Hierarchical, Feature-Based Text Generation

Stanford CS224N Custom Project

Caitlin Hogan
Department of Computer Science
Stanford University
cahogan@stanford.edu

Abstract

We introduce the creation and use of full-text, distributed feature maps as the basis for the hierarchical generation of long-form text. The problem of story generation, a challenge that consists of generating narratively-coherent passages of text about a particular topic, can be described as one of the most difficult challenges currently posed in text generation, as stories require long-range dependencies, creativity, and a high-level plot. Previous efforts note that story generation often fails to meet one or more of these requirements; generated stories are frequently repetitive, and typically lack any kind of a broader arc. We find that the use of automatically-generated "emotion maps" as a basis for hierarchical generation achieve perplexity scores comparable to previous efforts, despite using a numerical input rather than a textual one. Additionally, we introduce a new story generation dataset, consisting of 100,000 one thousand word stories, each paired with a series of tags which contain genre, character, and other feature information. We demonstrate that use of fully quantifiable feature maps as a conditional basis for generation achieves results comparable to the state of the art on multiple datasets. We also introduce a method for quantifying feature map/story relationship, and use this metric to show that the feature maps have a limited, but extant relationship to the generated text. Future use of quantitative analysis in hierarchical generation will aid researchers in effectively constructing and using first-step prompts for story generation.

- Mentor: Angelica Sun
- Sharing project: None

1 Introduction

The introduction explains the problem, why it's difficult, interesting, or important, how and why current methods succeed/fail at the problem, and explains the key ideas of your approach and results. Though an introduction covers similar material as an abstract, the introduction gives more space for motivation, detail, references to existing work, and to capture the reader's interest. Storytelling is considered one of the most complex tasks for models dedicated to text generation. Story generation is a challenge that consists of generating long, narratively-coherent passages of text about a particular topic, differentiating it from other text generation problems. Even models which successfully generate fluent text struggle to include long-range dependencies and maintain subject matter consistency. One more recent approach to text generation is to take ideas from the field of text summarization, such as the seq2seq-based method used by Liu et al. [1], and apply the methods "in the opposite direction," implicitly assuming the existence of an underlying manifold of textual information. Similarly to image compression, learning which pieces of text data are salient and which to omit build better representations of language by assisting in the understanding

of which features are most critical, and the mechanisms by which they matter. Hierarchical generation uses seq2seq-based models to translate some textual prompt into longer-form text. In the paper "Hierarchical Neural Story Generation", Fan et al. collect a paired dataset of prompt/story pairs, and demonstrate improved storytelling in text generated using a convolutional seq2seq model. This hierarchical model essentially involves first generating a base template of sorts (in this case, a sentence called "the prompt"), and then using that prompt or conditioning on that prompt to generate a longer piece of text (in this case, "the story"). The authors claim that using this kind of hierarchical generation reduces the tendency of the model to drift off topic, and also heightens the ability of the model to maintain narrative coherency. However, there is currently no way of quantitatively describing the relationship between a story and the writing prompt used to inspire it. Thus, we propose the use of "emotion maps" – – to generate more coherent stories, with the process outlined in Figure 1. The quantitative nature of the emotion maps allows for quantitative analysis of the

