

Visual Story Ordering with a Bidirectional Writer

Wei-Rou Lin

Department of Computer Science and
Information Engineering,
National Taiwan University
Taipei, Taiwan
wrlin@nlg.csie.ntu.edu.tw

Hen-Hsen Huang

Department of Computer Science,
National Chengchi University
MOST Joint Research Center for AI
Technology and All Vista Healthcare
Taipei, Taiwan
hhhuang@nccu.edu.tw

Hsin-Hsi Chen

Department of Computer Science and
Information Engineering,
National Taiwan University
MOST Joint Research Center for AI
Technology and All Vista Healthcare
Taipei, Taiwan
hhchen@ntu.edu.tw

ABSTRACT

This paper introduces visual story ordering, a challenging task in which images and text are ordered in a visual story jointly. We propose a neural network model based on the reader-processor-writer architecture with a self-attention mechanism. A novel bidirectional decoder is further proposed with bidirectional beam search. Experimental results show the effectiveness of the approach. The information gained from multimodal learning is presented and discussed. We also find that the proposed embedding narrows the distance between images and their corresponding story sentences, even though we do not align the two modalities explicitly. As it addresses a general issue in generative models, the proposed bidirectional inference mechanism applies to a variety of applications.

CCS CONCEPTS

• **Computing methodologies** → **Cognitive robotics**; *Discourse, dialogue and pragmatics; Image representations.*

KEYWORDS

Multimodal modeling, temporal information ordering, sentence ordering, visual-semantic representation

ACM Reference Format:

Wei-Rou Lin, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. Visual Story Ordering with a Bidirectional Writer. In *Proceedings of the 2020 International Conference on Multimedia Retrieval (ICMR '20), June 8–11, 2020, Dublin, Ireland*. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3372278.3390735>

1 INTRODUCTION

On social media sites such as Instagram and Flickr, and blogging platforms such as Medium, stories are recorded and created as a hybrid of text and images, since text and images play different roles in organizing and comprehending these stories. This work investigates machine comprehension by leveraging multimodal information from both text and images. We propose visual story

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ICMR '20, June 8–11, 2020, Dublin, Ireland

© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7087-5/20/06...\$15.00
<https://doi.org/10.1145/3372278.3390735>

ordering, a multimodal task involving sentence ordering and temporal image ordering. Here, we define a visual story as a significant ordered sequence of plot points, each of which can be a few sentences, a clause, or an image. Given a story whose plot points are disrupted to a random order, the aim of visual story ordering is to reconstruct it in the correct order. Figure 1 shows a visual story consisting of five plot points. The first plot point is a complete sentence, the second contains two sentences, the third one is a clause only, and the last two are images.

Temporal ordering is a high-level cognitive task for humans [6, 14]. One task that measures ability in verbal comprehension is the temporal ordering of text, such as sentence ordering; this is often found in language tests to assess reading ability. This evaluates a person’s knowledge of grammar and rhetoric in a specific language. Temporal ordering of visual information—such as picture arrangement—is a subset of some intelligence tests. The picture arrangement score indicates the subject’s ability to understand nonverbal behavior [5].

New challenging issues arise along the novelty of the visual story ordering problem. Previous studies on sentence ordering conducted experiments on corpora extracted from academic research papers or news [2, 4, 13, 17]. In contrast to the rigorous writing style of

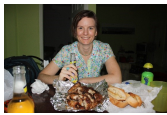

Order	Content
1.	One day while [male] was walking his dog, he had an idea.
2.	He should make his wife dinner and have a movie night with her! They wouldn’t even have to leave their little shack of a house, and it would be very romantic.
3.	So [male] started right away on cooking a fantastic dinner,
4.	
5.	

Figure 1: A visual story extracted from the VIST dataset

research papers and news, a visual story can be randomly generated by anyone in any circumstances without strict constraints, leading to a more casual textual form. Stories are looser in structure and less coherent, compared to corpora collected from research papers; however, since ordering methods rely on the coherence and linkage between sentences, this characteristic of stories only makes the problem more difficult.

