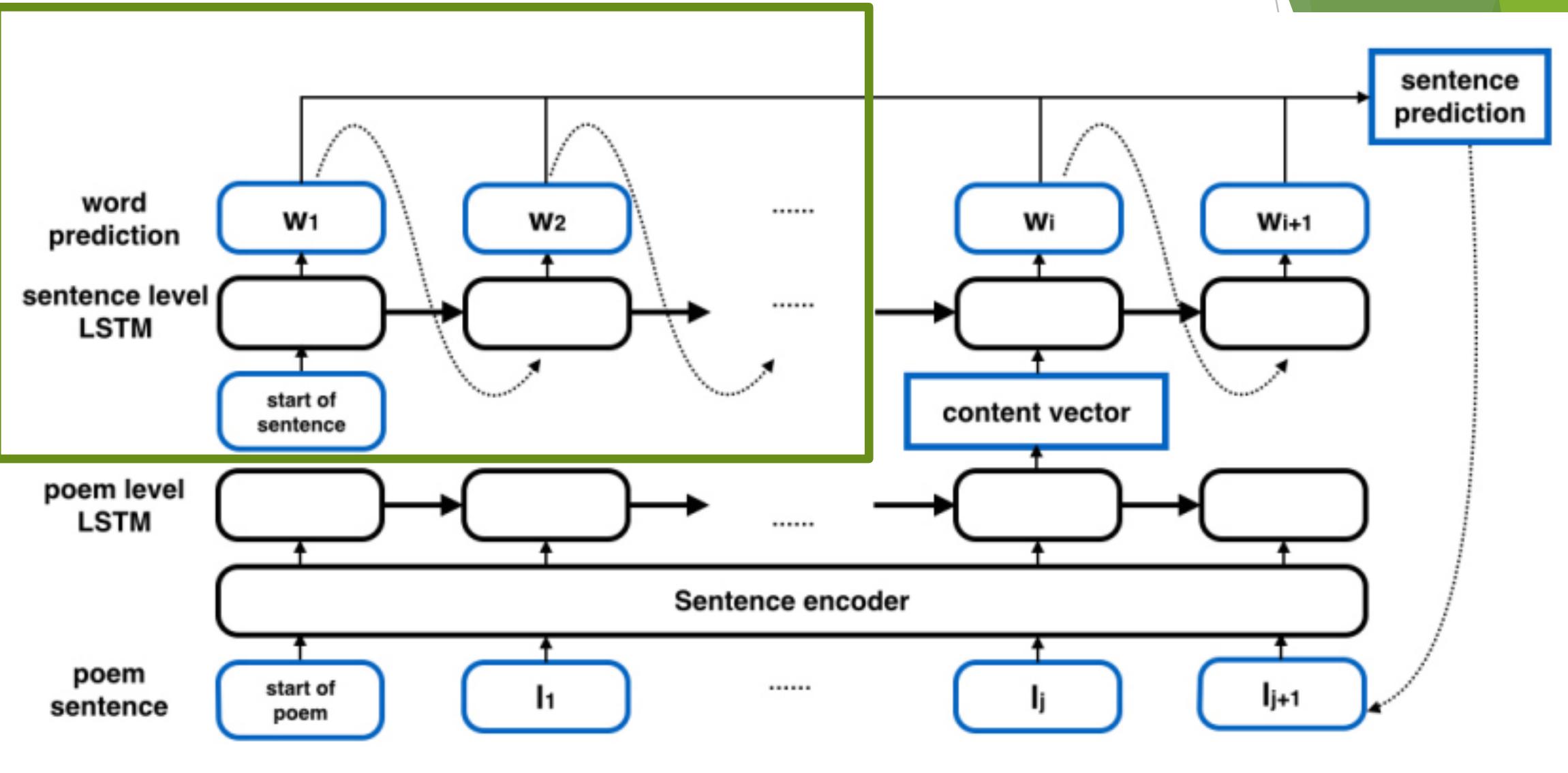
- ▶ 文字被編碼成一些固定編號 (方便計算)
- ▶ 透過有時間序列的輸入，模型可以按照序列將一個一個文字做數學運算
- ▶ 透過將前字作編碼輸入，預測下一個字。
- ▶ 機器在訓練
三毛的《遠方》
- ▶ 迭代訓練



句意層級的CharRNN訓練

- ▶ 在訓練完文字級別的RNN model之後，使用類似的架構訓練考量句意層級的CharRNN model
- ▶ 在原先的時間粒度為文字的模型，加入句子的encoding一起訓練
- ▶ 同時再訓練正向與反向生成的模型
(將文本倒著讀)