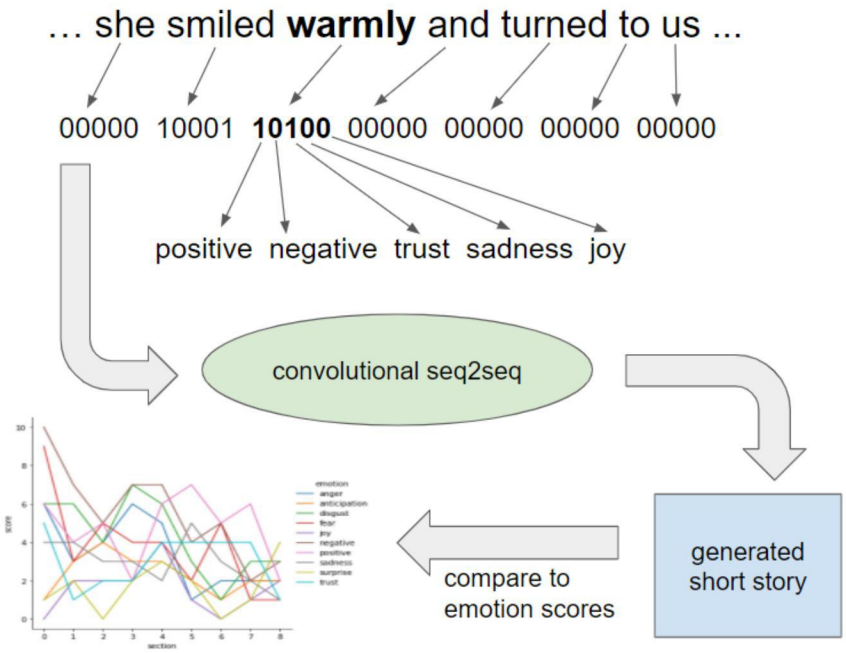

Figure 1: Graphical summary of the overall approach.

prompt-story relationship. Using a metric developed for this paper, the Average Emotional Similarity (AES), we observe a limited but present relationship between the emotion map and the generated story. The model perplexity seems to remain largely unchanged despite the significant difference between textual prompts and numerical arrays, and generated stories retain grammaticality and some coherency.

2 Related Work

The paper by Fan et al., introduced above, is the primary basis for the efforts described in this paper. In addition to exploring hierarchical story generation, and creating a dataset to do so, the authors note that recurrent architectures can be typically quite inefficient on modeling longer documents, and instead choose to use a convolutional seq2seq architecture, allowing for the encoding of stories in parallel. However, Fan et al. are by no means the only authors to attempt to emulate higher-order structure using hierarchical generation. Wu et al. describe successful use of hierarchical neural networks to generate musical melodies with long-term dependencies [2]. Using similar principles in the realm of text generation,

Li et al. describe the use of sentence-paragraph hierarchy to develop a hierarchical neural autoencoder for documents [3].

Finally, using semantic feature maps for hierarchical generation of text was partially inspired by similar tactics used in paired adversarial generation techniques. This approach has previously been applied to images, such as the Pix2Pix model developed by Isola et al. [4], which can generate realistic street images from semantically-segmented patterns as seen in Figure 2.

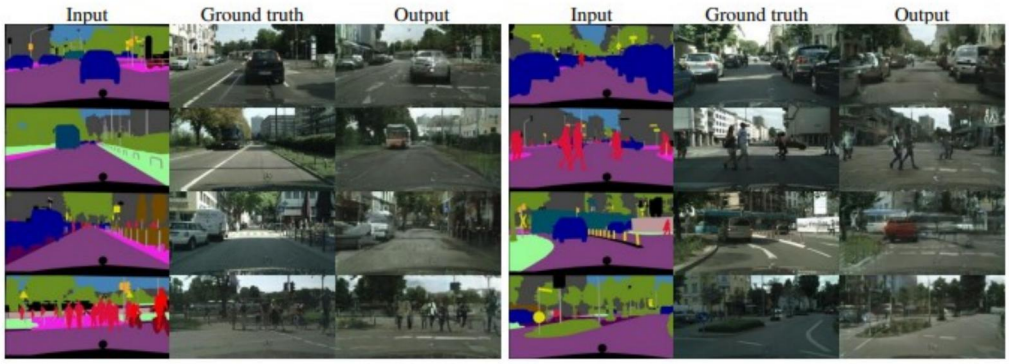

Figure 2: Image taken from the Pix2Pix paper, showing the generation of fake streets using semantic maps derived from images of real streets. [4]

3 Approach

To generate long-form text with higher-level dependencies using hierarchical generation, we use emotion maps rather than text prompts as the basis for generation, with an approach otherwise similar to Fan et al. [5]. The model is a convolutional seq2seq model, with the architecture introduced by Gehring et al. [6], as implemented by Fan et al. The model code used was taken directly from the Github repository provided by Fan et al., and only minor modifications were made in order to allow it to accept emotion-based feature maps rather than word prompts. These modifications did not affect the model architecture, a schematic of which can be found in Figure 3.

