Variational Recurrent Sequence-to-Sequence Retrieval for Stepwise Illustration

Vishwash Batra[1](✉), Aparajita Haldar[1], Yulan He[1], Hakan Ferhatosmanoglu[1], George Vogiatzis[2], and Tanaya Guha[1]

[1] University of Warwick, Coventry CV4 7AL, UK
{v.batra,aparajita.haldar,yulan.he}@warwick.ac.uk
[2] Aston University, Birmingham B4 7ET, UK

Abstract. We address and formalise the task of *sequence-to-sequence (seq2seq) cross-modal retrieval*. Given a sequence of text passages as query, the goal is to retrieve a sequence of images that best describes and aligns with the query. This new task extends the traditional cross-modal retrieval, where each image-text pair is treated independently ignoring broader context. We propose a novel *variational recurrent seq2seq (VRSS) retrieval model* for this seq2seq task. Unlike most cross-modal methods, we generate an image vector corresponding to the latent topic obtained from combining the text semantics and context. This synthetic image embedding point associated with every text embedding point can then be employed for either image generation or image retrieval as desired. We evaluate the model for the application of *stepwise illustration* of recipes, where a sequence of relevant images are retrieved to best match the steps described in the text. To this end, we build and release a new *Stepwise Recipe* dataset for research purposes, containing 10K recipes (sequences of image-text pairs) having a total of 67K image-text pairs. To our knowledge, it is the first publicly available dataset to offer rich semantic descriptions in a focused category such as food or recipes. Our model is shown to outperform several competitive and relevant baselines in the experiments. We also provide qualitative analysis of how semantically meaningful the results produced by our model are through human evaluation and comparison with relevant existing methods.

Keywords: Semantics · Multimodal datasets · Sequence retrieval

1 Introduction

There is growing interest in cross-modal analytics and search in multimodal data repositories. A fundamental problem is to associate images with some corresponding descriptive text. Such associations often rely on semantic understanding, beyond traditional similarity search or image labelling, to provide human-like visual understanding of the text and reflect abstract ideas in the image.

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12035, pp. 50–64, 2020.
https://doi.org/10.1007/978-3-030-45439-5_4

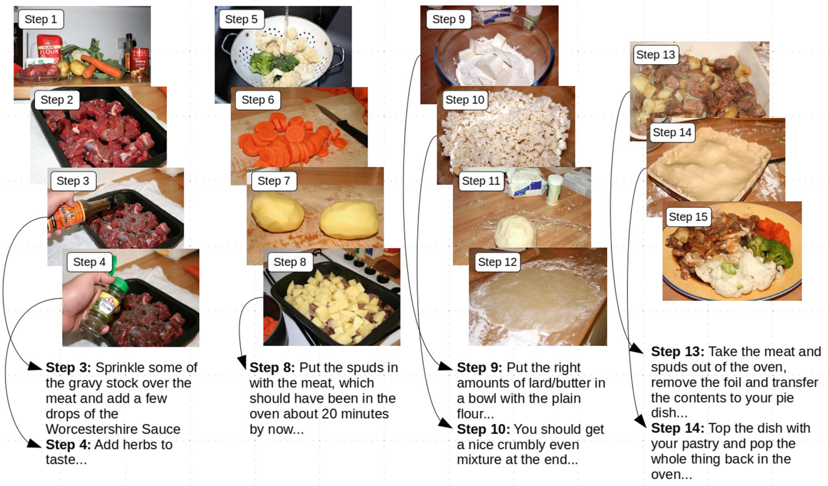

Step 1
Step 2
Step 3
Step 4
Step 5
Step 6
Step 7
Step 8
Step 9
Step 10
Step 11
Step 12
Step 13
Step 14
Step 15
Step 3: Sprinkle some of the gravy stock over the meat and add a few drops of the Worcestershire Sauce
Step 4: Add herbs to taste...
Step 8: Put the spuds in with the meat, which should have been in the oven about 20 minutes by now...
Step 9: Put the right amounts of lard/butter in a bowl with the plain flour...
Step 10: You should get a nice crumbly even mixture at the end...
Step 13: Take the meat and spuds out of the oven, remove the foil and transfer the contents to your pie dish...
Step 14: Top the dish with your pastry and pop the whole thing back in the oven...

Fig. 1. Stepwise Recipe illustration example showing a few text recipe instruction steps alongside one full sequence of recipe images. Note that retrieval of an accurate illustration of Step 4, for example, depends on previously acquired context information.

Cross-modal retrieval systems must return outputs of one modality from a data repository, while a different modality is used as the input query. The multimodal repository usually consists of paired objects from two modalities, but may be labelled or unlabelled. Classical approaches to compare data across modalities include canonical correlation analysis [12], partial least squares regression [28], and their numerous variants. More recently, various deep learning models have been developed to learn shared embedding spaces from paired image-text data, either unsupervised, or supervised using image class labels. The deep models popularly used include deep belief networks [23], correspondence autoencoders [9], deep metric learning [13], and convolutional neural networks (CNNs) [33]. With all these models it is expected that by learning from pairwise aligned data, the common representation space will capture semantic similarities across modalities.