生成第三步-自動評價

- ▶ 評價詩歌生成的困難點
- ▶ 因為隨機生成，關於“城市”，模型生成了許多不同的詩句

塞上城市的縫隙

兩個城市的邊緣

哈上城市的風暴遮沒

肉上的手掌殘留的城市與神

拉上城市的沙

舊城市和不同國地

有如沙粒的城市

城市愧對鄉村

城市的邊緣 虹影《輪盤賭》林莽

城市的砂暴遮沒 徐江《悲憫》

城市愧對鄉村 楊鍵《慚愧》

自動評價

- ▶ 根據句子的流暢度、詞性做評分並套用到生成中作為分數門檻。
- ▶ 使用額外的**語言模型**為句子計算分數來評判流暢度:比較常見、常使用的詞句、順序、組合會比較高分。
- ▶ 同時也為句子在詞性上的評分→將詞性視為字訓練的語言模型
- ▶ 完整/流暢的主謂關係句子→學習正確的主謂賓關係
- ▶ 做反向句子的評分，將句子倒著讀取計算整句的機率，在對句尾的修正有很大的幫助。

- ▶ 用詞是否足夠獨特？
- ▶ 使用的靈感是否容易有多樣的結果？

流暢度與正確性的自動評價

- ▶ 用機率表達寫好的詩句是有多符合一般文體的字詞組合邏輯
- ▶ 利用 n-gram 以及 skip n-gram

比較

“兩個城市的邊緣” >> 兩個城市 個城市的 城市的邊 市的邊緣...等4-gram組合

“有如沙粒的城市” >> 有如沙粒 如沙粒的 沙粒的城 粒的城市

“拉上城市的沙” >> 拉上城市 上城市的 城市的沙

“兩個城市的邊緣” >> 兩個-城市 兩個-市的 城市-邊緣 ...等skip bi-gram組合

“有如沙粒的城市” >> 有如-沙粒 有如-城市 又....

給定一個 Θ 機率模型 (閱讀大量的額外文本的N-gram LM)

$P(\text{兩個城市的邊緣}|\Theta) > P(\text{有如沙粒的城市}|\Theta) > P(\text{拉上城市的沙}|\Theta)$

來自動評斷文句的用字正確性

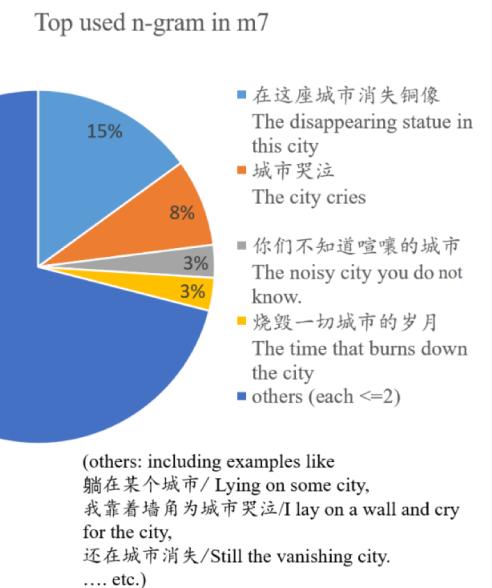
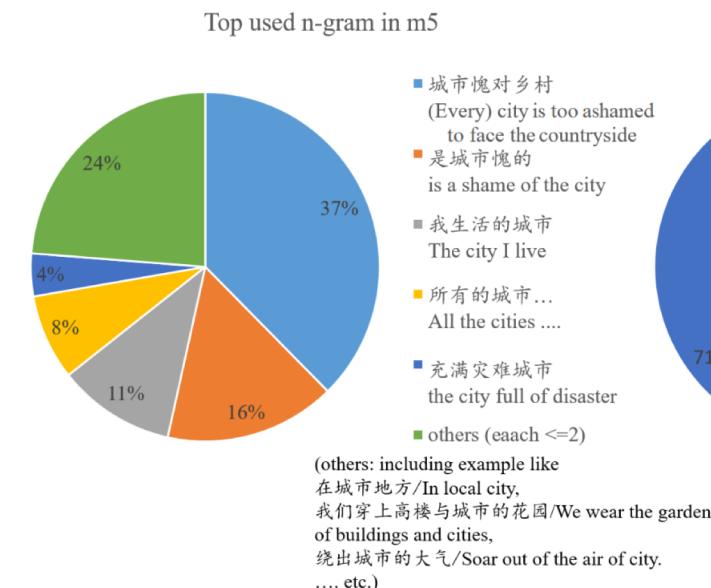
詞性與反向結尾的自動評價