From a visual point of view, temporal ordering is more complicated. To pair with the textual ordering task, the image order we describe here refers to the temporal order in which the photos were taken. When the temporal distances between the story images are small, for instance seconds apart, for humans the task becomes somewhat instinctive. To order image sequences such as the neighboring frames of a video, image forecasting and dynamic object detection strategies can be used. Meta-cognitive knowledge of ordering strategies, such as deduction from object position interpolation or extrapolation, can be acquired and even described; for instance, do water drops generally descend, or do they ascend?—this is a clue that can be used to order images. However, when photos are taken over a longer time interval, the problem becomes more difficult. The coherence of the image sequence drops as the time interval increases. When a human tries to order this kind of image sequence, a story is imagined depending on the person’s past experience and knowledge.

This work presents a model for visual story ordering based on an architecture consisting of a reader, a processor, and a writer with a self-attention mechanism. Both textual and visual plot points are represented within a single multimodal embedding. Furthermore, we also propose a novel bidirectional beam-search algorithm for decoding. We construct an evaluation dataset based on the VIST dataset [10]. Experimental results show the effectiveness of our approach. In addition, our multimodal embedding narrows the distance between an image and its corresponding sentence even though we do not explicitly align the two modalities.

The contributions of this work are threefold.

- (1) We introduce visual story ordering, a novel task by which to explore machine comprehension of multimodal information.
- (2) We propose a novel model with a bidirectional beam-search decoder for sequence ordering, and confirm the information that is captured in the multimodal representation.
- (3) As the proposed bidirectional decoder addresses a general issue of generation tasks, the approach can be applied to other tasks such as machine translation and caption generation.

2 RELATED WORK

2.1 Sentence Ordering

Sentence ordering is an essential and challenging problem. Hand-crafted features and traditional statistical approaches are not useful for the task due to the highly abstractive nature of paragraphs [13]. A primary application of sentence ordering is found in multi-document summarization systems [2, 4].

Lapata [13] learns the probability of sentence pair adjacencies to order sentences. Barzilay and Lapata [3] represent each text by an entity grid to build a coherence model, which is then used to measure the possibilities of each order for a sentence set. Logeswaran et al. [17] compare several neural network models for

sentence ordering in academic paper abstracts, finding the pointer network [20] to be the best.

Agrawal et al. [1] introduces the “Sort Story” task to rank a set of images. The textual information from the caption of each image is extracted as features. Different from their work, our task is to rank among images and captions, intertwinedly. Each element in the input of our task is only either an image or a sentence, but not both. Zhou et al. [24] explore story ending selection.

2.2 Temporal Image Modeling

The model architecture for learning unsupervised video representations can be similar to that for learning sentence embeddings. Srivastava et al. [18] use an encoder-decoder architecture, which predicts the future sequence of frames and reconstructs the input frames to learn video representations.

Zeng et al. [23] explore the task of visual forecasting that is aimed at predicting the next image given a number of previous ones. Ledig et al. [15] learn video representations from future frame prediction with more diversified models. They adapt generative adversarial networks (GANs) [8] for the frame prediction problem with a new gradient difference loss to sharpen image prediction.

2.3 Visual-Semantic Embeddings

Many different methods are proposed to learn the visual-semantic embedding models by forcing each image-word pair sharing similar semantic information to align in a single vector space. Weston et al. [22] employ an online learning-to-rank algorithm to train a model which maps image features to the joint embedding space. Frome et al. [7] instead propose a semi-supervised method. Some explore visual-semantic embeddings beyond one-on-one alignments between an image and a single word. Kiros et al. [12] implement an encoder-decoder pipeline learning a multimodal embedding space with image-description pairs. Karpathy and Li [11] align images, their object regions, and words enriched with the context in their captions to a multimodal embedding.

3 DATASET

We constructed the visual story ordering dataset based on the Visual Storytelling (VIST) dataset, which was designed for sequential vision-to-language modeling [10]. Note that we use the term *story* to represent an entry of the dataset and the term *visual story* to represent our definition. With photos extracted from Flickr albums, Huang et al. [10] used crowdsourcing in the construction of VIST to collect stories with text aligned to images and descriptions of the images. Triples composed of an image, a story sentence, and a description sentence constitute each story in the dataset; each story consists of five triples. The sequential image-text triples provided by VIST could be transformed into visual stories easily by selecting a modality in each triple of the story as a plot point in the resulting visual story. A total of 40,155, 4,990, and 5,055 stories are created for training, validation, and testing, respectively.