We have written code to adapt any story dataset to a paired emotion map-story dataset. Practically, these emotion maps consist of an array of floats, where each column corresponds to a word (or a section of text), each row corresponds to a quality of interest, such as emotional valence, and the value at a particular point contains the normalized score for the attribute over the source text. Emotion maps are automatically generated using the NRC Emotion Lexicon [7], which extracts sentiment-containing words for ten word valences, ranging from the generic (positive and negative) to the much more specific (joyful and disgusted).

4 Experiments

We describe the collection and features of a new dataset with tag-story pairs, and use our new code to automatically generate emotion-based feature maps over any text using the NRC Emotion Lexicon [7]. Then, we report results achieved using emotion maps as part of a hierarchical model, on both the newly collected dataset and the Reddit dataset used by Fan et al. Using purely numerical emotion maps rather than textual prompts, we achieve a perplexity comparable to the results reported by Fan et al., and samples of comparable qualitative coherency.

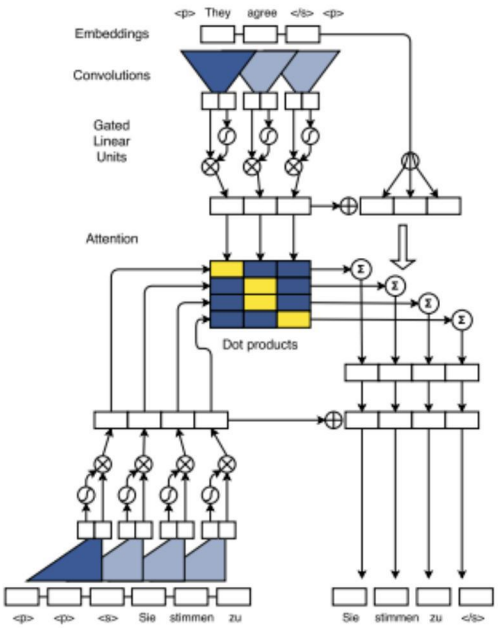

Figure 3: The convolutional seq2seq model introduced by Gehring et al., and used for these experiments to generate stories from emotion maps. [6]

4.1 Data

To establish a strong basis for comparison, we initially use the story and prompt paired dataset used by Fan et al. [5], generated by scraping the Reddit forum r/writingprompts. This dataset consists of short (less than one hundred words) prompts, paired with stories written using them (with an average story length of 734 words), in total 303,358 human-generated stories. It has allegedly already been preprocessed for quality and appropriateness, but the dataset as provided contained numerous typos, ubiquitous swearing, and a number of stories with very concerning or inappropriate subject matter. In order to most faithfully reproduce the results in the paper, we did not add any additional preprocessing steps. Subjective exploration of the dataset reveals that there are also frequent instances of repetition or off-topic prompt-story pairs, which are cited as common failure modes in the generated stories, raising the question of whether the model might be learning these behaviors to some extent.

