組長 GitHub ID：ming024
組員
- r08922080　資工所碩一　簡仲明
- r08921062　電機所碩一　黃健祐
- b04501127　土木四　凌于凱

5-1

Configuration

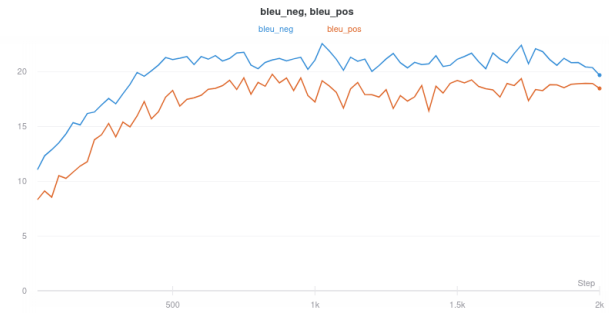
我們將 slf factor 設為 0.4，cyc factor 設為 0.7，其他都使用預設參數。

Loss Plot

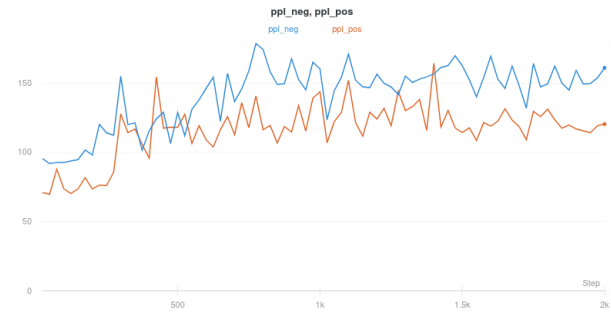
Loss 的變化如 Figure 1a 所示，基本上模型的訓練大致穩定。

(a) Loss

(b) Accuracy



(c) BLEU score



(d) Perplexity

Figure 1: Some important information about the style transformer.

Metrics

就 Figure 1 而言，accuracy 以及 BLEU score 在訓練後期成長逐漸停滯 (Figure 1b 及 Figure 1c)，而 perplexity 則相對沒有太大的起伏 (Figure 1d)。最終的數據則如 Table 1

所呈現。

Type	Accuracy	Ref-BLEU	Perplexity
Positive	0.922	18.489	120.47
Negative	0.874	19.702	161.11
Average	0.898	19.095	140.79

Table 1: Statistics about the style transformer and its outputs.

Examples

結果如 Table 2 及 Table 3 所示。如果有一些強烈肯定或是否定的詞彙，如 bad、great、terrible、excellent，都能夠轉換得很好，但如果有兩個以上這類的詞則可能會同時轉換，造成雙重否定，語意沒有改變。此外，如果沒有強烈立場的詞彙的話，model 有可能就只是單純將正向的詞轉成負面，但意思沒有什麼關係，也有可能就直接輸出一模一樣的語句。

Good Results	
[gold]	super good deals and very friendly staff .
[raw]	super good deals and very friendly staff .
[rev]	super bad deals and very rude staff .
[gold]	a great place to watch sports !
[raw]	a great place to watch sports !
[rev]	a terrible place to watch sports .
Problematic Results	
[gold]	loved the menu and the drinks .
[raw]	loved the menu and the drinks .
[rev]	regret the menu were the drinks .
[gold]	she was so patient , kind and understanding .
[raw]	she was so patient , kind and understanding .
[rev]	she was so overrated , plain and understanding .

Table 2: Examples of style transfer from **positive** sentences to **negative** ones given by style transformer.

Good Results	
[gold]	there chips are ok , but their salsa is really bland .
[raw]	there chips are ok , but their salsa is really bland .
[rev]	there chips are ok , but their salsa is really excellent .
[gold]	the sales people here are terrible .
[raw]	the sales people here are terrible .
[rev]	the sales people here are great .
Problematic Results	
[gold]	the wine was very average and the food was even less .
[raw]	the wine was very average and the food was even less .
[rev]	the wine is very efficient and the food was most than .
[gold]	she said she 'd be back and disappeared for a few minutes .
[raw]	she said she 'd be back and disappeared for a few minutes .
[rev]	she said she 'd be back and disappeared for a few !
[gold]	there is definitely not enough room in that part of the venue .
[raw]	there is definitely not enough room in that part of the venue .
[rev]	there is definitely not enough room in that part of the venue .