4 METHOD

The input of our model is a sequence of plots $\mathbf{x} = \{x_1, x_2, x_3, \dots, x_n\}$, where each plot x_i is either a sentence or an image. The output is a sequence of indices $\mathbf{y} = \{y_1, y_2, y_3, \dots, y_n\}$, denoting the order of

the input sequence. We attempt to learn a model to compute the conditional probability $p(y|\mathbf{x})$ with the chain rule as follows.

$$p(y|\mathbf{x}) = \prod_{i=1}^n p(y_i|y_1, y_2, y_3, \dots, y_{i-1}, \mathbf{x})$$

, where $p(y_i|y_1, y_2, y_3, \dots, y_{i-1}, \mathbf{x})$, the conditional probability at each time step i , is usually modeled by RNN. Pointer networks [20, 21] addresses the task of set to sequence. Plot points are fed to an encoder to calculate the plot point representation and generate story vectors, which are then fed to the decoder, which calculates an output vector, which is supplied to the pointer attention unit along with the sentence vectors to determine the probabilistic distribution of the next sentence. Compared to the vanilla pointer network, our model introduces an one-step self-attention mechanism for modeling the data at both the plot level and the story level. The details are shown in Section 4.1. In addition, we propose a bidirectional beam-search algorithm to mitigate multi-head inference, as explained in Section 4.2.

4.1 Multimodal Set to Sequence Model

Since the pointer network encodes each randomly ordered sequence using the recurrent network, the encoded sequence vector varies with the input order, which should not contain any information. A reader-processor-writer architecture excludes random noise from the input order, making the architecture more robust [20]. Our model is also free from random input order noise. Although the proposed approach is different from previous ones, we still borrow the terminology to describe our model.

4.1.1 Reader. The reader encodes the input plot points respectively, and can be split into the image reader and text reader in the experiments with visual plot points. The reader first encodes the input plot point x_i to a number of context vectors and then uses self-attention [16] to combine these context vectors into a single plot vector \vec{e}_i .

For the textual data, we use the bidirectional LSTM (BiLSTM) with a hidden size of h to transform word vectors to context vectors $\vec{c}_{i,1}, \vec{c}_{i,2}, \dots, \vec{c}_{i,m}$ as follows.

$$\vec{c}_{i,j} = \text{LSTM}(x_{i,j}, \vec{c}_{i,j-1}) \quad (1)$$

where $x_{i,j}$ is the j th token in x_i , and m is the length of the longest sentence in a training batch. Padding is applied to other sentences in the same batch. The dimension of each context vector is $2h$ since the context vector consists of hidden states in both directions. Let the context matrix $C_i = (\vec{c}_{i,1}, \vec{c}_{i,2}, \vec{c}_{i,3}, \dots, \vec{c}_{i,m})$. The attention weights of C_i at the plot level is computed as a matrix A_i :

$$A_i = \text{softmax}(W_1^P \tanh(W_2^P C_i^T)) \quad (2)$$

, where the matrices $W_1^P \in \mathbb{R}^{r \times d}$ and $W_2^P \in \mathbb{R}^{d \times 2h}$ are parameters to be optimized during training, and h , d , and r are hyperparameters. Finally, the plot vector $\vec{e}_i \in \mathbb{R}^r$ is computed as follows.

$$\vec{e}_i = A_i C_i \quad (3)$$

As a result, the representation of the whole plot is a vector with a constant dimension of r . We did attempt to pre-train the word vectors using word2vec with our corpus, but there was no obvious difference in comparison to the random initialization case.

For visual data, the plot representation is a sequential visual semantic embedding. To align with texts, we define image context vectors as well. The image context vectors are derived from a CNN containing a pre-trained ResNet [9] and additional convolution layers. We tried several different settings and found the concatenation of the first three blocks of ResNet and three layers of convolution, each of which is followed by a batch normalization and a ReLU activation, is more suitable. The output channels of the last convolution layer are then mapped to context vectors using a fully-connected layer. For textual plots in pairwise neural network approach, we also tried to adopt InferSent (Conneau et al., 2017) as a reader, but it was unable to perform better than the self-attentive reader.