Most such systems, however, do not consider sequences of related data in the query or result. In traditional image retrieval using text queries, for example, each image-text pair is considered in isolation ignoring any broader 'context'. A context-aware image-from-text retrieval model must look at pairwise associations and also consider sequential relationships. Such *sequence-to-sequence (seq2seq) cross-modal retrieval* is possible when contextual information and semantic meaning are both encoded and used to inform the retrieval step.

For *stepwise recipe illustration*, an effective retrieval system must identify and align a set of relevant images corresponding to each step of a given text sequence of recipe instructions. More generally, for the task of automatic *story picturing*, a series of suitable images must be chosen to illustrate the events and abstract concepts found in a sequential text taken from a story. An example of the instruction steps and illustrations of a recipe taken from our new Stepwise Recipe dataset is shown in Fig. 1.

In this paper, we present a variational recurrent learning model to enable seq2seq retrieval, called Variational Recurrent Sequence-to-Sequence (VRSS) model. VRSS produces a joint representation of the image-text repository, where the semantic associations are grounded in context by making use of the sequential nature of the data. Stepwise query results are then obtained by searching this representation space. More concretely, we incorporate the global context information encoded in the entire text sequence (through the attention mechanism) into a variational autoencoder (VAE) at each time step, which converts the input text into an image representation in the image embedding space. To capture the semantics of the images retrieved so far (in a story/recipe), we assume the prior of the distribution of the topic given the text input follows the distribution conditional on the latent topic from the previous time step. By doing so, our model can naturally capture sequential semantic structure.

Our main contributions can be summarised below:

- We formalise the task of *sequence-to-sequence (seq2seq) retrieval* for stepwise illustration of text.
- We propose a new *variational recurrent seq2seq (VRSS) retrieval model* for seq2seq retrieval, which employs temporally-dependent latent variables to capture the sequential semantic structure of text-image sequences.
- We release a new *Stepwise Recipe* dataset ($10K$ recipes, $67K$ total image-text pairs) for research purposes, and show that VRSS outperforms several cross-modal retrieval alternatives on this dataset, using various performance metrics.

2 Related Work

Our work is related to: cross-modal retrieval, story picturing, variational recurrent neural networks, and cooking recipe datasets.

Cross-Modal Retrieval. A number of pairwise-based methods over the years have attempted to address the cross-modal retrieval problem in different ways, such as metric learning [26] and deep neural networks [32]. For instance, an alignment model [16] was devised that learns inter-modal correspondences using MS-COCO [19] and Flickr-30k [25] datasets. Other work [18] proposed unifying joint image-text embedding models with multimodal neural language models, using an encoder-decoder pipeline. A later method [8] used hard negatives to improve their ranking loss function, which yielded significant gains in retrieval performance. Such systems focus only on isolated image retrieval when given a text query, and do not address the seq2seq retrieval problem that we study here.

In a slight variation [2], the goal was to retrieve an image-text multimodal unit when given a text query. For this, they proposed a gated neural architecture to create an embedding space from the query texts and query images along with the multimodal units that form the retrieval results set, and then performed semantic matching in this space. The training minimized structured hinge loss, and there was no sequential nature to the data used.

Story Picturing. An early story picturing system [15] retrieved landscape and art images to illustrate ten short stories based on key terms in the stories and image descriptions as well as a similarity linking of images. The idea was pursued further with a system [11] for helping people with limited literacy to read, which split a sentence into three categories and then retrieved a set of explanatory pictorial icons for each category.

To our knowledge, an application [17] that ranks and retrieves image sequences based on longer text paragraphs as queries was the first to extend the pairwise image-text relationship to matching image sequences with longer paragraphs. They employed a structural ranking support vector machine with latent variables and used a custom-built Disneyland dataset, consisting of blog posts with associated images as the parallel corpus from which to learn joint embeddings. We follow a similar approach, creating our parallel corpus from sequential stepwise cooking recipes rather than unstructured blog posts, and design an entirely new seq2seq model to learn our embeddings.

The Visual Storytelling Dataset (VIST) [14] was built with a motivation similar to our own, but for generating text descriptions of image sequences rather than the other way around. Relying on human annotators to generate captions, VIST contains sequential image-text pairs with a focus on abstract visual concepts, temporal event relations, and storytelling. In our work, we produce a similar sequenced dataset in a simple, automated manner.

A recent joint sequence-to-sequence model [20] learned a common image-text semantic space and generated paragraphs to describe photo streams. This bidirectional attention recurrent neural network was evaluated on both the above datasets. Despite being unsuitable for our inverse problem, VIST has also been used for retrieving images when given text, in work related to ours. In an approach called Coherent Neural Story Illustration (CNSI), an encoder-decoder network [27] was built to first encode sentences using a hierarchical two-level sentence-story gated recurrent unit (GRU), and then sequentially decode into a corresponding sequence of illustrative images. A previously proposed coherence model [24] was used to explicitly model co-references between sentences.