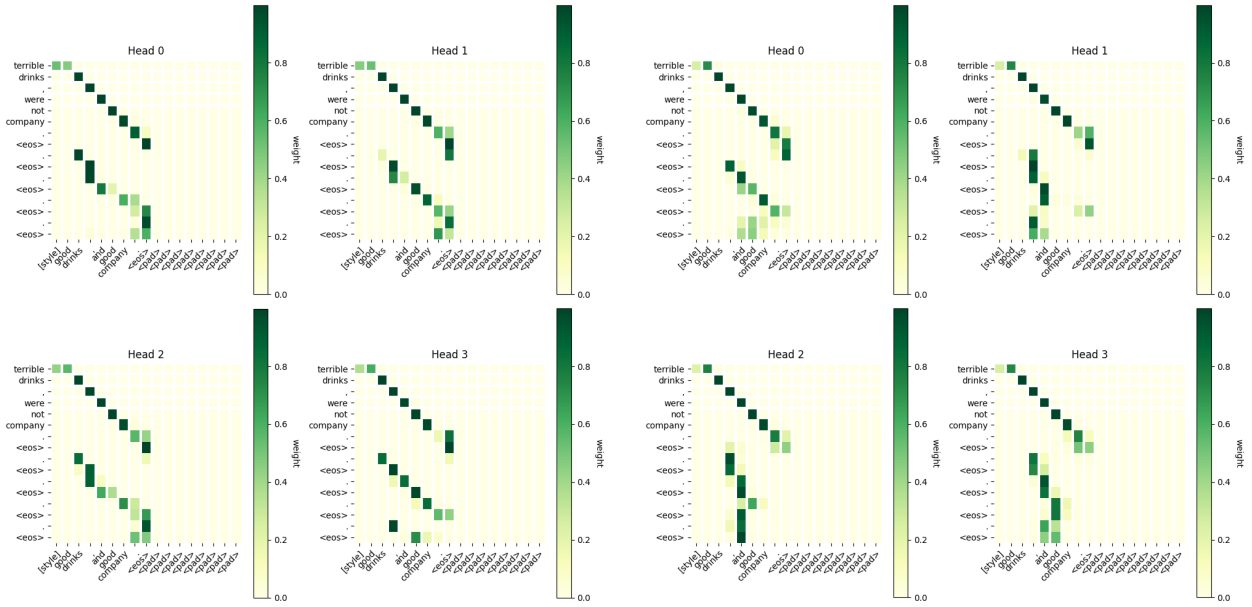
Table 3: Examples of style transfer from **negative** sentences to **positive** ones given by style transformer.

5-2

1. Visualize the attention weights on memory while decoding (for each head and layer). The style token seems not being attended to while decoding, but the style is actually being transferred. Do you think it's reasonable? Why or why not?

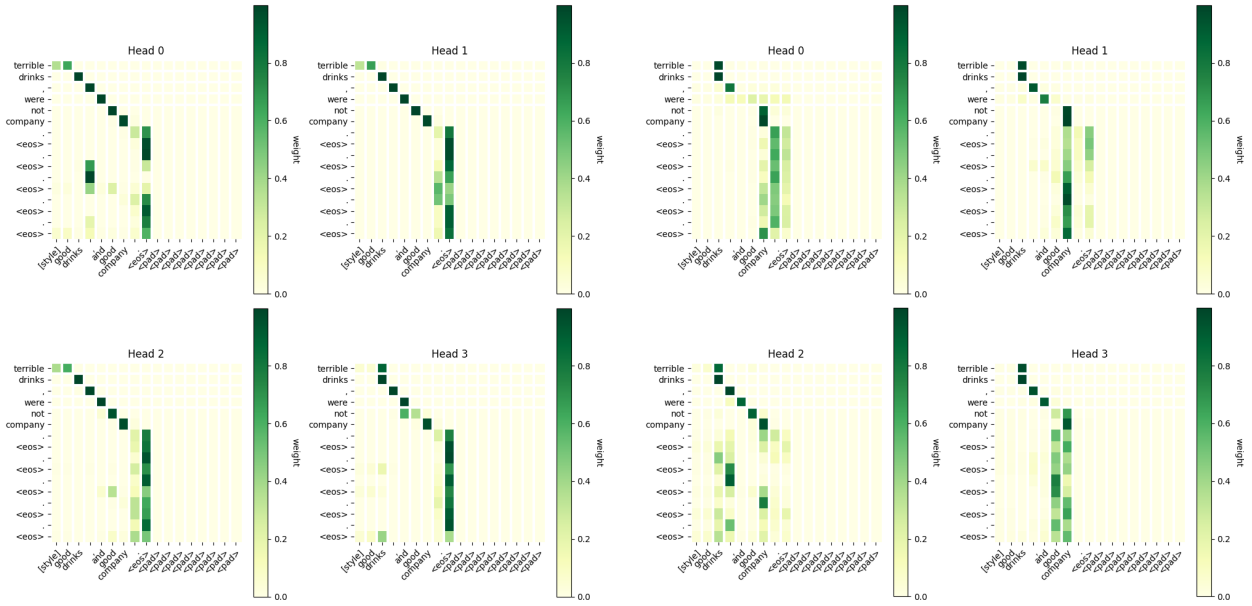
Type	**Text**
[GOLD]	good drinks , and good company .
[REV]	terrible drinks , were not company .

