IP Report: Multimodal Procedural Knowledge Learning using WikiHow articles

Anil Batra, Frank Keller
University of Edinburgh

Abstract

Procedural learning with multi-modal *'how-to'* articles is beneficial to enable AI systems with an ability to perform goal oriented tasks. Learning the temporal event structure in procedures through only-text based datasets fails to capture the implicit information among events e.g. missing object of an action. We hypothesize that the visual data is adequate to augment the missing information and extend the text based dataset (Zhang et al., 2020) with visual data. Towards our goal, we study pairwise event ordering with architectures pre-trained on uni and multi modal data. Surprisingly, we find that joining the features from architectures (Resnet-50 + BERT) which are pre-trained on uni-modal data, is superior to state-of-the-art multi-modal architectures (LXMERT and UNITER) towards temporal structure learning. Furthermore, we enhance the event relation learning with an attention mechanism. Our experiments on the extended pairwise step-order dataset shows that our approach benefit in learning the perfect order by 1.67% in comparison to text-only datasets.

1 Introduction

Time is a crucial dimension, apart from three spatial dimensions, for understanding the world dynamics and enable us to learn the evolution of object relationships and states. For this reason, it is essential for AI system to learn the temporal knowledge in uni or multi modal data i.e. text and/or images and benefit various applications such as summarization of documents or videos (Zhang et al., 2016), describing the images/videos (Yao et al., 2015), story understanding (Han et al., 2019), timeline construction (Zhou et al., 2020), question-answering and causal inference (Christiansen et al., 2020). An essential step to learn such temporal knowledge is identifying actions or events performed by objects and relations among different events stretched over time.

Figure 1: (a) Sequential Order of events, (b) Shows the EVENT ordering task as a binary classifier. Visual and Text modality provide complementary context to learn the event order, as in second step of (b) shows that where we need to empty the flavouring packet and what to stir.

Among the different relation types that could exist in the events, we focus on the sequential event order to perform a certain goal oriented task. Such an encoded temporal structure is widely referred as *procedural knowledge*, and defined as the stereotypical activities in our daily life to complete a goal e.g. sequential steps to make maggi. These sequential steps follow the arrow of time i.e. each step followed the next. For example, *'bring the mixture to boil'* can occur only after we have created the mixture i.e. after step *'stir masala and add water'* towards the goal of *making maggi*.

Modeling procedural knowledge with sequential events is challenging with respect to both vision and language understanding. For language, it requires understanding of predicate-argument structure and coreference. The *common sense* information is rarely mentioned in steps explicitly e.g. *'turn on the hob'* is not explicitly mentioned in the goal of *making maggi*, however such an information is implicit from step *'saute garlic'*. For visual understanding, AI systems require to perform recognition (e.g. detecting objects and their attributes) and cognition-level reasoning (e.g. inferring the likely intents about past and future).

Moreover, the appearance, view and state of the objects evolve over time, which might fail the current state-of-the-art recognition models to re-identify the objects. We hypothesize that it is necessary for both the modalities to interact to find the missing information. Consider the multi-modal step *'stir in the flavoring packet'* with visual image as shown in Figure.1, where humans are able to infer the implicit action *'add'* along with its arguments *'flavoring packet contents'* from textual modality and *'into skillet'* from visual modality. Our work is motivated by this complementary relation among events to learn procedural knowledge.

In order to model *procedural knowledge*, the problem is usually formulated as a binary classification to determine whether the given two event pairs are correctly or incorrectly ordered. Recently, Zhang et al. (2020) introduce `STEP-STEP` temporal relations dataset from `WikiHow` on large scale utilizing the textual step information. In this work, we are extending their work and propose to include visual modality in binary formulation of event ordering task. Our intuition is that visual modality provides complementary context to textual information and improve the learning of procedural knowledge. As discussed in our motivational example of identifying the *'skillet'* from visual modality as an argument to implicit *'add'* action help to predict order of the event *'stir in the flavoring packet'* only after adding the vegetables and water to create the mixture. Towards this goal we use transformer style event interaction followed by binary classifier, which is fed with the concatenated ResNet features for images and BERT features for text. This event interaction provides consistent gain in our method. Moreover, to further explore the full order prediction using the binary pair prediction, we utilize brute-force technique to learn the correct order of steps with beam based startegy (Chen et al., 2016). With qualitative analysis, we proposed to ensemble the outputs from our multi-modal architecture and BERT, leading to improve the *perfect match ratio (PMR)* and *Kendal Tau* by 1.67% and 1% respectively.

2 Related Work

In order to model *procedural knowledge* and address its challenges, the research community use statistical and probabilistic techniques. Specifically, the problem is formulated with two major task: temporal event ordering (Modi, 2016; Lin et al., 2020; Zhang et al., 2020) and future event prediction (Chambers and Jurafsky, 2008, 2009; Pichotta and Mooney, 2016b; Lee et al., 2020). In this work we consider the event ordering task.