Variational Recurrent Neural Networks. Our model is partly inspired by the variational recurrent neural network (VRNN) [6], which introduces latent random variables into the hidden state of an RNN by combining it with a variational autoencoder (VAE). They showed that using high level latent random variables, VRNN can model the variability observed in structured sequential data such as natural speech and handwriting. VRNN has recently been applied to other sequential modelling tasks such as machine translation [31].

Our proposed VRSS model introduces temporally-dependent latent variables to capture the sequential semantic structure of text/image sequences. Different from existing approaches, we take into account the global context information encoded in the entire query sequence. We use VAE for cross-modal generation by converting the text into a representation in the image embedding space instead of using it to reconstruct the text input. Finally, we use the max-margin hinge loss to enforce similarity between text and paired image representations.

Cooking Recipe Datasets. The first attempt at automatic classification of food images was the Food-101 dataset [3] having $101K$ images across 101 categories. Since then, the new Recipe1M dataset [29] gained wide attention, which paired each recipe with several images to build a collection of $13M$ food images for $1M$ recipes. Recent work [4] proposed a cross-modal retrieval model that aligns Recipe1M images and recipes in a shared representation space. As this dataset does not offer any sequential data for stepwise illustration, this association is between images of the final dish and the corresponding entire recipe text. Our Stepwise Recipe dataset, by comparison, provides an image for each instruction step, resulting in a sequence of image-text pairs for each recipe.

In [5] they release a dataset of sequenced image-text pairs in the cooking domain, with focus on text generation conditioned on images. RecipeQA [34] is another popular dataset, used for multimodal comprehension and reasoning, with 36K questions about the 20K recipes and illustrative images for each step of the recipes. Recent work [1] used it to analyse image-text coherence relations, thereby producing a human-annotated corpus with coherence labels to characterise different inferential relationships. The RecipeQA dataset reveals associations between image-text pairs much like our Stepwise Recipe dataset, and we therefore utilise it to augment our own dataset.

3 Stepwise Recipe Dataset Construction

We construct the *Stepwise Recipe* dataset, composed of illustrated, step-by-step recipes from three websites[1]. Recipes were automatically web-scraped and cleaned of HTML tags. The information about data and scripts will be made available on GitHub[2]. The construction of such an image-text parallel corpus has several challenges as highlighted in previous work [17]. The text is often unstructured, without information about the canonical association between image-text pairs. Each image is semantically associated with some portion of the text in the same recipe, and we assume that the images chosen by the author to augment the text are semantically meaningful. We thus perform text segmentation to divide the recipe text and associate segments with a single image each.

We perform text-based filtering [30] to ensure text quality: (1) descriptions should have a high unique word ratio covering various part-of-speech tags, therefore descriptions with high noun ratio are discarded; (2) descriptions with high repetition of tokens are discarded; and (3) some predefined boiler-plate prefix-suffix sequences are removed. Our constructed dataset consists of about 2K recipes with 44K associated images.

[1] simplyrecipes.com, visualrecipes.com, olgasflavorfactory.com.

[2] https://github.com/vishwerine/StepRecipe.

Furthermore, we augment our parallel corpus using similarly filtered RecipeQA data [34], which contains images for each step of the recipes in addition to visual question answering data. The final dataset contains over 10K recipes in total and 67K images.

4 Variational Recurrent Seq2seq (VRSS) Retrieval Model

The seq2seq retrieval task is formalised as follows: given a sequence of text passages, $\boldsymbol{x} = \{x_1, x_2, ..., x_T\}$, retrieve a sequence of images $\boldsymbol{i} = \{i_1, i_2, ..., i_T\}$ (from a data repository) which best describes the semantic meanings of the text passages, i.e., $p(\boldsymbol{i}|\boldsymbol{x}) = \prod_{t=1}^{T} p(i_t|\boldsymbol{x}, i_{<t})$. The training set (e.g., recipes or stories) is $S = \{S^1, S^2, \cdots S^N\}$, where each S^n consists of a sequence of images and their associated text. Each such sequence $S^n = \{(x_1^n, i_1^n), (x_2^n, i_2^n), \cdots, (x_{|S^n|}^n, i_{|S^n|}^n)\}$ is paired element-wise where each text sequence $\boldsymbol{x^n} = \{x_1^n, x_2^n, ..., x_T^n\}$ and each image sequence $\boldsymbol{i^n} = \{i_1^n, i_2^n, ..., i_T^n\}$.

