CS 6220 HW 5

Full Name: Wenhao Wang

Student ID: 903665693

Paper #: 1

Section 1: Title of the paper you choose to critique: Multimodal Neural Language Models

Problems: What are the problems addressed in each of the two papers

The paper is trying to bring a new proposed model consisted of multiple neural language model. The authors introduced the importance of the multimodal by explaining the role of a descriptive in a product advertisement. They depicted a scenario that when the customers are shopping online. A model can provide both text and pictures to the customer can greatly make the shopping experience more convenient and efficient by providing the double-sided reference, from their text to the image and back forth. In this paper, the authors brought a multimodal neural language model which is tailored for the modalities.

Speaking of the methodologies, the authors firstly introduced a simple model, “the Log-Bilinear Model (LBL)”, which can predict the next word representation given a real word vector ‘w’, a word representation vectors R, and a context parameter C. The estimation equation is given by

$$\hat{\mathbf{r}} = \sum_{i=1}^{n-1} \mathbf{C}^{(i)} \mathbf{r}_{w_i},$$

And the condition probability is given by

$$P(w_n = i | w_{1:n-1}) = \frac{\exp(\hat{\mathbf{r}}^T \mathbf{r}_i + b_i)}{\sum_{j=1}^{K} \exp(\hat{\mathbf{r}}^T \mathbf{r}_j + b_j)},$$

They also covered that the output layer is used a softmax activation function. Thus, the feature learning can be done through the standard backpropagation. Based on this simple model, the authors introduced their revised model to strengthen the ability of jointly learning the text and image features. The Modality-Biased Log-Bilinear Model (MLBL-B) is designed to add more bias to the word representation that is being predicted. Next, their more powerful model Factored 3-way Log-Bilinear Model (MLBL-F) is to incorporate modality conditionings. Given the learning both image and text features are computationally demanding, the authors innovatively followed other researchers’ methods of using r x r randomly selected patches from images and k-means to speed up the computations.

Section 2: New Idea and Strengths of the paper

Speaking on a high level of the models' application and potential usage, one of the main innovative ideas that the authors have proposed is they incorporate the feature learning involving both the text and the image features. Their newly proposed multimodal language model can generate the modalities' descriptions without using templates, structured models and synthetic trees (Kiros et al., 2014). Instead of searching key words on the internet, the model can link the similar modalities. Their newly proposed model can also automatically generate descriptive text for the newly listed images, which can greatly save the labor efforts to describe the online modalities.

When it comes to the innovations in the methodologies, the authors also conducted various and comprehensive experiments, including image captioning and retrieval tasks, showing the accuracy and capabilities of their multimodal language model.

One of the biggest innovations with respect to the methods is that they creatively combine the convolutional neural network with recurrent neural network making their model can not only be capable of dealing with words sequence but also are able to deal with image data, which is usually in 3 dimensions (W x D x color). Another innovation in the aspects of methods is their loss functions. They innovatively utilized the k-means methods on learning on a small group of features which can significantly reduce the need of computational power and thus speed up the computational speed.

Further, the scalability is one of the strengths of the paper. Usually, the scalability can be an issue for some model given their model's ability can only work on a small sample of dataset. However, the authors in this paper does not rely on the pre-conditions that the model's capability of working on a small sample, but they showcased the model's potential of applying on a real work large dataset.

Section 3 Weaknesses of the proposed solutions and your suggestions if any.

One of the weaknesses of this paper is that it lacks real human reviewing on the results that the model has generated. Given the product will be used for online shopping costumers, the real human review on its capability is crucial for its success. The evaluation metrics primarily rely on the fixed metrics like Bleu which will result the model bias towards the better scores in such metrics and may cause the discrepancy with human feelings and the metric scores.

Another weakness is the paper navigational subtitles. The readers of the paper might need to have quick grasp of the main idea of each section. Given the titles of the model's name, reader might not be able to directly identify their proposed methods and the methods they were comparing with. Adding more navigational function subtitles can greatly improve the reading experience of this paper.

Also, the baseline model is limited. If there are more baseline model to compare with to showcase the performance of their models, it might be more convincing.