2.1 Event Ordering

Learning event sequences is introduced in early 1980's by Schank and Abelson (1977) using narrative texts. With the availability of large language corpora, Chambers and Jurafsky (2009) propose count based methods to learn narrative chains i.e. learn co-occurrence count of pairwise events in a large corpus to predict the possible event sequence. The count based models has zero or low probability for unseen event combinations which are missing during training, leading to poor inference at test time. With the success of neural networks, Pichotta and Mooney (2016a) proposed RNN-LSTM based architecture to learn the event sequence by converting the textual event into `verb-argument` structure. Pichotta and Mooney (2016b) further introduced sentence level language modeling to learn the event sequences. Later, Modi (2016) proposed event specific representation via learning individual embedding of `verb` and `arguments` (subject and object). However, most of these work focused on general narrative structure from newswire and literature, except (Modi, 2016) which focused on stereotypical human activities such as *'visiting a doctor'*. In contrast, we focus on goal oriented event sequences for intelligent systems. The closely related work to ours is Zhang et al. (2020), which explore `WikiHow` resources to learn procedural knowledge and introduced `STEP-STEP` temporal relations dataset. In contrast to our work of utilizing the multi-modal information, the focus of their work is solely based on textual information. In addition to text, visual modality is explored in the context of instructional videos available on YouTube to learn the knowledge (Xu et al., 2020), however, computational processing and annotating a video is challenging and expensive. Consequently, we are exploring to learn the knowledge using image-text pairs available on `WikiHow` articles which are efficient to train and scalable.

2.2 Multi-modal Representation Learning

The success of attention mechanism in the NLP community motivated them to pre-train models in multi-modal settings for wide range of downstream tasks, such as visual question answering, visual reasoning and image captioning. Similar to BERT

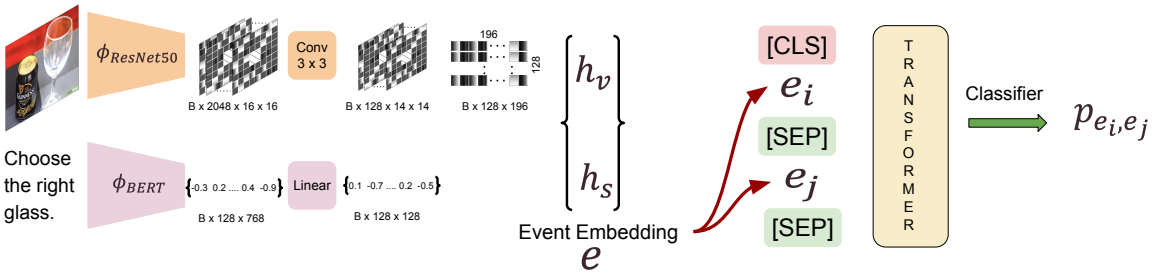


Figure 2: Architecture to learn order of pairwise steps of WikiHow article. We extract the features of muli-modal event (image and text) using pre-trained models (*Resnet50*: Images and *BERT*: Text). Finally, using Transformer based attention we learn event interaction to predict the order between two events.

(Devlin et al., 2018), the common method is to use a single transformer architecture to jointly encode text and image such as VisualBERT (Li et al., 2019), Uniter (Chen et al., 2019) and VL-BERT (Su et al., 2019). Alternatively, ViLBERT (Lu et al., 2019) and LXMERT (Tan and Bansal, 2019) introduced the two-stream architecture, where two transformers are applied to images and text independently, which is fused by a third transformer in a later stage. However, these models typically consume object detection features. In contrast to these multi-modal architectures, we utilize the individual components from uni-modal pre-trained architectures. The equivalent architecture is employed by Kiela et al. (2019) for image-text pair interaction, however, we employed at event level interaction.

3 Method

We are interested in learning the timeline of sequential events to perform a task from real-world web articles such as WikiHow. The key technical challenge is how to learn the full sequential structure of events directly from unstructured and noisy multi-modal data available on web. We take the pairwise ordering approach i.e. learn the order of steps using binary classification. We will first define the problem and how to address it using a pairwise order classification approach. We will then discuss how we learn the relation between two events to correctly order them on the timeline.

3.1 Problem Formulation

The articles in *'WikiHow'* has multiple events and a goal g, with each event having image-text pair. Consider this as dataset $\mathscr{D}$ of $\mathscr{N}$ articles with each article is represented as

$$\begin{aligned}\mathcal{D} &= \{e_1, \ldots, e_i, \ldots, e_m\} \\ &= \{(s_1, v_1), \ldots, (s_i, v_i), \ldots, (s_m, v_m)\}\end{aligned}$$

where e_i represents an event in the article of m steps and each event consist of textual and visual data represented as (s_i, v_i). To learn the procedural knowledge, AI system need to understand the dynamics of objects from the multi-modal data and place each event on the timeline to achieve the goal. Thus, we can consider the problem of identifying the order of each event and localize on the sequential timeline as shown in the Figure.1. Consequently, the ordering task can be defined as to find the correct order $\hat{o}$ of shuffled events s.t.

$$e_{\hat{o}_1} \succ e_{\hat{o}_2} \succ e_{\hat{o}_3} \succ \cdots \succ e_{\hat{o}_n}$$

Cohen et al. (1997); Fürnkranz and Hüllermeier (2003) model the ordering task as pairwise ranking model i.e. predict the order of any two event pair (e_i, e_j). Motivated by this we can model our problem as binary classification such that event e_i precedes e_j or not. Mathematically, we consider N data samples generated from $\mathscr{N}$ articles and one sample is represented as $x = \{(s_i, v_i), (s_j, v_j), y\}$ where $y \in \{0, 1\}$ and learn the order of each sample by learning the function Ψ between the two event representation.