4.1.2 Processor. In the original reader-processor-writer architecture, the processor also uses the attention mechanism to combine the plot vectors $\vec{e}_1, \vec{e}_2, \vec{e}_3, \dots, \vec{e}_n$ into a single story vector \vec{e} , representing the information of the whole \mathbf{x} . In contrast to the processor proposed in previous work [20], we feed the context vectors (i.e. $\vec{c}_{1,1}, \dots, \vec{c}_{1,m}, \vec{c}_{2,1}, \dots, \vec{c}_{2,m}, \dots, \vec{c}_{n,1}, \dots, \vec{c}_{n,m}$)—rather than the output of the reader (i.e. $\vec{e}_1, \dots, \vec{e}_n$)—directly into the processor. The architecture of the proposed self-attentive processor is similar to the self-attentive reader, but the processor takes into account all context vectors in a story. We define the context matrix at the story level $C = (\vec{c}_{1,1}, \dots, \vec{c}_{1,m}, \vec{c}_{2,1}, \dots, \vec{c}_{2,m}, \dots, \vec{c}_{n,1}, \dots, \vec{c}_{n,m})$. The attention weights at the story level and the resulting story vector \vec{e} are computed as follows.

$$B = \text{softmax}(W_1^S \tanh(W_2^S C^T)) \quad (4)$$

$$\vec{e} = BC \quad (5)$$

In our architecture, the four matrices W_1^P , W_2^P , W_1^S , and W_2^S are parameters to train. The flexibility of this model lies in this passing the context vectors through the reader and feeding them directly to the processor. Each of our stories contains five plot points, and the self-attention merely involves calculating the weighted sum from the five plot vectors to generate a story vector without the context vectors passing. With the passing presenting, each context vector provides the top one outcome of our model.

4.1.3 Writer. In the decoding step, we propose a novel bidirectional pointer writer. Noting that the validation accuracy of the first and last plot points in unidirectional pointer writer is higher than the others, we believe that decoding the output jointly from the two directions could contribute to better performance. The bidirectional pointer writer requires as input two story representations; we generate multiple story representations using the self-attention mechanism. Apart from the inputs extension and replacing the LSTM in the unidirectional writer with a BiLSTM, the training process of the bidirectional writer remains the same as that for the unidirectional one.

4.2 Bidirectional Beam Search

The biggest problem with the bidirectional decoder lies in the process of inference, as we do not know what should be fed to the next step before the model decides an output for current step. Therefore, we must resort to greedy algorithms such as beam search to find the next possible output, and a jumping of prediction may not be acceptable in this configuration. However, after using two sets of

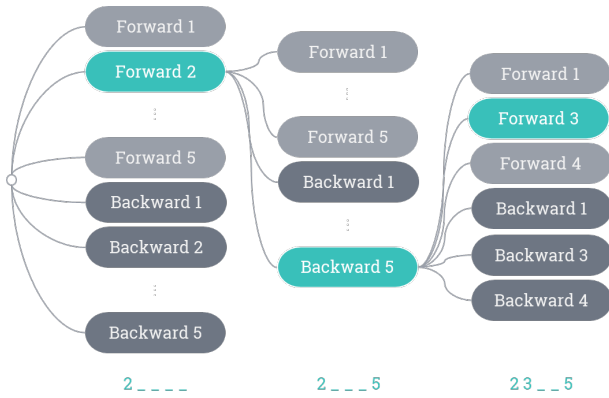


Figure 2: Bidirectional beam search

beams—forward and backward—in each step of the beam search, the situation is improved [19].

Figure 2 shows part of the bidirectional beam search process, in which the bottom line shows the decoded numbers and their positions. At each expansion, the model generates two sets of new nodes—forward and backward—from the candidate list of the expanded node. That is, each of the generated node contains the information about its selection of the next output and the decoding direction; the selected output is deleted from the candidate list of the node for the next expansion step. For example, in Figure 2, as FORWARD 2 is selected in the first step, in the second step, the possible selections are FORWARD 1, BACKWARD 1, FORWARD 3 to FORWARD 5, and BACKWARD 3 to BACKWARD 5.

