Multimodal Models for Learning to Order Sentences in Manga

Brian Lim
Department of Computer Science
University of California, Santa Barbara
Santa Barbara, CA 93107
blim@ucsb.edu

Abstract

We present a multimodal problem to order text in a comic's pages. We modified two state of the art models to include multimodal inputs, the first of which is a transformer-based model, and the second learns relative orders which can be ordered using topological sort. These models are tested on Japanese Wikipedia and the Manga109 dataset. We show that a multimodal model performs better than a text-only model, and verify that relative order-based models outperform transformer models in the sentence ordering task.

1 Introduction

Comics are popular all over the world, and there are many different forms, from superhero comics, manga, webtoons, manhwa, webcomics, all of which have different characteristics. The world's media has become so interconnected that people from different cultures have easy access to read comics from other cultures through fan translations, that is, if the comics are interesting enough to be translated by readers that know multiple languages. In order for a comic to succeed on the market now, it needs to be either published in a medium that has the capability to have an official translation, or it needs to be interesting enough that fans translate it for free and provide it worldwide reach, in which scenario new readers that also find it interesting will buy the source material to collect. This is why there is an ongoing effort to provide an end-to-end pipeline for fully automated comic translation so that it is easier for interesting comics to surface on the internet without the need for preexisting multilingual fanbases or large companies to provide official translations.

One particular issue with the current end-to-end pipeline in automated manga translation developed by Hinami, et al (2020) [2] is that ordering speech bubbles has only a 91.9% accuracy, which on its own is fine, but with many additional components to the pipeline, this small inaccuracy will compound with the errors in OCR and machine translation. The method proposed by them is a simple recursive algorithm that breaks down the page into vertical and horizontal strips and attempts to calculate the proper text order using that recursion. This problem is a reformulated *sentence ordering* problem where a model is given a randomly ordered list of sentences and has to predict the correct order. With the medium being comics, there is also additional information present in the comic images, leading to the potential of a multimodal sentence ordering model. There has been no research done on this, which requires us to create a new dataset that orders comic book text.

1.1 Contributions

This paper addresses the problem of ordering text in manga specifically, but it can be extended to other comics given a proper dataset. We make the following contribution:

Preprint. Under review.

Multimodal sentence ordering model . The primary contribution is a model that is able to order text in manga using the text and the comic book page images. We demonstrate it is a viable method for ordering text in manga, enabling us to use it in multimodal translation pipelines.

2 Related Work

2.1 Attention order networks (RankTxNet)

The previous state of the art model for sentence ordering is RankTxNet created by Kumar et al. (2020)[4]. It uses BERT as a sentence embedding extractor and a transformer to predict the order of sentences within a paragraph. This model was trained and tested on several research paper abstracts, captions, and story paragraphs.

2.2 Sentence ordering using topological sort

The current state of the art model beating RankTxNet for best method of ordering sentences is a relative ordering that gets evaluated using a depth-first topological sort created by Prabhumoye et al. (2020)[5]. It uses either an LSTM or BERT to determine the relative ordering between each pair of sentences, then sorts them to produce an ordering. This method has a downside that it takes longer to compute than RankTxNet due to the n^2 number of pairs for a paragraph with n sentences and isn't parameterized on the entire paragraph to create orderings, but it has the highest accuracy.

2.3 Multimodal bitransformers

Facebook research created a model that is able to classify a combination of images and text that is able to combine pretrained image neural networks and pretrained BERT-like architectures (Kiela et al. (2020)[3]). This model was able to be performed on par with state of the art multimodal models on various multimodal texts.

3 Technical Approach

In this paper, we propose 2 multimodal models that are extensions of previous state of the art models.

3.1 Multimodal attention order network (MMRankTxNet)

Similar to MMBT [3], we propose an attention order network that extracts sentence embeddings and uses ResNet to learn image embeddings of the same size as the image. These embeddings are run through a transformer and fully connected network that outputs a score for every sentence embedding, of which they will be sorted in ascending order.

Pointwise ranking loss The loss function is the mean squared error of the predicted score z_i and the target score y_i. The target score y_i is defined as $\frac{(k-1)}{n-1}$ where $k \geq 1$ is the index of the sentence in the gold order and $n \geq 2$ is the total number of sentences in the paragraph. The loss is calculated as $\sum_{k=1}^{n}(y_k - z_k)^2$.