$$\hat{y} = \Psi(\phi(s_i, v_i), \phi(s_j, v_j))$$

where $\hat{y}$ is predicted order and $\phi(.)$ is an encoding function to learn the embedding of multi-modal event, which can modeled using uni or multi modal deep neural architectures. The order function Ψ can be learned by optimizing using simple *Cross-Entropy* loss over N data samples as follows:

$$\mathcal{L}(y, \hat{y}) = \frac{1}{N} \sum_{N} \sum_{i}^{2} y_i \cdot log(\Psi(x_i))$$

3.2 Event Embedding

Many self-supervised learning approaches are explored by NLP and Vision community to learn the

individual or joint representation of both the modalities. In this work, we propose to use highly effective BERT (Devlin et al., 2018) representation for textual step and combine with the power of CNN based ResNet (He et al., 2016) architecture for visual information in an event, as shown in Figure.2. Mathematically, textual ($\mathbf{h_s}$) and visual ($\mathbf{h_v}$) event embeddings are extracted by respective pre-trained models as follows:

$$\mathbf{h_s} = \mathsf{Linear}(\phi_{BERT}(s))$$
$$\mathbf{h_v} = \mathsf{Conv3 \times 3}(\phi_{R50}(v))$$
$$\mathbf{e} = \mathsf{concat}(\mathbf{h_s}, \mathbf{h_v})$$

where $\phi_{BERT} \in \mathbb{R}^{S \times D}$ is word level embedding output from last layer of BERT with S: *sequence length*, D: embedding dimension. And $\phi_{R50} \in R^{C \times H \times W}$ is the output of 3^{rd} encoder layer in ResNet50 architecture with C: *channel length*, $H \times W$: *spatial feature dimension*. These embeddings are further projected to lower dimension space using 3×3 $conv$ layer and $linear$ layer for visual and textual feature respectively. Finally, the event embedding is represented by concatenating the low dimensional outputs $\mathbf{h_s}$ and $\mathbf{h_v}$.

3.3 Event Interaction

In this section, we discuss how to learn the order function $\Psi(.)$ among two event embeddings. Formally, we decompose the order function into two sub functions (1) event interaction function $f(.)$, and (2) linear classification function. The naive approach to learn interaction function is by concatenating the event embedding. The final concatenated long embedding vector provides the opportunity to each modality in an event to interact with each other via linear classification layer. Such an interaction is un-restricted to specifically focus on the identifying the order between two events. Hence, our new Transformer style attention technique focus on learning the interaction with explicit event separation by special token *[SEP]* as shown in Figure.2. Specifically, it takes the multimodal event embedding $\mathbf{e}$ in the BERT style format i.e. $([CLS], \mathbf{e}_1, [SEP], \mathbf{e}_2, [SEP])$ before the final classification layer to predict the binary decision.

3.4 Full Event Order

Inferring the correct order from pairwise comparisons is combinatorial hard problem and is an active research area in statistics (Shah and Wainwright, 2017; Heckel et al., 2018; Gao and Zhang, 2019; Chen et al., 2021). In the current work, we attempted the brute-force approach with beam strategies inspired by the work from Chen et al. (2016) for sentence ordering task. Specifically, following Chen et al. (2016) notation we calculated the score of event order in an article using log-likelihood maximization problem.

$$score(e, o, i, j) = log(p_{e_i, e_j})$$
$$Score(a, o) = \sum_{i=1} \sum_{j=i+1} score(e, o, i, j)$$
$$\hat{o} = \arg\max_{o} Score(a, o)$$

where $score(e, o, i, j)$ represents indicates the score for event pairs (e_i, e_j) with probability p_{e_i, e_j} obtained from classifier output and $Score(a, o)$ indicates the score of all event permutations for an article a. Finally, the predicted order $\hat{o}$ is obtained with maximum score.

4 Experimental Setup

4.1 Dataset

The annotated dataset to acquire the procedural knowledge are restricted to textual domain (Regneri et al., 2010; Chambers and Jurafsky, 2009). Recently, Zhang et al. (2020) introduced the `WikiHow` articles and utilize the textual steps in the article as self-supervision task. Each article consists of multiple methods to perform a task and each method comprises of textual headline, description and visual image. In the current work we consider textual headline and image as a single event to perform the task. However, we focus on the articles with base category of *'Food and Entertaining'*. We use the dataset splits created by Zhang et al. (2020) and exclude the deleted articles or articles with missing images. Furthermore, Zhang et al. (2020) create a set of examples by sampling every adjacent pair steps as candidates in the `WikiHow` article with binary labels (0 - correct order, 1 - incorrect order) and then randomly shuffle the candidates to balance the dataset. Finally, we describe the statistics of the dataset splits in Table.1 and Figure.3 summarizes the distribution of articles per step count for Valid and Test set.

4.2 Evaluation Metrics

We evaluate the performance of our approach for pair wise event order classification using *Accuracy*,

	Train	Valid	Test
WikiHow Articles	1,044	174	174
Number of Event Pairs	27,688	4,742	4,900
Unique Textual Events	10,856	1,788	1,848
Unique Images	11,293	1,817	1,889

Table 1: Dataset Statitscs for Pairwise Event Order task to learn procedures via WikiHow articles.