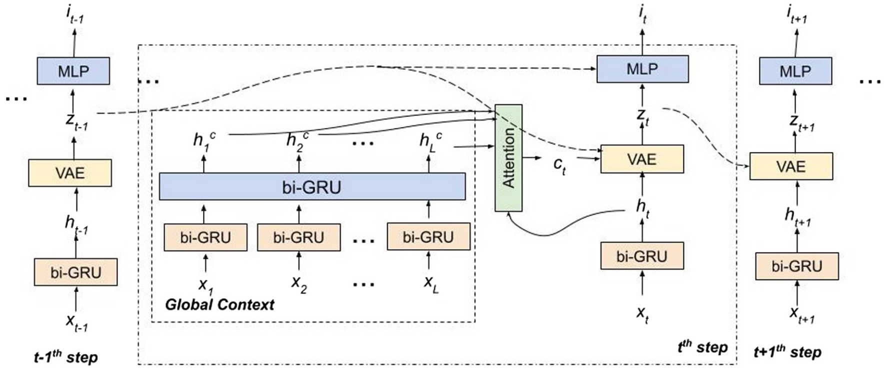


Fig. 2. Variational Recurrent Sequence-to-Sequence (VRSS) model architecture.

We address the seq2seq retrieval problem by considering three aspects: (1) encoding the contextual information of text passages; (2) capturing the semantics of the images retrieved (in a story/recipe); and (3) learning the relatedness between each text passage and its corresponding image.

It is natural to use RNNs to encode a sequence of text passages. Here, we encode a text sequence using a bi-directional GRU (bi-GRU). Given a text passage, we use the attention mechanism to capture the contextual information of the whole recipe. We map the text embedding into a latent topic z_t by using a VAE. In order to capture the semantics of the images retrieved so far (in a story/recipe), we assume the prior of the distribution of the topic given the text input follows a distribution conditional on the latent topic z_{t-1} from the previous step. We decode the corresponding image vector i_t conditional on the latent topic, to learn the relatedness between text and image with a multi-layer perceptron and obtain a synthetic image embedding point generated from its associated text embedding point. Our proposed *Variational Recurrent Seq2seq (VRSS) model* is illustrated in Fig. 2.

Below, we describe each of the main components of the VRSS model.

Text Encoder. We use a bi-GRU to learn the hidden representations of the text passage (e.g. one recipe instruction) in the forward and backward directions. The two learned hidden states are then concatenated to form the text segment representation $\{x_t = [\overrightarrow{h_T}, \overleftarrow{h_T}]\}$. To encode a sequence of such text passages (e.g. one recipe), a hierarchical bi-GRU is used which first encodes each text segment and subsequently combines them.

Image Encoder. To generate the vector representation of an image, we use the pre-trained modified ResNet50 CNN [22]. In experiments, this model produced a well distributed feature space when trained on the limited domain, namely food related images. This was verified using t-SNE visualisations [21], which showed less clustering in the generated embedding space as compared to embeddings obtained from models pre-trained on ImageNet [7].

Incorporating Context. To capture global context, we feed the bi-GRU encodings into a top level bi-GRU. Assuming the hidden state output of each text passage x_l in the global context is h_l^c, we use an attention mechanism to capture its similarity with the hidden state output of the t^{th} text passage h_t as $\alpha_l = \text{softmax}(h_t^T W h_l^c)$. The context vector is encoded as the combination of L text passages weighted by the attentions as $c_t = \sum_{l=1}^{L} \alpha_l h_l^c$. This ensures that any given text passage is influenced more by others that are semantically similar.

Latent Topic Modeling. At the t^{th} step text x_t of the text sequence, the bi-GRU output h_t is combined with the context c_t and fed into a VAE to generate the latent topic z_t. Two prior networks f_{μ_θ} and f_{Σ_θ} define the prior distribution of z_t conditional on the previous z_{t-1}. We also define two inference networks f_{μ_ϕ} and f_{Σ_ϕ} which are functions of h_t, c_t, and z_{t-1}:

$$p_\theta(z_t|\boldsymbol{z}_{<t}, \boldsymbol{x}_{<t}) = \mathcal{N}(z_t|f_{\mu_\theta}(z_{t-1}), f_{\Sigma_\theta}(z_{t-1})) \tag{1}$$

$$q_\phi(z_t|\boldsymbol{z}_{<t}, \boldsymbol{x}_{\leq t}) = \mathcal{N}(z_t|f_{\mu_\phi}(z_{t-1}, h_t, c_t), f_{\Sigma_\phi}(z_{t-1}, h_t, c_t)) \tag{2}$$

Unlike the typical VAE setup where the text input x_t is reconstructed by generation networks, here we generate the corresponding image vector i_t. To generate the image vector conditional on z_t, the generation networks are defined which are also conditional on z_{t-1}:

$$p_\varphi(i_t|\boldsymbol{z}_{\leq t}, \boldsymbol{x}_{\leq t}) = \mathcal{N}(i_t|f_{\mu_\varphi}(z_{t-1}, z_t), f_{\Sigma_\varphi}(z_{t-1}, z_t)) \tag{3}$$

The generation loss for image i_t is then:

$$\begin{aligned}\mathcal{L}_{recons.}(i_t) = \mathbb{E}_{q(\boldsymbol{z}_{\leq T}|\boldsymbol{x}_{\leq T})} \log p(i_t|\boldsymbol{z}_{\leq t}, \boldsymbol{x}_{<t}) \\ - KL(q(z_t|\boldsymbol{x}_{\leq t}, \boldsymbol{z}_{<t}) \| p(z_t|\boldsymbol{x}_{<t}, \boldsymbol{z}_{<t}))\end{aligned} \tag{4}$$