3.2 MMTopoSortNet

Similar to the topological sort model proposed by Prabhumoye et al. [5], we can extend that model to combine a dual sentence embedding with an image embedding and predict which of the two sentences came first. This prediction determines the arrow that defines the graph between all nodes. This graph output can be sorted using a depth-first search topological sort to order the sentences.

4 Experiments

We conduct an analysis of both models on two datasets. Since this is a new problem, there are no existing benchmarks, thus both datasets had to be preprocessed[1].

[1] Code repository: `https://github.com/BLimmie/manga_ordering`

Figure 1: MMRankTxNet page encoder.

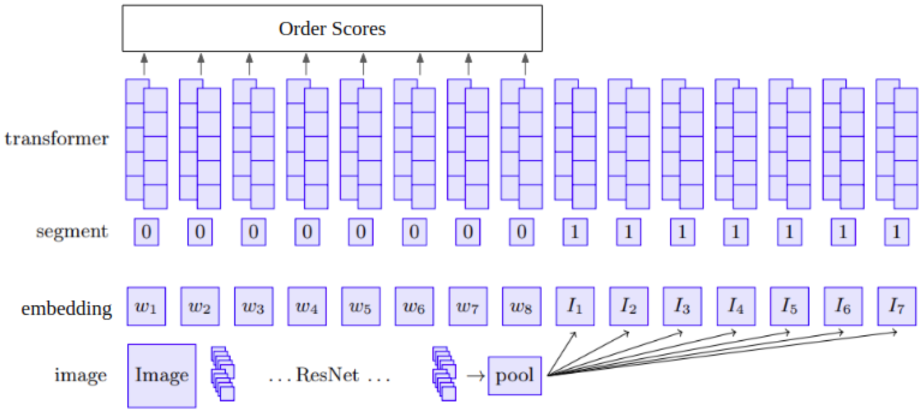
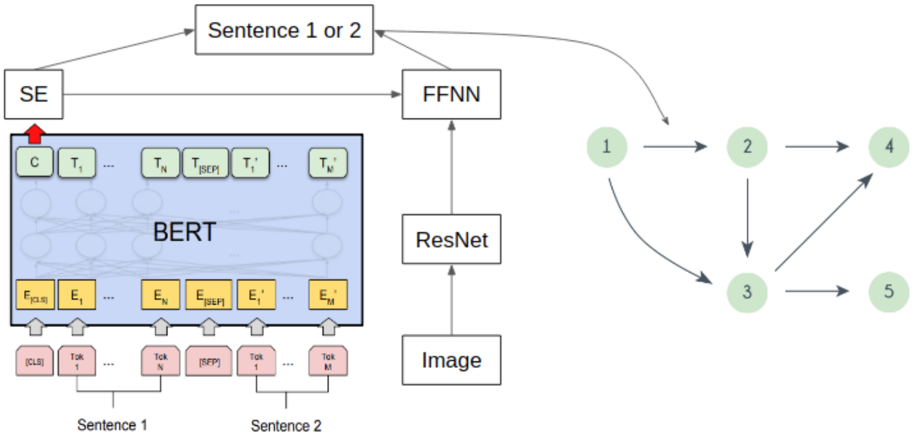


Figure 2: MMTopoSortNet architecture and pipeline.