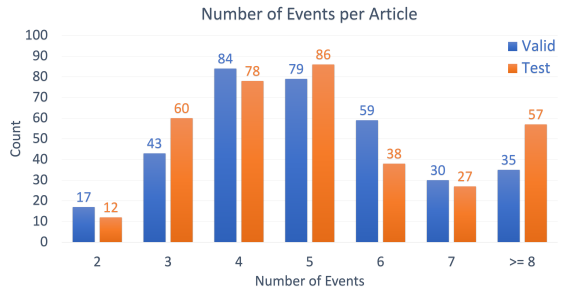


Figure 3: Step Count distribution for valid and test set. It shows that valid and test set has more WikiHow articles with event count of 3,4,5, and 6 per article.

Precision, *Recall* and *F1-score* metrics. Furthermore to infer the full step order of given steps we borrow the ideas from sentence ordering task and use the following metrics.

Perfect Match Ratio (PMR): calculates the percentage of samples for which the entire step sequence was correctly predicted (Che et al., 2019).

$$\tau = \frac{1}{N}\sum_{i}^{N} \mathbb{1}\{\hat{o}_i = o_i^*\}$$

where $\hat{o}_i$ is predicted step order, o_i^* is ground truth step order and N is total number of articles in the dataset.

Kendal Tau (τ): quantifies the distance between the predicted order and the correct order in terms of the number of inversions (Lapata, 2006).

$$\tau_i = 1 - \frac{2 \times I}{\binom{n_i}{2}} \quad (1)$$

where I is the number of step-pairs in the predicted order with incorrect relative order and n_i is the total number of steps in i^{th} WikiHow article.

Longest Common Sequence (LCS): calculates the ratio of longest common step-sequence (Gong et al., 2016) between the predicted order and the given order (consecutive steps are not necessary, and higher is better).

4.3 Baselines

Our goal in the work is to learn procedural knowledge using the multi-modal data, thus we explore the following state-of-the-art neural network architectures based on their modality usage.

BERT It is an attention-based bidirectional language model, introduced by (Devlin et al., 2018). The BERT model is trained end-to-end on a large language-corpus under two tasks: *masked language modelling* and *next sentence prediction*. This pre-training on a large language corpus helps BERT to be very effective for transfer learning on multiple tasks. BERT model can be fine tuned with one text segment A or two text segments (A, B). The one text segment is passed to model with format: $(CLS, w_{A_1}, \ldots, w_{A_n}, SEP)$ and two text segments are passed to model in format: $(CLS, w_{A_1}, \ldots, w_{A_n}, SEP, w_{B_1}, \ldots, w_{B_n}, SEP)$, where w_{A_i}, w_{B_i} are tokens in the text segment and (CLS, SEP) are special tokens. Specifically, one text segment input is employed to learn the token level representation of text segment and two text segment is used to predict whether second text segment follows the first in the source document.

In the current work we utilize the BERT-base model with both input scenarios. Firstly, we consider the two textual steps from the WikiHow article as two text segments from a single document and predict the order of events. Secondly, we learn event representation of each step in the WikiHow article by passing the single textual step to model and finally concatenate the two event representations to learn the event order by linear classification.

ResNet50 It is a traditional convolutional feed-forward network introduced by (He et al., 2016) and is pre-trained on large object recognition dataset *ImageNet* (Deng et al., 2009). In computer vision community, it is the standard practice to initialize the convolutional networks with model pre-trained on *ImageNet* for fine tuning the downstream tasks. Consequently, this model has achieved high performance on different challenging tasks such as image localization, semantic segmentation, and object detection tasks. Following the similar trend, we use ResNet50 variant to learn the visual features of events int the WikiHow article and learn event order with linear classification layer, which is fed by concatenating the two events.

LXMERT It is proposed by (Tan and Bansal, 2019) to solve different Vision and Language tasks

Method	Accuracy		Precision		Recall		F1	
	Concat	*Tx*	*Concat*	*Tx*	*Concat*	*Tx*	*Concat*	*Tx*
Bert + NSP (Devlin et al., 2018)	0.7639	-	0.7504	-	0.7693	-	0.7598	-
Bert (Devlin et al., 2018)	0.7381	0.7463	0.7427	0.7286	0.7340	0.7533	0.7383	0.7407
ResNet50 (He et al., 2016)	0.5240	0.5807	0.4777	0.5550	0.5236	0.5825	0.4996	0.5684
Lxmert (Tan and Bansal, 2019)	0.7056	0.7031	0.6895	0.7161	0.7102	0.7121	0.6997	0.7141
Uniter (Chen et al., 2019)	0.7123	0.7157	0.6486	0.6813	0.7409	0.7294	0.6917	0.7046
ResNet50 + Bert	0.7479	**0.7488**	**0.7441**	0.7395	0.7478	**0.7515**	**0.7459**	0.7454

Table 2: Comparison of step pair classification to learn the order between two steps. It shows the improvement in metrics by fusing the features (either $Concat$ or $Transformer(Tx)$) is from unimodal pre-trained neural architectures. **BERT + NSP**: Using the BERT Pretrained model with next-sentence-prediction task. In the current, we consider the **BERT** model as our baseline with input of one textual event due to unavailability of pre-trained multi-modal architecture with similar task for fair comparison.

by exploiting the attention based transformers. LXMERT has two individual feature learning encoders for vision and language, and has an additional transformer based cross attention module to learn the joint representation of two modalities. The visual encoder in the model built upon the output of (Anderson et al., 2018) i.e. use the features of detected objects or regions as the input embeddings of images to LXMERT. In the current work we explore the model to jointly predict the event order in the `WikiHow` article.