Thus, in this paper, we present the second dataset dedicated to story generation: a dataset of 100,000 one-thousand-word stories taken from the fanfiction website Archive of Our Own (AO3), each paired with its associated list of content tags and some useful metadata. Content tags contain information both related to the concrete features of a story (such as main characters, eg. "Harry Potter") as well as more nebulous characteristics (such as genre, eg. "Humor"). We wrote a web scraper and preprecessor, included in the supplementary materials, which allow any user to specify an initial page of the website, as well as the number of works to scrape. Because the website has an extremely robust content filtering system, works with inappropriate or undesired material can be easily excluded – the dataset presented in this paper only contains appropriate stories, and has been further processed for quality using story-associated metadata. A qualitative examination of this quality-controlled dataset reveals a generally higher quality of writing samples. Additionally, the average number of typos per story and average length of repeated substrings are much lower in a random sample of stories from the AO3 dataset, in comparison to the Reddit dataset. However, this new dataset has some notable limitations and unique characteristics. There is a higher incidence of <unk> tokens (unknown or low-frequency words) in the AO3 dataset, likely due to the increased domain-specific vocabulary (such as character and place names, i.e. "Hogwarts"). The dataset also is domain-specific, having been drawn from a mixture of *Harry Potter* and

Lord of the Rings fanfictions; fantasy elements appear much more frequently in the generated stories. However, it is possible that this specificity is an asset, as the presence of consistent characters and places might enable models to learn and reproduce domain-specific features, likely resulting in more consistent stories.

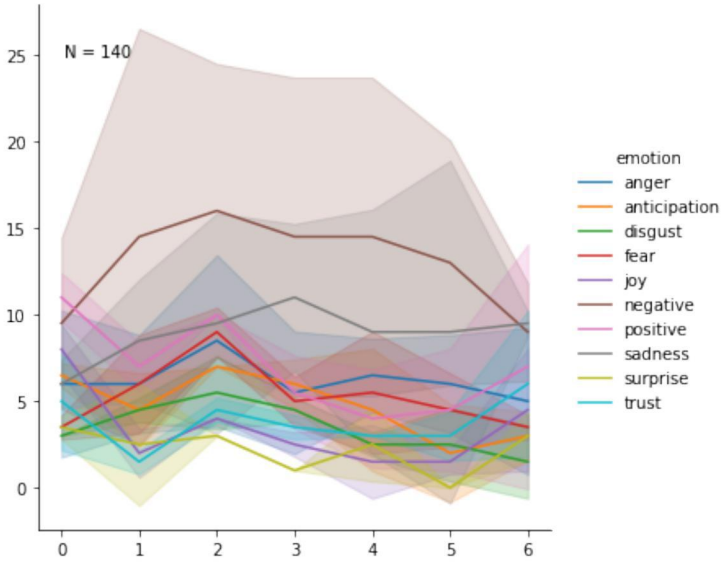

Figure 4: Average emotion map characteristics for stories tagged "Depression" in the new dataset. Note the strong prevalence of negative emotion.

4.2 Evaluation method

For automatic evaluation, we use the exact implementation of calculations of the model perplexity used by Fan et al. In addition to this, Equation 1 describes the new metric, Average Emotional Similarity (AES), that we use to measure the adherence of a given story to its emotion map prompt, where $window - size$ is the number of words used per batch and $num - emotions$ refers to the number of emotions available (in the case of this paper, $num - emotions = 10$).

$$AES = avg(\sum_{w=0}^{window-size} avg(\sum_{e=0}^{num-emotions} \frac{count(w[e])_{generated} - count(w[e])_{original}}{count(w[e])_{generated}) + count(w[e])_{original}}) \quad (1)$$

In simple terms, the story is first converted to an emotion map itself. Then, the emotion map prompt and the generated story's emotion map are sliced into pieces of size $window - size$; the count of occurrences for each emotion is summed, then the percentage difference between the two feature maps is calculated. The final AES score is an average of the subscore, taken across all windows. Practically speaking, an AES score closer to zero is more desirable. We also directly compare the count of emotion-associated words in the story with the count of non-zero elements in the emotion map. Finally, some limited human evaluation was used to evaluate story realness, as discussed in the Analysis section.