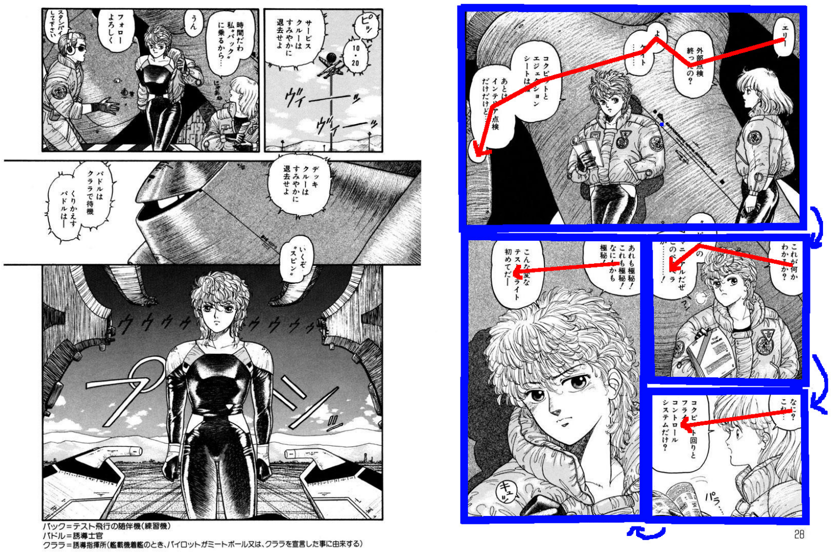


4.1 Datasets

Japanese Wikipedia paragraphs The entire archive of the Japanese Wikipedia was split into paragraphs and sentences given orders. This was used as a baseline dataset for the other dataset's metrics and as a pretraining dataset for other models.

Manga109 The Manga109 dataset created by Aizawa et al. (2020) [1] contains the pixel locations of all frames, textboxes, and character models. Using the top-right and bottom-left locations of each frame it was possible to create a set of rules to create a relative ordering set by trial and error which cover 99% of cases found in the dataset. The remaining 1% of cases had their annotations manually tweaked to fit the existing ruleset. Then the frames are ordered by topological sort. For the purposes of this dataset, this algorithm works, but it cannot be assumed to extend to other comic book datasets due to the amount of manual tweaking necessary. Within each frame, each speech bubble, which will for the purposes of this experiment will be called sentences, is then ordered by shortest total distance calculated using a traveling salesman problem solver, starting the sentence with Manhattan distance closest to the top-right corner of the frame. The images then had their text masked. I estimate the accuracy of this method is around 95%, but since there is no publically available dataset to compare against, it is difficult to say how accurate it really is.

Figure 3: Example ordering of half a page.


エリー
外部点検終ったの?
よ
ウエイト……
コクピットとエジェクションシートは
あとはインテリア点検だけだけど……
これが何かわかるか?
このペラのマニュアルだぜ……!
あれも極秘! これも極秘! なにもかも極秘!
こんな変なテストフライト初めてだー
なに? これ……
コクピット回りとフライトコントロールシステムだけ?
キュッ
パラ……
28
ピッ
10・20
サービスクルーはすみやかに退去せよ
ヴィーーッ
ヴィーーッ
時間だわ 私〝パック〟に乗るから…
うん
じゃあね
フォローよろしく
スタンバイして下さい
デッキクルーはすみやかに退去せよ
パドルはクララで待機
くりかえすパドルは—
いくぞ〝スピン〟
ウウウウ
ウウウウ
パック=テスト飛行の随伴機(練習機)
パドル=誘導士官
クララ=誘導指揮所(艦載機着艦のとき、パイロットがミートボール又は、クララを宣言した事に由来する)

4.2 Hyperparameters

MMRankTxNet Since sentence ordering can quickly result in large models, we chose a size for MMRankTxNet that would fit on a powerful consumer GPU. This results in a sentence encoder with 12 transformer blocks, hidden size 768, 12 attention heads, and intermediate layer size 3072. The page encoder consists of 4 transformer blocks, and the same hyperparameters set for the rest of the transformer. The page decoder consists of a 4 layer, 768 intermediate size fully-connected network. The image encoder uses ResNet18 due to the size constraints and outputs 8 embedding vectors (768 × 8). This was trained with an AdamW optimizer with an initial learning rate of 5×10^{-5} for the sentence encoder and 5×10^{-3} elsewhere.

MMTopoSortNet The dual sentence embedding uses the same architecture as MMRankTxNet. The image encoder uses ResNet18 due to size as well. The decoder uses an 8 layer, 768 intermediate size fully-connected network. This was pretrained with an AdamW optimizer with an initial learning rate of 5×10^{-6} and trained on multimodal input with a learning rate of 5×10^{-5}.

4.3 Evaluation

Kendall's Tau (τ) For a sequence of length n, Kendall's Tau (τ) is defined as $\tau = 1 - 2 \times (\# \, inversions)/\binom{n}{k}$. This metric correlates with human judgements.