UNITER It is BERT style pre-trained model to learn the joint embeddings of text and images (Chen et al., 2019). It is single stream architecture with a new pre-training task of word region alignment in contrast to BERT. This model follows the BERT like input format: $(CLS, v_1, \ldots, v_m, SEP, w_1, \ldots, w_n, SEP)$ where v_i, w_i are visual and word tokens. Similar to LXMERT, it considers the features of detected objects in images as input to the network. We fine tune the model to obtain the multimodal event embeddings to learn the order of events.

4.4 Implementation Details

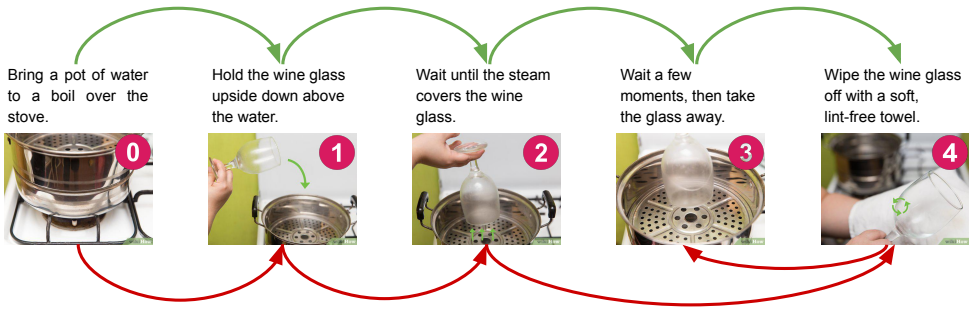
We resize the image to 320×320 and take random crops of size 256×256 from the image of an event or step in `WikiHow` article during training and resize the images to 256×256 at inference time. To improve the generalization of network, random horizontal flip, and random crop is employed as data augmentation to convolutional based models. We train the joint network with a batch size of 36 for 30 epochs using 4 GPUs (GTX 1080; 11GB VRAM). We optimize the models with AdamW optimizer having initial learning rate of 5e-6 with step scheduler having drop factor of 10 at epochs 20, 25. The embedding of each modality is reduced to 128 before classification to fit the model on GPU with batch size of 9 per GPU. The textual modality and detected object features are downsampled by linear projection layer. However, in convolutional networks the visual features are downsampled by convolutional kernel of size 3×3 having 128 channel width. For transformer based fusion of two event embeddings, we use 6 attention heads with the depth of 6 layers. Our models are implemented in PyTorch.

5 Results and Discussion

We aim to investigate the role of visual information to learn the procedural knowledge from `WikiHow` articles, i.e. in our problem formulation we are interested to correctly identify the order between two events of an article using multimodal information and infer the full order of events from pairwise comparison. Towards this goal, firstly we perform comparison with baseline models and finally, perform analysis on results and discuss our reflection on the failure articles.

5.1 Comparison with Baseline

We choose strong baseline models based on the input modality type to the neural network and report the results in Table.2. We fine tune the BERT-base model with two tasks i.e. learning the event embedding and directly identifying the event order (similar to next sentence prediction). Interestingly, we observe that BERT with input format of next sentence prediction task perform better than the fine tuning the event embedding followed by classifying the order between two events using simple concatenation. Our hypothesis on the success of BERT+NSP is due to large pre-training on sentences which has predicate-argument structure similar to two events in our scenario. However, such



(a) Goal: How to Clean Wine Glasses

(b) Goal: How to Serve and Enjoy Guinness

Figure 4: Qualitative comparison of full event inference of WikiHow article. **Green Arrows**: Represents the correct sequential order by our method of using multi- modal (*Resnet50* and *BERT*) features. **Red Arrows**: Represents the order predicted by text (*BERT*) features. Both models use our proposed Transformer based interaction.

a large pre-trained model using multimodal data with next sentence prediction task is unavailable due to lack of data. Consequently, we consider fine tuning the BERT to learn event embedding with one textual segment i.e. one textual event from the article for fair comparison with multi-modal data.

To study the role of images, we fine tune the image-only model (ResNet50) and our results significantly dropped suggesting learning the order between images is hard. Furthermore, we fine tune the multi-modal architectures (LXMERT and UNITER) trained to learn better vision-language representation and observe the improvement on image-only model. However, the pairwise metrics still lacks far behind the text only model i.e. BERT. Our hypothesis for the failure of such models is out-of-domain images data, additionally claimed by Hendrycks et al. (2020). Moreover, WikiHow articles contains *'cartoons'*, *'drawings'*, and *'graphics'* etc. in contrast to image based models are pre-trained with wild and outdoor images. However, Kiela et al. (2019) shows that ImageNet pre-trained models can outperform the multi-modal architectures to learn joint embedding of image-text data by employing transformer style training. We compare the baseline architectures with proposed event interaction using the transformer with minimal modifications due GPU memory constraints. The results in Table.2 under the column *Tx* shows consistent and marginal improvement in the pairwise metrics.