Table 4: Text used in Figure 2.



(a) Layer 0

(b) Layer 1



(c) Layer 2

(d) Layer 3

Figure 2: Attention weights on memory while decoding for each head and layer.

模型各層的 attention weights 如 Figure 2 所示。Style token 在 decoding 階段時，並沒有一直被 attended，我認為這是可以有合理解釋的。首先，transformer 在 decode 時為 auto-regressive 的，因此就算只有前面幾個 token 有 attend 到 style token，後面的 output 還是很有機會被 style token 所影響。其次，這個 model 在訓練時並沒有做

disentanglement，因此 token 的 embedding 很有可能就帶有 style 的資訊，因此就算 decode 時不全部依靠 style token，model 仍然能夠產生出相應的語句。

2. Transfer the style of some sentences sampled from the dataset. Collect the embedding of CLS token and use t-SNE to visualize the distributions. Does the result look reasonable?

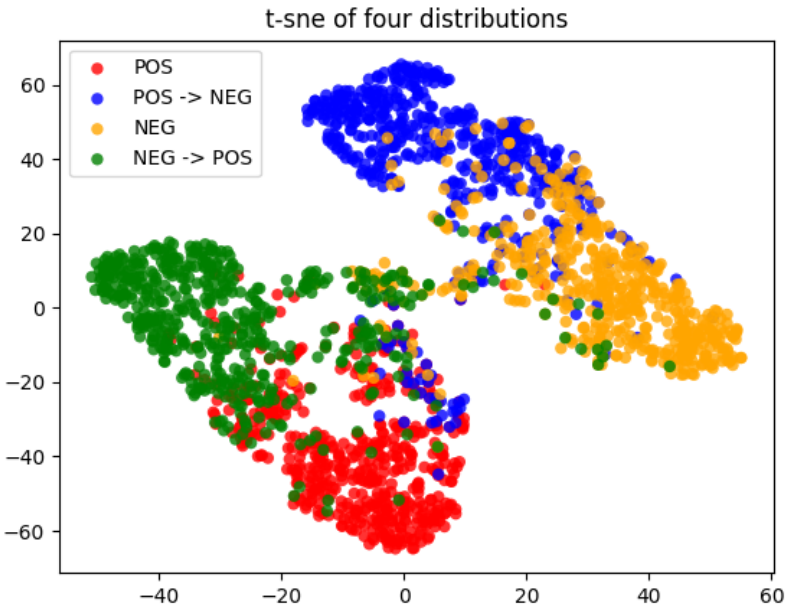


Figure 3: Visualization of embeddings using t-SNE.

降維後的結果如 Figure 3 所示，可以看到 positive 及 negative 的 embeddings 明顯分成兩群；雖然 real data 與 style-transferred data 的 embeddings 還是看得出來有一點區別，但經過 style transfer 的 embeddings 是有和相對應的 style token embeddings 靠在一起的，如 POS 和 NEG → POS，因此我認為這樣的結果還算合理。

3. Perform style transfer with one position masked each time. Show the result and your discovery.

Type	Text
[ORG]	they have delicious soups everyday .
[REV]	they have bland soups everyday .
[ORG]	<unk> have delicious soups everyday .
[REV]	they have bland soups everyday .
[ORG]	they <unk> delicious soups everyday .
[REV]	they were bland soups everyday .
[ORG]	they have <unk> soups everyday .
[REV]	they have no soups everyday .
[ORG]	they have delicious <unk> everyday .
[REV]	they have dry and everyday .
[ORG]	they have delicious soups <unk> .
[REV]	they have bland soups too .

Table 5: Examples of style transfer when a token is randomly masked at each time.