- ▶ 同樣的利用N-gram模型與RNN模型（但訓練在詞性標注的文本上）對詩句的詞性正確性與流暢度做自動評價
- ▶ 反向結尾的自動評價
針對所有詩的最後一句做特別的訓練：反向閱讀。
- ▶ 提升結尾的完整性

重複性高或原創詩詞的評價方法

- ▶ 透過一個簡單的binary classifier，判別詩詞的重複度與原創度（由人類標註的900句詩詞，針對用詞、句法是否獨特做的標註），測試準確率為78%
- ▶ 同時將詩詞生成結果透過網際網路搜尋，尋找高相似的內容。
 - ▶ 相似結果但類型卻很廣泛→並非原創的詞句，乃詩詞創作常見的用語
 - ▶ 反之，相似、結果卻特別指向某些詩人作品→詩句為該詩人原創的可能性高，則應避免使用。

- ▶ 透過這兩個後製篩選，至少在保證 5字詞以上的詩句是有原創度的。
- ▶ 與原詩相比的創新性？
與生成結果相比的多樣性？



第四步：接續詩句的生成

- ▶ 將上一句的資訊轉化成編碼傳給下一句
- ▶ 重複第2-3步來完成後續的每句詩詞。
 - ▶ 句子層級訓練的語言模型：將結果轉達成句意表達，之後的生成中會參考這些特徵來生成。
 - ▶ 句跟句之間的連貫性
- ▶ 接續生成搭配的關鍵字也直覺地影響多樣性

Diversity of the 2 nd Sentence from	(a) High diversity (1 st keyword)	(b) Medium diversity (1 st keyword)	(c) Low diversity (1 st keyword)
(I) High diversity keyword	0.321	0.329	0.299
(II) Medium diversity keyword	0.327	0.290	0.259
(III) Low diversity keyword	0.232	0.193	0.205

人工評價

▶ 評價的目的

- ▶ 圖像的關聯性
- ▶ 可讀
- ▶ 語意連貫
- ▶ 想像力
- ▶ 感染力

▶ 絕對與相對評價

▶ General guideline :

用詞可讀為基本滿足標準

基本之上如果還可以滿足更高標準的可以酌情加分
如果句子有錯誤影響理解，可以酌情減分。

1為與圖無關

▶ 圖片與評價的關聯性

評價一致性的提升

▶ A/B test, latin square 的方式分配任務



城市 美好 象 地

像每一座城市愧对乡村
才有一个美好的完成
霎时逃出我们的想象
地球我的母亲

城市 美好 风 象

城市愧对乡村
比一比美好的心
浓缩了一种风景
象随风晃动

城市 美好 风 最

绕出城市的边缘
美好的
在风里
最轻微的触动

城市 美好 地 人

这座城市愧对乡村
想着美好的气味
朋友在我住的地方
这个迷路的人

请根据请根据诗句是否通顺，诗句是否连贯，是否与图片相关，诗歌的意境等为诗歌评分：（5分最高，1分最低）

5 (喜欢)

4

3

2

1 (不喜欢)

诗句与图片不符
 关键字与图片不符

5 (喜欢)

4

3

2

1 (不喜欢)

诗句与图片不符
 关键字与图片不符

5 (喜欢)

4

3

2

1 (不喜欢)

诗句与图片不符
 关键字与图片不符

5 (喜欢)

4

3

2

1 (不喜欢)

诗句与图片不符
 关键字与图片不符

寫詩的缺陷 (2017)

- ▶ 慣用用詞:沙灘、靈魂
- ▶ 不能控制較多變化:關鍵字生成
(沒有控制主題、情緒生成等等)
- ▶ 特定關鍵字生成困難: 電腦
(只能透過關鍵字擴張來修正)
- ▶ 圖片辨識錯誤: 水(螢幕的藍色)
- ▶ 圖片不能辨識出”動作”



太空梭、潮、水、宇宙 (未修正)
這個空間裡經向大地沉醉太空梭
人生在每一瞬間達到高潮
耳聽屋簷滴水的聲音
朝向宇宙點燃

未來可以如何提升寫詩的質量

- ▶ 更為強大的模型基礎 (用transformer 取代 RNN)
- ▶ 結合強化學習的概念加入人工評價的結果
- ▶ 結合深度學習在圖片與文字的學習
- ▶ 更大量的詩句語料庫

Two sentences to conclude

- ▶ AI離不開人類的養料，AI是一種傳承，以一種新鮮的方式去呈現
- ▶ 人工智能的侷限

We are hiring !

- ▶ Data Scientist
- ▶ Data Engineer
- ▶ Data Analyst
- ▶ NLP Engineer

Thank you

QA

johnson.wu@linecorp.com