Infer Full Order: We compare the baseline models to infer the full order of events in an article from pairwise comparison using Beam strategy. The results in Table.3 shows that our proposed event interaction using transformer based attention mechanism improve the metrics consistently. However, the gain in our multi-modal (R50+BERT) is competitive with text only model. We hypothesize that either one of the modality enhances the noise in the multi-modal representation towards the learning of order among events. Towards this hypothesis, we

Method	PMR		LCS		Kendal Tau	
	Concat	*Tx*	*Concat*	*Tx*	*Concat*	*Tx*
Bert	22.62	25.14	69.36	69.84	0.4901	0.5062
Resnet50	7.54	9.21	55.55	57.46	0.0635	0.1761
Lxmert	18.43	19.27	65.66	66.56	0.4104	0.4240
Uniter	15.36	16.48	62.22	65.29	0.3192	0.3818
R50+Bert	23.46	24.58	70.21	69.63	0.5125	0.5053
Ensemble	22.91	**26.81**	**70.42**	69.73	0.5159	**0.5162**

Table 3: Comparison of full step order inference by Top1 - Beam Search (Chen et al., 2016) on pair classification predictions using the beam size of 128. ***Ensemble***: It combines the inference output of our multimodal (R50+Bert) and Bert output by picking the Top1 beam prediction with highest probability.

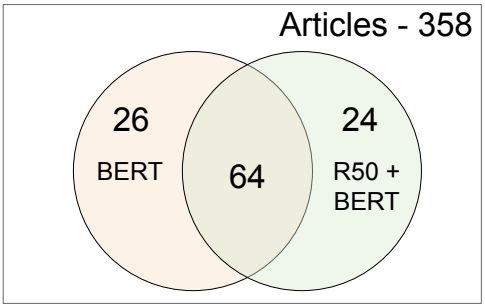


Figure 5: Venn Diagram visualizing the number of articles correctly ordered by BERT and ours (R50+BERT). It shows that both models able to correctly order 64 articles, however they differ on equal number of articles leading to similar *PMR* in Table.3 under the column *Tx*.

visualize the number of articles correctly ordered by both models via a venn diagram, as shown in Figure.5. This shows that visual information benefits additional distinct 24 articles to correctly identify the order where only-text based model (BERT) failed. However, the visual information adds noise in joint learning of another set of 26 articles and in-correctly order the steps, in contrast to correctly ordered by BERT. To address this gap, we propose to ensemble the inference output of text-only and our model. Specifically, we prefer the order with highest probability for Top-1 beam prediction i.e. pick the order from R50+BERT if the probability of Top-1 beam order is greater than the BERT. The last row in Table.3 shows that ensemble technique improves *PMR* and *Kendal Tau* metrics, which suggest that both models predict order correctly for different articles.

5.2 Discussion

The qualitative example of our approach is shown in Figure.4 contrasting with BERT. In the first example (see Figure.4(a)), the visual cues from similar objects i.e. pot and inverted glass in steps 3 & 4, benefit the multi-modal architecture to correctly order the step for the article. Likewise, in example Figure.4(b) the progressive increase of liquid in the glass (i.e. from step 0,1,2) provide a signal to the model to identify the correct order. Our hypothesis for the failure of BERT is `verb-argument` structure as in the first example the action or verb is similar for steps 2 & 3, enforcing BERT to prefer different `verb`. In the example Figure.4(b), the BERT failed due to bias in the dataset, as we observe that `pour` and `add` verb occurs in the initial steps in comparison to `choose` and `serve` in the final steps. The additional positive and failure examples are provide in Appendix (refer Figure.6 & 7) with following observations.

1. We observe that the visual information is useful only if there is gradual change in the object dynamics or visual appearance along the timeline of steps, as shown in 6(b), (c).
2. The model fail to focus on the specific object to correctly order the steps if the visual images are enriched with number of objects (see 7(a)).
3. As discussed earlier, ResNet failed to learn better visual features for cartoon images due to pre-training on natural images, which hinder the sequential order learning (see Figure.7(c)).

6 Conclusion

In this work, we investigate the role of visual information to learn the procedures using pairwise comparison of steps of `WikiHow` articles. Towards this goal, we propose attention based mechanism to perform interaction among multi-modal steps embedding. Moreover, our architecture consists of components that are pre-trained individually as unimodal tasks, surpassing the performance of using state-of-the-art multi-modal architectures. Additionally, we conduct analysis on the results and observe that the irrelevant visual images add noise to multi-modal data leading to restricted the learning of procedures. Finally, we propose ensemble approach to learn the procedures by taking benefits of both worlds.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610.

Wanxiang Che, Longxu Dou, Yang Xu, Yuxuan Wang, Yijia Liu, and Ting Liu. 2019. HIT-SCIR at MRP

2019: A unified pipeline for meaning representation parsing via efficient training and effective encoding. In *Proceedings of the Shared Task on Cross-Framework Meaning Representation Parsing at the 2019 Conference on Natural Language Learning*, pages 76–85, Hong Kong. Association for Computational Linguistics.

Pinhan Chen, Chao Gao, and Anderson Y Zhang. 2021. Optimal full ranking from pairwise comparisons. *arXiv preprint arXiv:2101.08421*.

Xinchi Chen, Xipeng Qiu, and Xuanjing Huang. 2016. Neural sentence ordering. *arXiv preprint arXiv:1607.06952*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Uniter: Learning universal image-text representations.

Rune Christiansen, Matthias Baumann, Tobias Kuemmerle, Miguel D Mahecha, and Jonas Peters. 2020. Towards causal inference for spatio-temporal data: Conflict and forest loss in colombia. *arXiv preprint arXiv:2005.08639*.

