Learning Capsule Networks with Images and Text

Yufei Feng
ECE, Queen's University
feng.yufei@queensu.ca

Xiaodan Zhu
ECE, Queen's University
xiaodan.zhu@queensu.ca

Yifeng Li
National Research Council Canada
yifeng.li@nrc-cnrc.gc.ca

Yuping Ruan
University of Science and Technology of China
ypruan@mail.ustc.edu.cn

Michael Greenspan
ECE, Queen's University
michael.greenspan@queensu.ca

Abstract

Among its potential but fundamental advantages, a capsule network attempts to better model part-whole relations in images by aggregating agreed votes from lower-level sub-entities. It is still, however, challenging to enforce capsules to represent meaningful parts and therefore the network to capture the intended part-whole composition. In this work we investigate enriched capsule nets on image classification, guided by external textual knowledge that specifies and constrains the desired parts and whole. Our preliminary results show a better performance of the proposed model over the original capsule networks on the ShapeNet screenshot dataset, which renders 3D objects with salient part-whole hierarchies.

1 Introduction

Capsule networks [6, 7, 19] have recently been proposed to address limitations of convolutional neural networks. The networks explicitly group neurons into separate capsules where higher layer capsules are instantiated from votes of lower-level capsules. Research [7, 19] shows that capsule nets can extract robust and salient features to overcome some shortcomings of convolutional neural networks. For example, while convolutional neural networks [5, 10, 20] have achieved great successes on a wide range of tasks, they are still challenged by some basic limitations such as effectiveness of modelling spatial hierarchies and part-whole relations and composition.

Among its potential but fundamental advantages, a capsule network attempts to model part-whole relations in images by aggregating agreed votes from lower-level sub-entities. While capsule nets are developed upon such an intention, it is however still challenging to enforce to learn intended parts and therefore so-formed part-whole composition.

In this work we explore enriched capsule nets on image classification, guided by external textual knowledge that provides information (e.g., common sense) on parts and whole. We aim to enrich capsule nets so that capsules are encouraged to be relevant to entities and sub-components, e.g., those specified in meronyms, and hence part-whole composition can be better captured. Even though neural networks may automatically extract hierarchical features from training data, guiding the network with knowledge embedded in text can take advantage of complementary visual and text information. Our preliminary results show that the proposed model achieves better performances over the original capsule networks on the ShapeNet screenshot dataset, which renders 3D objects with salient part-whole hierarchies.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

2 Related Work

Convolutional neural networks (CNN) and variants [5, 10, 11, 20] have achieved considerable success in a wide range of problems and tasks. Capsule networks [19] have been proposed recently to address some basic limitations of CNN, e.g. those in modelling spatial hierarchies and part-whole composition. More recently the work of [7] further introduces dynamic routing with the EM algorithm and shows that networks with matrix capsules can achieve a significant improvement on the NORB dataset [12]. It is shown in [1, 22] that capsules can achieve good performance on more complex dataset.

Capsule networks have also been adapted to solve natural language processing problems and achieved promising results on tasks such as text classification [24] and natural language inference [23]. Independent of capsule nets, there have also been extensive research efforts studying visual and language representation learning, including but not limited to, visual question answering [2, 14], visual grounding [4, 8], visual relation extraction [13], and image classification [18].

In this work, we focus on the fundamental potentials of capsule nets in modelling part-whole relations. Specifically we explore and leverage external composition knowledge to enrich capsule nets and encourage different capsules to be relevant to part-whole composition specified in textual sources and show that the model achieves better performance on the ShapeNet screenshot dataset.

3 Our Approach

The high-level view of our approach is depicted in Figure 1. Below we discuss the details of the model from lower-level feature extractor (left) to high-level capsules and routing layers (right).

Incorporating Image and Text in Capsules Our model uses shallowly stacked convolutional layers with rectified linear activation as the lower-level feature extractor. We then group learned features into capsules as in [19].

In order to guide lower-layer capsules to learn features related to parts, we add external textual knowledge. Specifically, We use the part-whole concepts encoded in common sense database, the WordNet [16]. For example, a ‘chair’ consists of ‘leg’, ‘arm’, and ‘back’. When building the network, relations as such are incorporated into the models.

We introduce the aggregation layer to integrate image and textual information. We aim to aggregate all features related to certain parts into a single capsule vector representation in the higher layer. It is not straightforward to impose this type of guided routing into the original dynamic routing process, as dynamic routing treats capsules at the higher layer interchangeably, but in our model we expect the capsules to carry different hidden parts.