Image Retrieval. We enable the search process by a timestep-wise hinge loss to model $p(i_t|\boldsymbol{x}, z_t, i_{<t})$. The latent semantic variable z_t is used to predict the image at the given timestep t, with a hinge loss max-margin objective:

$$\mathcal{L}_{HL}(i_t) = \sum_{j} \max(0, \alpha - s(i_t, \hat{i}_t) + s(i_j, \hat{i}_t)) \tag{5}$$

where α is the margin parameter, i_t is the image vector generated by the model, $\hat{i}_t$ is the vector representation of the gold-standard image at time step t, i_j is the negative images, and $s(\cdot)$ denotes the similarity measurement function. In our experiments, we use the cosine distance function.

Overall Objective Function. The overall objective function is the total of the image reconstruction loss and the image retrieval hinge loss summing over all the time steps for the whole image sequence, with β as the weighting factor:

$$\mathcal{L}_{overall} = \sum_{t=1}^{T} \mathcal{L}_{recons.}(i_t) + \beta \mathcal{L}_{HL}(i_t) \tag{6}$$

Parameter Configuration. As the initial parameter setting of the VRSS architecture, we use bi-GRU with the hidden dimension of 500 and set the dimension of latent topics to 500. We also introduce a dropout layer in the RNNs with probability of 0.3. Each word in the text is represented in the 500 dimensional embedding space. The image encoder projects images to a $2,048$ dimensional feature space. For training the objective function, we use AdaDelta optimisation function, with a learning rate of 1.0. The values of hyperparameters α and β were set to be 0.2 and 1.7 respectively.

5 Experimental Setup

We create a train-test split of $60k/6k$ image-text pairs and $9k/1k$ recipes in the Stepwise Recipe dataset. The split is done author-wise to ensure style consistency, but having overlapping authors in train and test splits.

5.1 Models for Comparison

- **LDA.** We re-implement the topic modelling based approach [10] to jointly generate words in text and visual words in image assuming each image-text pair share the same set of topics.
- **Visual Semantic Embeddings (VSE++).** Following [8], we implement a deep neural network approach which maps the text representations and image vectors into the same semantic embedding space.
- **Coherence Neural Story Illustration (CNSI).** We use the encoder-decoder CNSI model proposed in [27], with coherence capturing the co-reference relations among sentences, to retrieve a sequence of images illustrating a passage of text.

- **VRSS-VAE.** This follows the same encoder-decoder architecture of our VRSS model, using two bi-GRU architectures as encoders and decoders with the same learning objective, but without latent variables. Therefore, it is treated as an ablation study of our VRSS model without the VAE module.
- **VRSS-globalCon.** This is a variant of our VRSS model without the incorporation of the global context.

In all the neural models evaluated here, the image representation are extracted using the ResNet50 model [22] pre-trained on food-related images.

5.2 Evaluation Methods

Recall@k indicates that the retrieved image was among the top k best matches out of the set of candidate images. We also define *Story Recall@k*, which considers the retrieved image as correct if it is from the same data sequence. Further, we provide *Visual Saliency Recall@k* values. We implement *Visual Saliency Recall* following [27] and train a VGG-19 network to classify the images of the story test set, with visual features from [22] for initialization. We also report *Visual Feature Similarity* using the average cosine similarity between gold-standard image and retrieved image, considering image features generated by [22].

Previous work [27] highlights that existing quantitative retrieval metrics may be too harsh for a task of this description. Therefore, it is imperative that we use human evaluators to judge how appropriate and coherent the retrieved illustration sequences are. For our human evaluation, we pick a random sample (164 recipes, 1564 image-text pairs) from the test set ($1K$ recipes, $6K$ image-text pairs). We present each evaluator with a sequence of recipe instruction steps that make up one complete recipe. Alongside each text segment, they are given three possible illustrations that depict that step, which are randomly shuffled images of the gold-standard, the non-context model, and the proposed VRSS model. The evaluator is asked to select all image options that may be appropriate illustrations for the corresponding text segment. A total of $5.1K$ ratings are obtained from 12 evaluators, ensuring that every sample receives at least 2 ratings.

6 Results and Discussion

6.1 Automatic Evaluation

Table 1 reports the retrieval performance of different methods using *Recall@k* and *Story Recall@k* metrics. LDA gives the worst results, which shows that using a generative model for capturing the semantic topics from both text and image does not work well in the seq2seq retrieval task. By mapping both text and image into the same embedding space, VSE++ outperforms LDA. Our VRSS model without the VAE component (VRSS-VAE) gives similar performance compared

Table 1. Text illustration performance using *Recall@k (R@k)* and *Story Recall@k (StR@k)* and *Visual Saliency Recall@k (VSR@k)* on the Stepwise Recipe dataset. The best result in each column is highlighted in **bold**.