William W Cohen, Robert E Schapire, and Yoram Singer. 1997. Learning to order things. *Advances in neural information processing systems*, 10:451–457.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Johannes Fürnkranz and Eyke Hüllermeier. 2003. Pairwise preference learning and ranking. In *European conference on machine learning*, pages 145–156. Springer.

Chao Gao and Anderson Y Zhang. 2019. Iterative algorithm for discrete structure recovery. *arXiv preprint arXiv:1911.01018*.

Jingjing Gong, Xinchi Chen, Xipeng Qiu, and Xuanjing Huang. 2016. End-to-end neural sentence ordering using pointer network. *arXiv preprint arXiv:1611.04953*.

Rujun Han, Mengyue Liang, Bashar Alhafni, and Nanyun Peng. 2019. Contextualized word embeddings enhanced event temporal relation extraction for story understanding. *arXiv preprint arXiv:1904.11942*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Reinhard Heckel, Max Simchowitz, Kannan Ramchandran, and Martin Wainwright. 2018. Approximate ranking from pairwise comparisons. In *International Conference on Artificial Intelligence and Statistics*, pages 1057–1066. PMLR.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. 2020. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*.

Mirella Lapata. 2006. Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, 32(4):471–484.

I-Ta Lee, María Leonor Pacheco, and Dan Goldwasser. 2020. Weakly-supervised modeling of contextualized event embedding for discourse relations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4962–4972.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Shih-Ting Lin, Nathanael Chambers, and Greg Durrett. 2020. Conditional generation of temporally-ordered event sequences. *arXiv preprint arXiv:2012.15786*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*.

Ashutosh Modi. 2016. Event embeddings for semantic script modeling. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 75–83.

Karl Pichotta and Raymond Mooney. 2016a. Statistical script learning with recurrent neural networks. In *Proceedings of the Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods*, pages 11–16.

Karl Pichotta and Raymond J. Mooney. 2016b. Using sentence-level LSTM language models for script inference. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 279–289, Berlin, Germany. Association for Computational Linguistics.

Michaela Regneri, Alexander Koller, and Manfred Pinkal. 2010. Learning script knowledge with web experiments. In *Proceedings of the 48th Annual*

Meeting of the Association for Computational Linguistics, pages 979–988, Uppsala, Sweden. Association for Computational Linguistics.

Roger C. Schank and Robert P. Abelson. 1977. *Scripts, plans, goals, and understanding : an inquiry into human knowledge structures*. Hillsdale, N.J. : Lawrence Erlbaum Associates.

Nihar B Shah and Martin J Wainwright. 2017. Simple, robust and optimal ranking from pairwise comparisons. *The Journal of Machine Learning Research*, 18(1):7246–7283.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

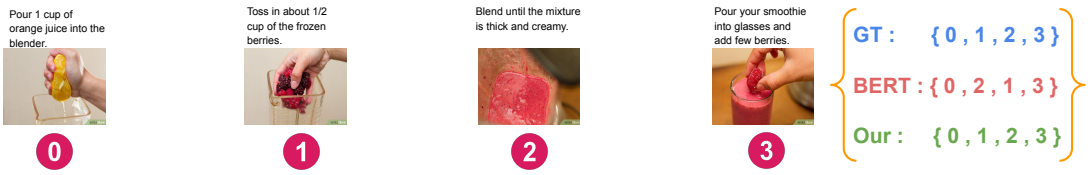
Frank F Xu, Lei Ji, Botian Shi, Junyi Du, Graham Neubig, Yonatan Bisk, and Nan Duan. 2020. A benchmark for structured procedural knowledge extraction from cooking videos. *arXiv preprint arXiv:2005.00706*.

Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*, pages 4507–4515.

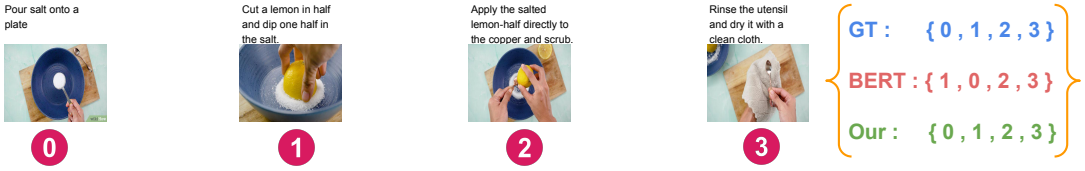
Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. 2016. Video summarization with long short-term memory. In *European conference on computer vision*, pages 766–782. Springer.

Li Zhang, Qing Lyu, and Chris Callison-Burch. 2020. Reasoning about goals, steps, and temporal ordering with WikiHow. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4630–4639, Online. Association for Computational Linguistics.

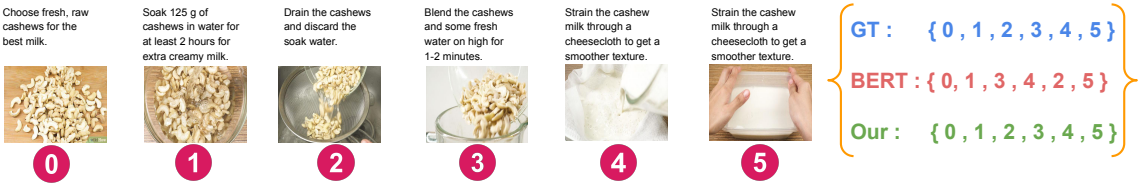
Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2020. Temporal reasoning on implicit events from distant supervision. *arXiv preprint arXiv:2010.12753*.