5 EXPERIMENTS

We conduct two evaluations are described below.

- **Text Only:** Contains only the text in the story. The goal is to test if any of the applied strategies affect the sentence ordering results. Models trained with and without images are evaluated using this subset.
- **Text-Image Intertwined:** Each story contains two visual plot points and three textual plot points, and the interval between the two images varies from 0 to 3.

We used two evaluation metrics from previous work [13] to assess our ordering, including perfect match accuracy (Acc_p) and the mean of Kendall’s τ . Acc_p is the ratio of the number of perfect match orders, i.e., the generated order totally agrees with the story order in the dataset. For random guessing, Acc_p is 0.0008 on our dataset. The mean of Kendall’s τ is a common evaluation metric for ordering problems, assessing an order by the number of inversions. The value of τ ranges from -1 to 1. $\tau = 0$ is the expectation value of random guessing. Hyper-parameters are optimized as follows. The input image size is [224, 224]. Adam with a learning rate of 0.0001 and β of (0.9, 0.99) is used. The batch size is 32. The beam size is 120. the dimension of the word vectors is 300, the size of the self-attentive hidden layer was 100, the number of hops was 1, and the size of the LSTM unit for the writer was 600.

Table 1 reports the performance of our model on the Text Only evaluation, compared with three baseline models. The Coreference

Model	Acc_p	τ
Pairwise Pointer Network	0.1217	0.5006
Unidirectional Pointer Network	0.1306	0.5206
Our Model	0.1501	0.5245

Table 1: Results for text only evaluation without visual information given in the training data

k	Unidirectional		Bidirectional	
	Forward	Backward	Forward	Backward
0	0.7395	0.7266	0.6898	0.4144
1	0.4651	0.4805	0.5183	0.8150
2	0.5019	0.5201	0.9600	0.9439
3	0.7490	0.7282		
4	0.9642	0.9600		

Table 2: Accuracy of predicting next plot point given k previous plot points

model is a text-only model that ranks the textual plot points by calculating the directions of coreference linkings. The pairwise pointer network model is a slightly modified pointer network that learns to rank the visual story ordering in a pairwise fashion. The unidirectional pointer network is another pointer network that learns to rank in a listwise fashion with unidirectional decoding. The results show that the proposed model outperforms all baseline models; they also confirm the effectiveness of the proposed bidirectional approach.

To exam the effectiveness of the bidirectional beam search, Table 2 compares the accuracy scores of next plot point prediction at each time step with different numbers of previous plots given. We observe that the accuracy of the last plot point always is the highest. This may be because the model chooses to strengthen the feature of the last plot point in the story representation to ensure a higher accuracy. We would expect that in models dealing with longer temporal sequences, the improvements would be more obvious when replacing the unidirectional decoder with the proposed bidirectional decoder.

6 CONCLUSION

We introduce visual story ordering, a new task for modeling human cognition on multimodal information, and propose a novel model for the task. Our bidirectional inference mechanism is not limited to visual story ordering; it is general and can be applied to other sequence decoders such as those for machine translation and image-caption generation.

ACKNOWLEDGMENTS

This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST-106-2923-E-002-012-MY3, MOST-109-2634-F-002-040-, MOST-109-2634-F-002-034-, MOST-108-2218-E-009-051-, and by Academia Sinica, Taiwan, under grant AS-TP-107-M05.