	Recall@k			Story Recall@k	Visual Saliency Recall@k		
Models	R@1	R@5	R@10	StR@1	VSR@1	VSR@5	VSR@10
Non-context models							
LDA	1.4	3.4	8.9	4.1	3.2	6.7	12.5
VSE++	7.7	18.6	24.6	21.3	8.1	23.1	26.6
Context models							
CNSI	3.6	8.9	13.7	18.4	16.6	31.8	39.8
VRSS-VAE	6.4	19.7	23.1	18.1	11.3	29.2	33.2
VRSS-GlobalCon	5.2	19.9	26.5	21.1	15.1	28.9	32.7
VRSS	**8.2**	**21.3**	**29.8**	**24.4**	**18.4**	**33.4**	**45.1**

to the non-context model VSE++ despite considering the contextual information. VRSS without the incorporation of global context (VRSS-GlobalCon.) performs similarly as VRSS-VAE. CNSI gives worse results compared to both VRSS variants in *Recall@k* and *Story Recall@1*. Our new VRSS model, which maps each hidden state of the RNN into a latent topic and also further incorporates global context information, gives the best results across all metrics. This indicates the importance of representing semantics encoded in both text and images in a more abstract manner and the benefit of incorporating global context.

Recall@k and *Story Recall@k* metrics only measure the degree of exact matches of the retrieved images with regards to the gold-standard images. This might not be appropriate for our text illustration task since a given text segment could be illustrated by multiple images expressing similar semantics. Example image retrieval results are shown in Fig. 3 where both the gold-standard and the VRSS retrieved images are displayed for some recipe instructions. It can be observed that although VRSS failed to retrieve the gold-standard images in these examples, its output images are still appropriate illustrations of the corresponding texts. For this reason, we also report the evaluation results using more semantics-based and feature-based metric, *Visual Saliency Recall@k*.

It can be observed from Table 1 that VRSS performs significantly better than baselines on *Visual Saliency Recall@k*. These recall scores indicate that VRSS is able to retrieve images that are described by text segments that are semantically related to the query text, even if the images themselves do not match the gold-standard image. We also calculate the *Visual Feature Similarity* which measures the average cosine similarity between the gold-standard image and the retrieved image in the feature space. For VRSS, this is 0.51 and for VSE++ it is 0.37, and for CNSI it is 0.45 This confirms that VRSS retrieves illustrations that are visually similar to the gold-standard image.

6.2 Human Evaluation

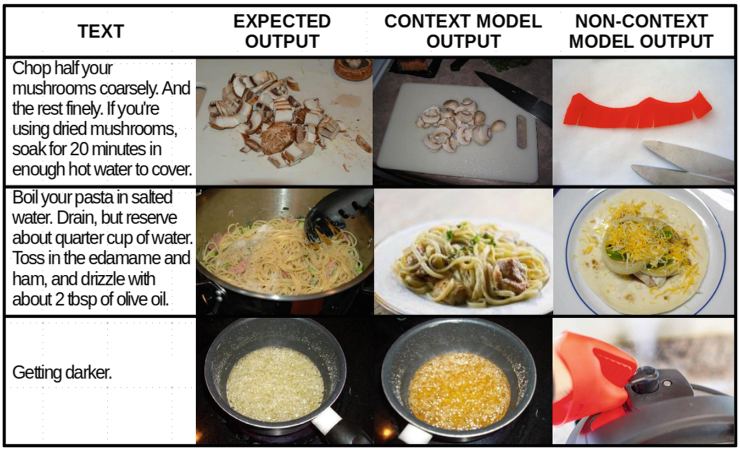
For the human evaluation, we count the number of votes received for the gold-standard images, the VRSS model output images, and the VSE++ (non-context based) model output images. We only count a vote if there is majority consensus among the evaluators. Hence, in Table 2, the '# Votes' column indicates the number that constitutes a majority among voters.

Table 2. Human Evaluation results. The cell values indicate the number of images output by the corresponding model(s) that receive x number of votes ($x \in \{2, 3, 4, 5\}$) as majority.

# Votes	2	3	4	5
Gold-standard only	0	442	171	47
Gold-standard and VRSS	255	41	0	0
Gold-standard and VSE++	88	9	0	0
Gold-standard, VRSS and VSE++	75	0	0	0