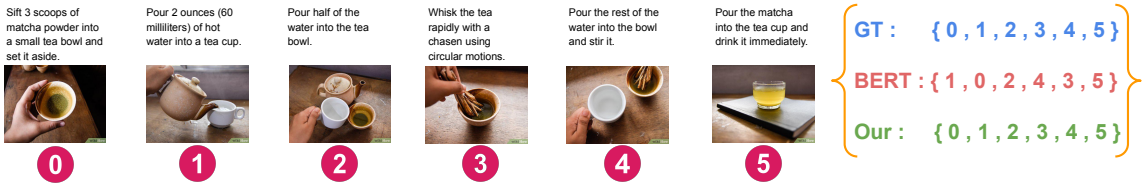
(a) Goal: How to Make a Berry Smoothie



(b) Goal: How to Clean Copper Utensils



(c) Goal: How to Make Cashew Milk



(d) Goal: How to Make Matcha Tea

Figure 6: Additional Qualitative examples of full event inference of `WikiHow` article, depicting the benefit of multi-modal data. In all the example articles, visual images has progressive change in object or object dynamics, leading to correctly order the events.

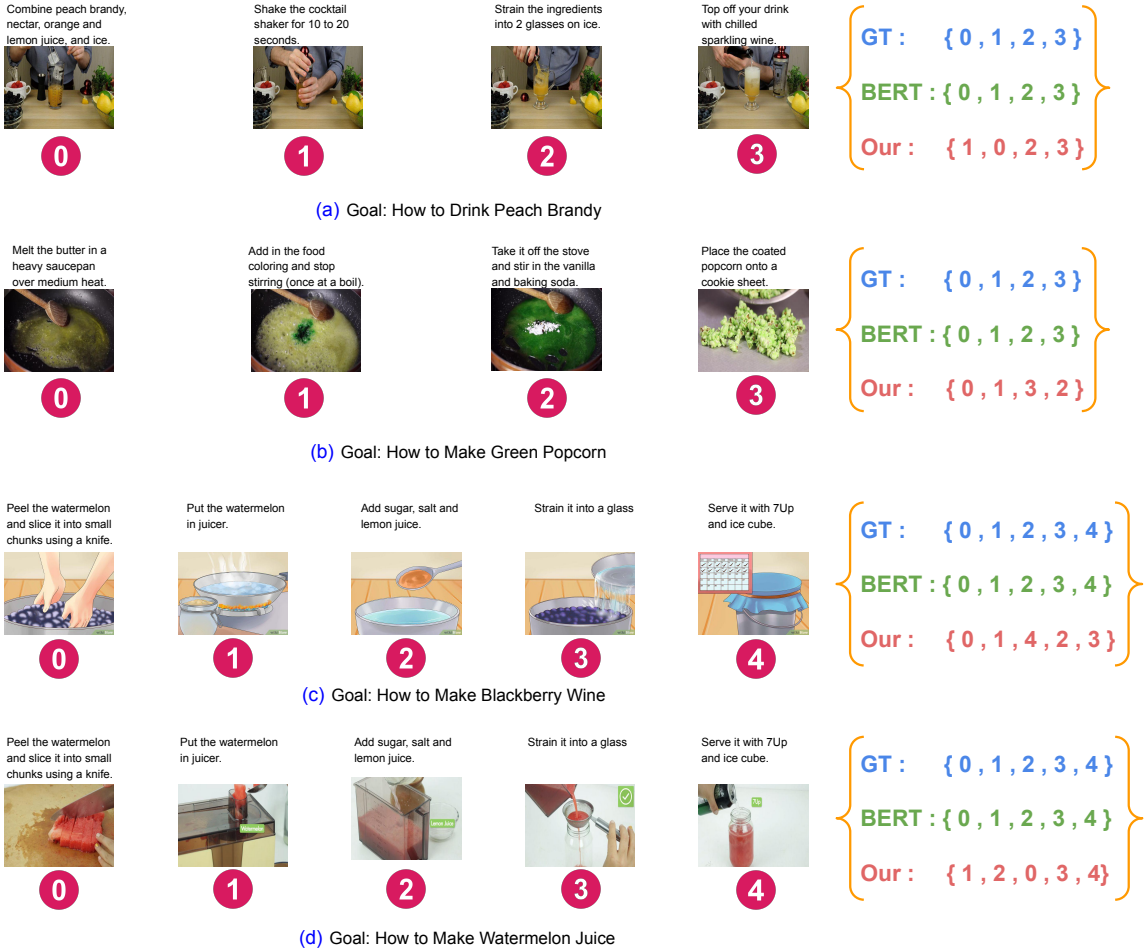


Figure 7: Qualitative examples of full event inference of `WikiHow` article, where visual information add noise in the learning process. **(a)** It is hard to emphasis on specific object dynamics for event order in the enriched visual images with variety of objects, **(b)** Color texture of popcorn in step 3 is gradual change from step 1, adding noise in learning. **(c)** From the sequence of cartoon images the extracted features from ResNet50 (pre-trained on natural images) add noise to correctly order the events. **(d)** The object in first visual image is different from others leading to incorrect order, however objects in event 1&2 and 3&4 has similar objects.