由 Table 5 可以發現，隨機 mask 掉任何一個 token 並不會有太大的影響，大部分的結果都能夠將 delicious 轉成 bland。當一個 token 被 mask 時，model 會傾向於補上相同或類似性質的詞，如當 they 被 mask 時，model 正確地補上了合理的主詞；而當 delicious 被 mask 時，model 雖然不知道究竟是怎麼樣的 soups，但推斷 no soups 應該是比 have 某種 soups 還要負面。不過當 soups 被 mask 時，model 輸出的結果比較奇怪，語句不太通順。雖然都是名詞，但 they 跟 soups 被 mask 的結果差異頗大，我推斷這是因為前者可以透過 have 來推斷主詞應該是人 (I、you、we 等等)，但食物相較於前者有很多種，所以光是以 delicious 難以推斷出到底是什麼東西好吃，model 只好輸出相對於 delicious 比較負面的 dry (食物如果乾乾的應該比較不好吃)。由上述結果也可以發現，對 model 而言，名詞及動詞的重要性應該比形容詞及副詞等重要。

5-3

Model Configuration

這題我們選擇實作了 NeurIPS 2019 的 paper - **Controllable Unsupervised Text Attribute Transfer via Editing Entangled Latent Representation**。這個 model 的特色是在 training 的 reconstruction 時並不使用到 positive/negative 的 label，而是直接用 teacher forcing 的方法來做 self-reconstruction，完全不做 cycle-reconstruction，也就是變成一個單純的 autoencoder。

當然，為了使 encoder 能夠學會將 positive 以及 negative sentence 的 embedding 平均分布在 latent space 中，還是需要在 z 空間中學一個 discriminator，這也是 label 在 training 過程

中唯一會被用到的地方。經過這樣的訓練後，這個 model 被期望能夠擁有將任何句子打到一個 latent space 中，並且 reconstruct 回來的能力，並且不管是 positive 或是 negative sentence，被打到這個空間中後，就會難以分辨原本的 label，因此句子在這個空間空的 latent embedding 幾乎不帶任何會含有 positive/negative 資訊的東西。

接著第二階段中，若餵給這個 model 一句 positive sentence，要求他轉為 negative，則首先會和 training 時一樣，將句子轉為一個不帶有 label 資訊的 z 向量。接著 model 會在 latent space 中進行某種操作，將 z 轉為在同一個空間中的 z'，而這個 z' 被相信更像是一個 negative sentence 的 embedding。最後只要從 z' 通過 decoder 還原，就可以得到一句 negative sentence。

而要如何得到 z' 呢？這篇 paper 的作法是使用類似 gradient ascent 的方法，將 discriminator 的 loss 一路傳回到 z，並對 z 做數次 gradient ascent 得到 z'，因此這個 z' 在 discriminator 眼中更像是一個 negative sentence 的 embedding。從 negative 句子轉為 positive 也是完全一樣的做法，只是所有 label 相反。

最後，因為完全沒有任何 disentangled latent distribution，這個 model 明顯是一個 entangled 的 model。

Loss Plot

這個 model 在訓練時只有 reconstruction loss 跟 discriminator loss 兩個 loss，整體相當穩定，如 Figure 4a 所示。

Metrics

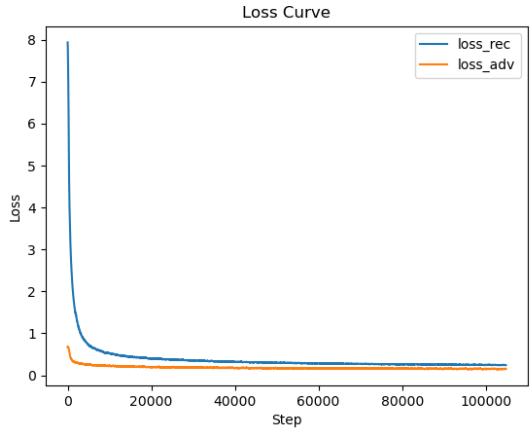
在沒有精調參數的情況下，這個 model 的 metrics 似乎比原本 style transformer 要差一些，但因為這個 model 沒有用到 transformer 架構，而只有用 recurrent 網路，所以也算是合理。雖然訓練了 30 個 epoch 以後 accuracy 看起來還在上升 (Figure 4b)，但是 perplexity 已經開始飆高 (Figure 4d)，ref-BLEU 也沒有明顯提升 (Figure 4c)，人工檢查文字結果甚至發現已經開始有點怪怪的，因此沒有繼續如 paper 內一樣訓練到 200 個 epoch。

Type	**Accuracy**	**Ref-BLEU**	**Perplexity**
Positive → Negative	0.798	9.220	185.76
Negative → Positive	0.580	14.035	142.69
Average	0.689	11.628	164.28

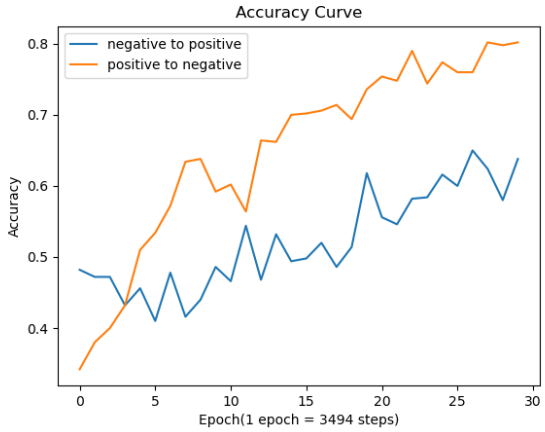
Table 6: Statistics about the style-transfer model and its outputs.