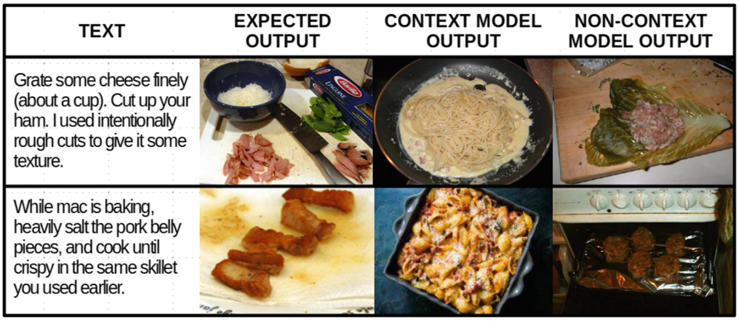
In Table 2, we see the preference results obtained from human evaluation of the retrieved recipe illustrations. Considering majority agreement as 2 votes, gold-standard was never preferred in isolation. Rather, in 61% of the cases, both the gold-standard image and the image retrieved using VRSS were deemed to be appropriate illustrations for the given text query. In 18% of the cases, gold-standard as well as the retrieved images from both models were considered appropriate. In the remaining 21% of the cases, the VRSS output was not judged as being appropriate. Taking 3 votes as the majority, gold-standard alone was picked in 88% of the cases, and picked in combination with the VRSS output in 8% of the remaining cases, with a negligible number of cases for the other combinations. Where the majority consensus is above 4 votes, evaluators chose gold-standard alone in every case. Therefore, VRSS outperforms other models particularly in ambiguous cases where the text is likely to contain an indirect description of the image. The VRSS output is about 3 times more likely to be selected compared to the VSE++ output. Over 60% of the time, at least 2 human evaluators believe that the VRSS output is as appropriate as the gold-standard image. These results indicate that the context based VRSS model significantly outperforms the non-context based model.

Figure 3 shows examples where the VRSS output was preferred by human evaluators. It also highlights cases where metrics other than recall are beneficial such as semantically related entities and paired images having *Visual Feature Similarity*. The last text segment implicitly refers to the previous, with this retrieved image counted favourably when using the context-aware *Story Recall* metric. We also perform a qualitative error analysis, and find that the attention mechanism sometimes misdirects the image retrieval (Figs. 4 and 5).



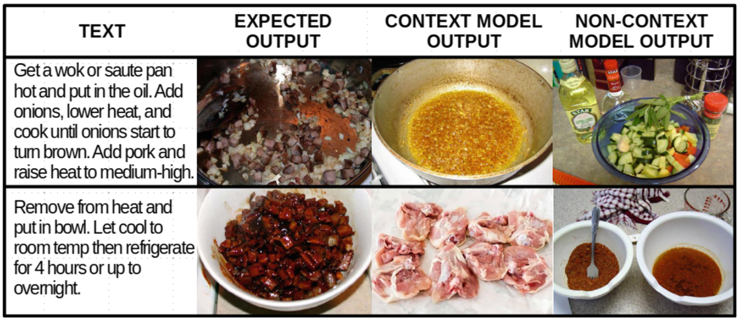
TEXT	EXPECTED OUTPUT	CONTEXT MODEL OUTPUT	NON-CONTEXT MODEL OUTPUT
Chop half your mushrooms coarsely. And the rest finely. If you're using dried mushrooms, soak for 20 minutes in enough hot water to cover.			
Boil your pasta in salted water. Drain, but reserve about quarter cup of water. Toss in the edamame and ham, and drizzle with about 2 tbsp of olive oil.			
Getting darker.			

Fig. 3. Illustrative comparison of non-context (VSE++) and context models (VRSS) - VRSS result preferred by human evaluators.



TEXT	EXPECTED OUTPUT	CONTEXT MODEL OUTPUT	NON-CONTEXT MODEL OUTPUT
Grate some cheese finely (about a cup). Cut up your ham. I used intentionally rough cuts to give it some texture.			
While mac is baking, heavily salt the pork belly pieces, and cook until crispy in the same skillet you used earlier.			

Fig. 4. Illustrative comparison of non-context (VSE++) and context models (VRSS) - VSE++(R) result preferred by human evaluators.



TEXT	EXPECTED OUTPUT	CONTEXT MODEL OUTPUT	NON-CONTEXT MODEL OUTPUT
Get a wok or saute pan hot and put in the oil. Add onions, lower heat, and cook until onions start to turn brown. Add pork and raise heat to medium-high.			
Remove from heat and put in bowl. Let cool to room temp then refrigerate for 4 hours or up to overnight.			

Fig. 5. Illustrative comparison of non-context (VSE++) and context models (VRSS) - Neither VRSS nor VSE++(R) result preferred by human evaluators.

7 Conclusion

We presented VRSS model that given a sequence of text passages, identifies a sequence of images best describing the semantic content of text. We introduced the Stepwise Recipe dataset to facilitate further research on this problem. Our results on the Stepwise Recipe dataset show that VRSS significantly outperforms competitive baselines in terms of both automatic and human evaluations.