REFERENCES

- [1] Harsh Agrawal, Arjun Chandrasekaran, Dhruv Batra, Devi Parikh, and Mohit Bansal. 2016. Sort Story: Sorting Jumbled Images and Captions into Stories. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 925–931. <https://aclweb.org/anthology/D16-1091>
- [2] Regina Barzilay, Noemie Elhadad, and Kathleen R. McKeown. 2002. Inferring Strategies for Sentence Ordering in Multidocument News Summarization. *J. Artif. Int. Res.* 17, 1 (Aug. 2002), 35–55. <http://dl.acm.org/citation.cfm?id=1622810.1622812>
- [3] Regina Barzilay and Mirella Lapata. 2008. Modeling Local Coherence: An Entity-based Approach. *Comput. Linguist.* 34, 1 (March 2008), 1–34. <https://doi.org/10.1162/coli.2008.34.1.1>
- [4] Danushka Bollegala, Naoaki Okazaki, and Mitsuru Ishizuka. 2010. A Bottom-up Approach to Sentence Ordering for Multi-document Summarization. *Inf. Process. Manage.* 46, 1 (Jan. 2010), 89–109. <https://doi.org/10.1016/j.ipm.2009.07.004>
- [5] Jonathan M. Campbell and David M. McCord. 1996. The WAIS-R Comprehension and Picture Arrangement Subtests as Measures of Social Intelligence: Testing Traditional Interpretations. *Journal of Psychoeducational Assessment* 14, 3 (1996), 240–249. <https://doi.org/10.1177/073428299601400305> arXiv:<https://doi.org/10.1177/073428299601400305>
- [6] Patricia Chisholm. 1970. The Structure of Intellect Theory: Implications for More Meaningful Mental Test Interpretations. (1970).
- [7] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc' Aurelio Ranzato, and Tomas Mikolov. [n.d.]. DeViSE: A Deep Visual-Semantic Embedding Model. In *Advances in Neural Information Processing Systems 26*.
- [8] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'14)*. MIT Press, Cambridge, MA, USA, 2672–2680. <http://dl.acm.org/citation.cfm?id=2969033.2969125>
- [9] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- [10] Ting-Hao (Kenneth) Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual Storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, 1233–1239. <http://www.aclweb.org/anthology/N16-1147>
- [11] A. Karpathy and Fei-Fei Li. 2017. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 4 (April 2017), 664–676. <https://doi.org/10.1109/TPAMI.2016.2598339>
- [12] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014. Multimodal Neural Language Models. In *Proceedings of the 31st International Conference on Machine Learning*. 595–603.
- [13] Mirella Lapata. 2003. Probabilistic Text Structuring: Experiments with Sentence Ordering. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Sapporo, Japan, 545–552. <https://doi.org/10.3115/1075096.1075165>
- [14] Karen Le. 2015. Narrative and Horticultural Imperative: Predicting Discourse Ability in Traumatic Brain Injury from Cognitive and Communicative Factors. (2015).
- [15] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2017. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4681–4690.
- [16] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130* (2017).
- [17] Lajanugen Logeswaran, Honglak Lee, and Dragomir Radev. 2018. Sentence Ordering and Coherence Modeling using Recurrent Neural Networks. <https://www.aaii.org/ocs/index.php/AAAI/AAAI18/paper/view/17011>
- [18] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. 2015. Unsupervised Learning of Video Representations Using LSTMs. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37 (ICML'15)*. JMLR.org, 843–852. <http://dl.acm.org/citation.cfm?id=3045118.3045209>
- [19] Qing Sun, Stefan Lee, and Dhruv Batra. 2017. Bidirectional Beam Search: Forward-Backward Inference in Neural Sequence Models for Fill-In-The-Blank Image Captioning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [20] Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. 2015. Order matters: Sequence to sequence for sets. *arXiv preprint arXiv:1511.06391* (2015).
- [21] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. [n.d.]. Pointer Networks. In *Advances in Neural Information Processing Systems 28*.
- [22] Jason Weston, Samy Bengio, and Nicolas Usunier. 2011. Wsabie: Scaling Up to Large Vocabulary Image Annotation. <https://www.aaii.org/ocs/index.php/IJCAI/IJCAI11/paper/view/2926>
- [23] Kuo-Hao Zeng, William B Shen, De-An Huang, Min Sun, and Juan Carlos Nieves. 2017. Visual forecasting by imitating dynamics in natural sequences. In *Proceedings of the IEEE International Conference on Computer Vision*. 2999–3008.
- [24] M. Zhou, M. Huang, and X. Zhu. 2019. Story Ending Selection by Finding Hints From Pairwise Candidate Endings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27, 4 (April 2019), 719–729. <https://doi.org/10.1109/TASLP.2019.2893499>