4.3 Experimental details

Experiments were run with close adherence to the setup reported by Fan et al., using the same learning rate (0.25), both gated and self attention mechanisms, and the same conditions for stopping (no change in loss for multiple epochs). Since training took, on average, about eighty hours per experiment, we did not conduct a hyperparameter search. Preprocessing of the Reddit dataset was kept intentionally minimal in order to best replicate the experimental conditions used by Fan et al, in order to use non-feature map performance on the dataset as a baseline. However, on both datasets, samples were manually curtailed to 1000 words,

as this step was used by Fan et al. and additionally has the benefit of standardizing the emotion map size.

4.4 Results

We achieve comparable perplexity to the values reported by Fan et al., as seen in Table 1.

Experiment	Valid Perplexity	Test Perplexity
seq2seq, Reddit, Fan et al.	37.37	37.94
seq2seq + fusion, Reddit, Fan et al.	36.08	36.56
seq2seq + emap, Reddit	37.04	37.87
seq2seq + emap, AO3	37.43	38.30

Table 1: Perplexity achieved for different experiments. The first two rows correspond to results previously reported by Fan et al, using the same underlying seq2seq architecture. Row 2 additionally uses a fusion mechanism ill-suited to the methods used in this paper, but is provided as a state-of-the-art metric, as it contains the best results reported by Fan et al. "Reddit" denotes usage of the r/writingprompts dataset, while "AO3" denotes use of the novel fanfiction dataset. "emap" refers to usage of emotion maps as the conditional basis for generation.

Stories generated using either dataset displayed coherent punctuation, and largely grammatical sentences, as can be seen in randomly selected Samples 1 and 2. In all samples, the <newline> token has been replaced with actual line breaks to improve human readability. Similarly, samples have been edited to remove spaces between words and punctuation where applicable, without editorializing (i.e., strictly following punctuation rules and strictly using generated punctuation).

> The old man was wearing the black suit . He was an old man .
> "I'm sorry, sir." The man said with a grin, "You're the one who has been working in the old world."
> The man sighed and took a sip of his coffee and pulled it in his pocket.
> "I've got some of you..."
> There, now, then, the man stood, and turned back to the old woman in the room. He was wearing a white t-shirt and tie, and a pair of blue dress, and was a woman.
> "Do you know the thing?"
> A woman's eyes were the same, and then she was a <unk> man. I looked in horror, and with that, a young woman, with a blue dress.

Sample 1: Generated using emotion maps, with Reddit dataset

> This was the first time in the whole situation.
> "Yeah," she said, staring at her in awe. "It is a good idea," he said, "I believe that you have been gone for you."
> "I am not."
> "I will be able to live with books and share a replacement," she said, "I am sorry. I am sorry, but I did not remember me," she is, he said. "I am not looking at me."
> "What?"
> "No, no," he said, "I am not."
> "No, no," she said, getting up and presses her against his chest. "I was just wondering if I can help it."
> "Because, I do not know."
> "I am not a freak."
> "Oh..."
> "Ah, and if you are not."
> "I know," he said, pausing on his face, and he turned to look at the back of his head.

"You are supposed to be the one who can not."
"I am not a good kisser, and said 'dear sir, you know it is not my fault you would be here."
Harry was outraged at him.
"You are not a wizard," he said, voice slightly. "I did not."
"I am not going to see you in the library.“

Sample 2: Generated using emotion maps, with AO3 dataset

5 Analysis

The generated samples were largely of comparable quality to the stories generated by Fan et al., but still contained typos, grammatical errors, and repetition. The samples generated using the AO3 dataset contained these flaws far less often. However, an analysis of the Reddit dataset suggests this may not be due to the use of emotion maps.