Perfect Match Ratio (PMR) PMR is the fraction of number of correct orderings over all the paragraphs. This doesn't consider partial matches, so it is a difficult evaluation metric to understand.

4.4 Results

Table 1 provides the results of experiments on both the Wikipedia and Manga109 dataset. We can see immediately that MMRankTxNet did not perform well and is unable to agree on the proper order. Meanwhile, topological sort was able to perform much better than the baseline and outperforms the original accuracy in the pipeline proposed by Hinami et al. [2] of 91.9%

5 Discussion

Model differences The most likely explanation for why a topological sort model performs better than a transformer based model is that the problem is broken down into individual pairs. Otherwise,

Table 1: Kendall's tau (τ) and perfect match ratio (PMR) on test set for both datasets.

Dataset		MMRankTxNet		MMTopoSortNet	
Name	Image Type	τ	PMR	τ	PMR
Wikipedia	No Image	0.0015	5.7%	0.852	79.7%
Manga109	No Image	0.002	1.0%	0.993	96.3%
Manga109	Masked Image	0.005	1.6%	0.999	99.5%
Manga109	Unmasked Image	0.024	2.3%	0.999	99.5%

the dimension of the inputs and hidden states is larger than neural networks are built to handle, taking into consideration that BERT is effectively being layered on top of itself multiple times along with ResNet. By reducing the inputs to pairs, we effectively decrease the model size to a manageable chunk and provide it a more easily trained problem of classification instead of ranking or regression. The most prominent downside of MMTopoSortNet is that preprocessing the sentences is very costly, requiring every pair of sentences to be tokenized using the BERT Tokenizer provided by HuggingFace.

Images improving the model We see in both transformers and topological sort models that adding images improves the accuracy of the model. It could be that the models are learning the ruleset used for processing the Manga109 dataset, but it doesn't discredit the fact that a multimodal model performs better than a text-only model.

Japanese as a language Japanese is notorious in machine translation for being a difficult language to translate due to the lack of context clues in the language, especially in conversational scenarios found in manga. Even humans have difficult with sentence ordering in Japanese so the results associated with MMRankTxNet are partially expected, but not nearly that bad.

Bugs in training time It is possible that the training for the multimodal versions of MMTopoSortNet had bugs in training time related to the learning rate scheduler provided by HuggingFace that cause it to overfit to picking the first sentence in every pair, since at test time, this was the only pairing evaluated due to time constraints of this paper being due[2]. This problem did not exist for the text-only model since it took significantly less time to train and test and was able to be properly debugged. There is reason to believe that the multimodal models do perform better, but the numbers seem suspiciously good.

6 Conclusion and Future Work

We have shown that in the domain of comics, a multimodal model performs better compared to a text-only model. Relative ordering and topological sort-based models seem to perform much better than Attention Order Networks due to the simplification of the problem statement. This method has limitations however, and future work is to see if these multimodal methods works on other kinds of comics besides manga.

Acknowledgments and Disclosure of Funding

This work was done as part of UCSB's CS291A course, Deep Learning under Professor William Wang.

References

[1] Kiyoharu Aizawa, Azuma Fujimoto, Atsushi Otsubo, Toru Ogawa, Yusuke Matsui, Koki Tsubota, and Hikaru Ikuta. Building a manga dataset "manga109" with annotations for multimedia applications. *IEEE MultiMedia*, 27(2):8–18, 2020.

[2] As a sidenote, this is an interesting problem for an undergraduate researcher to verify.

[2] Ryota Hinami, Shonosuke Ishiwatari, Kazuhiko Yasuda, and Yusuke Matsui. Towards fully automated manga translation. *arXiv preprint arXiv:2012.14271*, 2020.

[3] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. *CoRR*, abs/1909.02950, 2019.

[4] Pawan Kumar, Dhanajit Brahma, Harish Karnick, and Piyush Rai. Deep attentive ranking networks for learning to order sentences. In *AAAI*, pages 8115–8122, 2020.

[5] Shrimai Prabhumoye, Ruslan Salakhutdinov, and Alan W Black. Topological sort for sentence ordering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2783–2792, Online, July 2020. Association for Computational Linguistics.