Examples

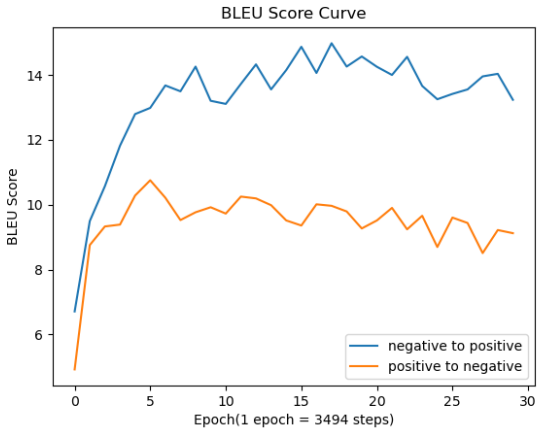
結果如 Table 7 及 Table 8 所示。整體來說，這個 model 有時候會直接複製整句話，但又在句尾生出一些無意義獲意義不明的後綴，可能是因為訓練 autoencoder 的時候的效果太強，所以 model 非常強烈傾向一定要把整句話都 reconstruct 出來，而不是根據文法去生成一個合理、合文法的語句。



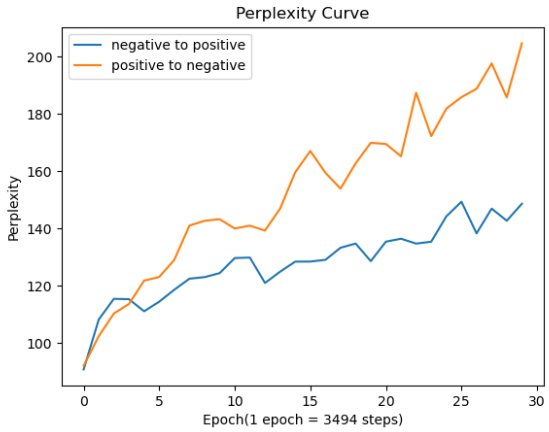
(a) Loss



(b) Accuracy



(c) BLEU score



(d) Perplexity

Figure 4: Some important information about the model.

Good Results	
Original	everyone that i spoke with was very helpful and kind .
Reconstructed	everyone that i spoke with was very helpful and kind kind .
Style-transferred	everyone that i spoke with was n't very helpful and kind .
Original	the biscuits and gravy were good .
Reconstructed	the biscuits and gravy were good especially my room are good .
Style-transferred	the biscuits and gravy were bad " we began for cry .
Problematic Results	
Original	this golf club is one of the best in my opinion .
Reconstructed	this golf club is one of the best in my opinion .
Style-transferred	this golf club is one of the worst in my room .
Original	excellent knowledge dentist and staff !
Reconstructed	excellent dentist and knowledgeable staff !
Style-transferred	poor blood dry address than i saw management and helpful !

Table 7: Examples of style transfer from **positive** sentences to **negative** ones.

Good Results	
Original	there chips are ok , but their salsa is really bland .
Reconstructed	there chips are ok , but their salsa is really bland .
Style-transferred	there chips are ok , but their salsa is really nice and delicious .
Original	blue cheese dressing was n't the best by any means .
Reconstructed	blue cheese dressing was n't the best by any means .
Style-transferred	blue cheese dressing was great by the best means any by .
Problematic Results	
Original	the wine was very average and the food was even less .
Reconstructed	the wine was very average and the food was even less .
Style-transferred	the wine was very delicious and the food was even less and always .
Original	moving past the shape , they were dry and truly tasteless .
Reconstructed	moving past the shape , they were dry and truly tasteless .
Style-transferred	moving past the shape , they are truly dry and tasteless as well .
Original	the burgers were over cooked to the point the meat was crunchy .
Reconstructed	the burgers were over cooked to the point the meat was crunchy .
Style-transferred	the burgers were over cooked to the point the meat was crunchy .

Table 8: Examples of style transfer from **negative** sentences to **positive** ones.