Eyes opened eyes opened eyes opened
Stay awake they say
Caffeine they say
Awake «unk», eyes opened on the world
Never
Close
(But why?)
No dreams for me they say
No sleep for me they say
Drugs
Keep me awake
Awake forever
Eyes opened eyes opened
Blinking
Do not blink too long

Sample 3: Human-generated sample from the Reddit training dataset, truncated to fifteen lines

Sample 3 is a real training example from the Reddit dataset, but is somewhat non-sensical and highly repetitive. Qualitatively, we showed the above sample and several other randomly selected samples to other humans (n = 3) and found that human-generated stories from the r/writingprompts dataset were frequently misclassified as machine-generated. The top reason cited for classifying the above story as machine-generated was its incoherency and repetition, although one participant added that the presence of the typos actually caused the above example to seem more realistic. In contrast, stories generated using AO3 were more frequently interpreted as real.

“ You should have gotten this . ”
<unk> . “ <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> . ”
<unk> . “ <unk> <unk> <unk> <unk> . <unk> <unk> . <unk> <unk> <unk> <unk> . ”
<unk> . “ <unk> <unk> . <unk> <unk> . <unk> <unk> . <unk> <unk> <unk> . <unk> <unk> . <unk> <unk> . “ <unk> <unk> . ”
<unk> . <unk> the <unk> <unk> <unk> <unk> . A little . <unk> . .
<unk> . <unk> <unk> . <unk> . <unk> . <unk> <unk> . <unk> . <unk> . *
<unk> , the <unk> <unk> <unk> . <unk> <unk> . <unk> . <unk> . <unk> . <unk> . <unk> . <unk> . <unk> . <unk> . <unk> . <unk> . <unk> . <unk> .

<unk> . <unk> . <unk> . <newline> <newline> <unk> . <unk> .
<unk> . <unk> . <unk> . <unk> . <unk> . <unk> . <unk> .
<unk> . <unk> . <unk> . <unk> . <unk> . <unk> . <unk> .
<unk> . <unk> . <unk> . <unk> .

Sample 4: Generated using emotion maps, with Reddit dataset

A major failure mode of the trained model is its performance on emotion maps derived from largely emotionless stories. Sample 4 shows a representative "story" generated from an emotion map containing only five non-zero elements. Emotion maps with a very low number of emotion-containing words usually result in stories filled with unknown characters and bizarre punctuation. This failure mode is more common in the Reddit dataset than the AO3 dataset, likely due to the higher prevalence of emotion-associated words in the AO3 dataset.

6 Conclusion

We find that stories generated using emotion maps are of comparable quality to stories generated based on a prompt, and that the level of emotion found in generated stories is similar to the level of emotion described by the emotion map, using a newly created metric (AES) for evaluating emotion map-story relationship. To overcome some of the limitations of existing datasets, such as poor sample quality, we also introduce a new story generation dataset which pairs genre and character information with stories.

One interesting potential avenue for future work would be to attempt adversarial text generation using emotion maps and the convolutional seq2seq model. This could emulate the mechanism used by Pix2Pix, in which the loss function accounts for the possibility of many semantically valid but visually different outputs (for example, coloring a car-coded segment as having white paint or blue paint), which has obvious similarities to the problem of story generation from semantic feature maps. Furthermore, the non-recurrent nature of the model architecture would solve the problem faced by many modern text generations – that recurrent models are not differentiable. Another interesting area for further research opened up by this work is to pursue alternative feature maps, using similarly quantifiable features, such as grammar, or even a combination of multiple features. We anticipate that methods which incorporate data from the distribution of the source text can enable models to better emulate long-term dependencies and common subjects in longer-form generated text.

References

[1] Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences. *CoRR*, abs/1801.10198, 2018.

[2] J. Wu, C. Hu, Y. Wang, X. Hu, and J. Zhu. A hierarchical recurrent neural network for symbolic melody generation. *IEEE Transactions on Cybernetics*, 50(6):2749–2757, 2020.

[3] Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. A hierarchical neural autoencoder for paragraphs and documents. *CoRR*, abs/1506.01057, 2015.

[4] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.

[5] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Conference of the Association for Computational Linguistics (ACL)*, 2018.

[6] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning. *CoRR*, abs/1705.03122, 2017.

[7] Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. 29(3):436–465, 2013.