References

1. Alikhani, M., Chowdhury, S.N., de Melo, G., Stone, M.: CITE: a corpus of image-text discourse relations. arXiv preprint arXiv:1904.06286 (2019)
2. Balaneshin-kordan, S., Kotov, A.: Deep neural architecture for multi-modal retrieval based on joint embedding space for text and images. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, pp. 28–36. ACM (2018)
3. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101 – mining discriminative components with random forests. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 446–461. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10599-4_29
4. Carvalho, M., Cadène, R., Picard, D., Soulier, L., Thome, N., Cord, M.: Cross-modal retrieval in the cooking context: learning semantic text-image embeddings. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 35–44. ACM (2018)
5. Chandu, K., Nyberg, E., Black, A.W.: Storyboarding of recipes: grounded contextual generation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 6040–6046. Association for Computational Linguistics, Florence, Italy, July 2019. https://doi.org/10.18653/v1/P19-1606. https://www.aclweb.org/anthology/P19-1606
6. Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A.C., Bengio, Y.: A recurrent latent variable model for sequential data. In: Advances in Neural Information Processing Systems, pp. 2980–2988 (2015)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
8. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: VSE++: improving visual-semantic embeddings with hard negatives. arXiv preprint arXiv:1707.05612 (2017)
9. Feng, F., Wang, X., Li, R.: Cross-modal retrieval with correspondence autoencoder. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 7–16. ACM (2014)
10. Feng, Y., Lapata, M.: Topic models for image annotation and text illustration. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT 2010), pp. 831–839. Association for Computational Linguistics, Stroudsburg, PA, USA (2010). http://dl.acm.org/citation.cfm?id=1857999.1858124
11. Goldberg, A.B., Zhu, X., Dyer, C.R., Eldawy, M., Heng, L.: Easy as ABC?: facilitating pictorial communication via semantically enhanced layout. In: Proceedings of the Twelfth Conference on Computational Natural Language Learning (CoNLL 2008), pp. 119–126. Association for Computational Linguistics, Stroudsburg, PA, USA (2008). http://dl.acm.org/citation.cfm?id=1596324.1596345

12. Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: an overview with application to learning methods. Neural Comput. **16**(12), 2639–2664 (2004)
13. He, Y., Xiang, S., Kang, C., Wang, J., Pan, C.: Cross-modal retrieval via deep and bidirectional representation learning. IEEE Trans. Multimed. **18**(7), 1363–1377 (2016)
14. Huang, T.H.K., et al.: Visual storytelling. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1233–1239 (2016)
15. Joshi, D., Wang, J.Z., Li, J.: The story picturing engine–a system for automatic text illustration. ACM Trans. Multimed. Comput. Commun. Appl. **2**(1), 68–89 (2006). https://doi.org/10.1145/1126004.1126008
16. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 664–676 (2017). https://doi.org/10.1109/TPAMI.2016.2598339
17. Kim, G., Moon, S., Sigal, L.: Ranking and retrieval of image sequences from multiple paragraph queries. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1993–2001 (2015)
18. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. CoRR abs/1411.2539 (2014). http://arxiv.org/abs/1411.2539
19. Lin, T., et al.: Microsoft COCO: common objects in context. CoRR abs/1405.0312 (2014). http://arxiv.org/abs/1405.0312
20. Liu, Y., Fu, J., Mei, T., Chen, C.W.: Let your photos talk: generating narrative paragraph for photo stream via bidirectional attention recurrent neural networks. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
21. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(Nov), 2579–2605 (2008)
22. Marin, J., et al.: Recipe1M+: a dataset for learning cross-modal embeddings for cooking recipes and food images. arXiv preprint arXiv:1810.06553 (2018)
23. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of the 28th International Conference on Machine Learning (ICML 2011), pp. 689–696 (2011)
24. Park, C.C., Kim, G.: Expressing an image stream with a sequence of natural sentences. In: Advances in Neural Information Processing Systems, pp. 73–81 (2015)
25. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 2641–2649 (2015)
26. Quadrianto, N., Lampert, C.: Learning multi-view neighborhood preserving projections. In: Proceedings of the 28th International Conference on Machine Learning, Washington, USA, 28 June–2 July 2011, pp. 425–432. Association for Computing Machinery (2011)
27. Ravi, H., Wang, L., Muniz, C., Sigal, L., Metaxas, D., Kapadia, M.: Show me a story: towards coherent neural story illustration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7613–7621 (2018)
28. Rosipal, R., Krämer, N.: Overview and recent advances in partial least squares. In: Saunders, C., Grobelnik, M., Gunn, S., Shawe-Taylor, J. (eds.) SLSFS 2005. LNCS, vol. 3940, pp. 34–51. Springer, Heidelberg (2006). https://doi.org/10.1007/11752790_2

29. Salvador, A., et al.: Learning cross-modal embeddings for cooking recipes and food images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3020–3028 (2017)
30. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: a cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2556–2565. Association for Computational Linguistics (2018). http://aclweb.org/anthology/P18-1238
31. Su, J., Wu, S., Xiong, D., Lu, Y., Han, X., Zhang, B.: Variational recurrent neural machine translation. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
32. Wang, J., He, Y., Kang, C., Xiang, S., Pan, C.: Image-text cross-modal retrieval via modality-specific feature learning. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, pp. 347–354. ACM (2015)
33. Wang, W., Yang, X., Ooi, B.C., Zhang, D., Zhuang, Y.: Effective deep learning-based multi-modal retrieval. VLDB J. **25**(1), 79–101 (2015). https://doi.org/10.1007/s00778-015-0391-4
34. Yagcioglu, S., Erdem, A., Erdem, E., Ikizler-Cinbis, N.: RecipeQA: a challenge dataset for multimodal comprehension of cooking recipes. arXiv preprint arXiv:1809.00812 (2018)