Citation:

Kiros, R., Salakhutdinov, R., & Zemel, R. (2014). Multimodal neural language models. *International Conference on Machine Learning*, 595–603. http://ece.duke.edu/~lcarin/dec4.24.2015.pdf

Full Name: Wenhao Wang

Student ID: 903665693

Paper #: 3

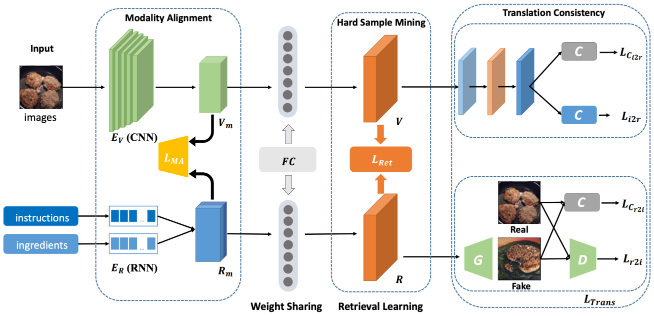
Section 1: Learning Cross-Modal Embeddings with Adversarial Networks for Cooking Recipes and Food Images

Section 1: Problems: What are the problems addressed in each of the two papers

The authors explain the main topics in the paper by first introducing the social media platform called "all-recipes", which can allow the chefs share their created unique cooking recipes and the relevant food images. On the other side of the internet, their followers can share their reproducing the recipes and their feeling about the shared dishes. This is an example of IoT can facilitate people's life. The authors argued that given the images of the food, their ideas are to create such model that can "analyze" the gradient of the food, its nutrition content and calories information. And they introduced their newly proposed cross-model retrieval in the food domain, which is an end-to-end framework named Adversarial Cross Modal Embedding (ACME) (Wang et al., 2019).

In the section of methodologies, the authors created a new embedding, which contains two sets. One set is a group of recipes, and the other group is food images. They created the link between these sets so that the embedding can work well with retrieval tasks.

Speaking of the architecture of the model, the authors established a new model that can take in the image and a words sequence representing the food recipes, like the following.



Section 2: New Idea and Strengths of the paper

In the high level of what the authors achieved in this paper is that they created a brand-new embedding to relate the food image and their recipes. Compare to the previous paper we discussed in this Homework, this paper proposed this new embedding to represent the relation between the food overall and its gradient, while the previous paper put more focus on generating descriptive text that can showcase the product. This paper also introduces the application of adversarial networks to learn the joint embedding, which bridge the gap between the 2 or more different modalities.

Speaking the domain of the methodologies, the authors argue that their model is end-to-end meaning the model can be trained all at once, combining both the embedding learning and adversarial parts into a single system. This makes it easier to optimize and helps the different types of data (like text and images) work together more effectively. This paper also introduced the application of LSTM and ResNet-50 model into the image-text domain which proved its wider application scenarios. In the dataset aspect, the authors leveraged a large dataset containing 238,999 image-recipes pairs for model training and testing, which increases its credibility of accuracy. Also, besides the large dataset selection, they incorporated several food classes in the training which help decreased the food bias and ensure the fairness of the resulting accuracy.

Further, the application of this model is closer to real life, which provide the model with enough training samples to grow and play a more important role in people's real life. Also, in the future-facing scenario, the application could be used for a wider range, such as automatic cooking assistant or dietary suggestion applications that keep people in a healthier lifestyle.

Speaking the result of the testing, the authors used a wide range of existing models as the baseline model to compare with their newly created model. This greatly enhance the effectiveness of their models.

Section 3 Weaknesses of the proposed solutions and your suggestions if any.

Again, given the application is facing customers. The result also needs to incorporate real human labor testing. Without the human testing and purely using the automated metrics to evaluate the result make lead to lacking persuasiveness when the application is provided to the market.

Speaking of the writing styles, the paper lacks the detail of such as training process, hyper-parameter choices. Without these details, it might be difficult for other researchers to reproduce the result and cause discrepancy in knowledge sharing. Also, training the adversarial network can be hard due to the nature of such models. The paper does not include the difficulties that the researchers have encountered, which might make people reproduce their work harder.

Speaking of the potential improvements of the model, which might be too harsh for the product. However, as we know that there might be different kinds of tomatoes or different product of tomatoes. For example, a dish can use the tomato paste or fresh tomatoes. This can lead to different dishes sometimes.

Be more accurate describing the kind and status of the gradient might be more helpful for sharing the chef’s brilliant ideas.

Citation:

Wang, H., Sahoo, D., Liu, C., Lim, E., & Hoi, S. C. (2019b). Learning cross-modal embeddings with adversarial networks for cooking recipes and food images. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11564–11573. https://doi.org/10.1109/cvpr.2019.01184