Scaled dot-product attention, the building block of recently proposed transformer models [21] and many other models, on the other hand, is not symmetric in this sense. We use scaled dot-product attention layer here, with word embedding $E = \{e_1, e_2, ..., e_n\}$ as the ‘Query’ [21] in the dot product, as indicated in Equation (1):

$$C = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

where matrix Q is a linear transformation of all word embedding e_i, K, V are both linear transformation of the input capsules, d_k is the dimension of K, and C is the output capsules of the attention layer. All the lower-level capsule inputs share the same transformation, thus reducing the number of parameters significantly.

Dynamic Routing with Prior After we aggregate similar features into capsules corresponding to certain symbolic parts, we perform voting and dynamic routing. Unlike in [19], we do not initialize the routing coefficients with constant values. Instead the initial routing coefficients between a lower level capsule and a higher one is multiplied by a factor γ before being normalized with softmax, to enhance the part-whole word symbols introduced with external resources. γ is treated as a hyper-parameter. By setting γ to be infinity we assume strict part-whole information flow. The norm of the final layer capsule represents the classification probability and we use the marginal loss proposed by [19].

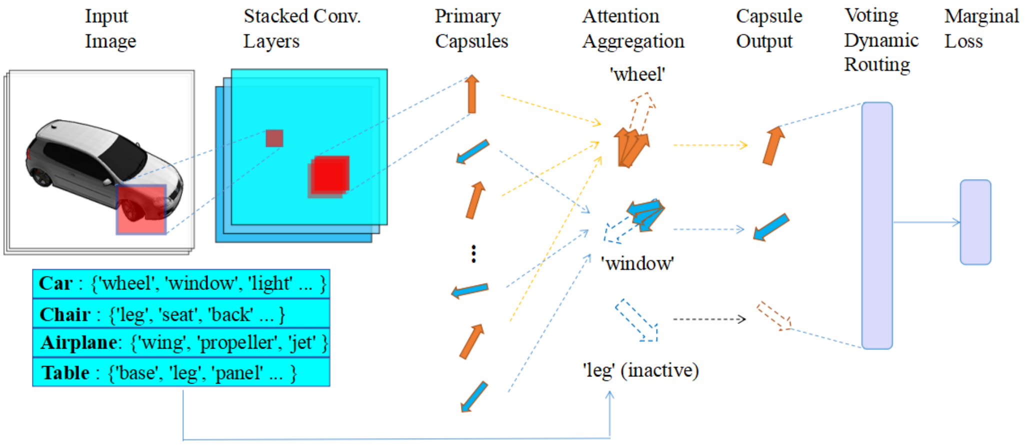
Input Image
Stacked Conv. Layers
Primary Capsules
Attention Aggregation
Capsule Output
Voting Dynamic Routing
Marginal Loss
'wheel'
'window'
'leg' (inactive)
Car : {'wheel', 'window', 'light' ... }
Chair : {'leg', 'seat', 'back' ... }
Airplane: {'wing', 'propeller', 'jet' }
Table : {'base', 'leg', 'panel' ... }

Figure 1: A capsule network with external part-whole knowledge and the dot-product attention.

External Knowledge In order to guide lower-layer capsules to learn features related to specific part concepts as described above, we fetch all related nouns from the gloss of WordNet [16] and meronyms for the categories appear in ShapeNet. We also manually add words expressing parts.

The word embedding learned from text is in a space different from what is potentially learned from images. Pre-trained word embedding [15, 17] places words in a continuous space where salient semantic relatedness information of words is maintained. In images we often do not explicitly get pre-trained features related to certain entities. To address this, the model will use two types of information in image data:

- The co-occurrence of image label and related features and sub-entities. The words corresponding to sub-entities of the main concept in an image will be correlated to the image label automatically throught their part-whole relations.
- The existence of similar image features and sub-entities across different classes. If the same sub-entities and similar features appear in two different classes of images, it is possible to establish a connection between them. This can help computers to give identities to sub-entities and find the correspondence to the words naming them.

We use GloVe [17] embedding vectors to guide the voting and routing process. Note that we build the pool of word vectors into the model rather than feeding them as input during training.

4 Experiments

We performed image classification using the ShapeNet screenshot data [3]. We choose this dataset since it contains clean, artificial images that are rendered from 3D CAD models with the objects and their part-whole information more salient compared with real-life images.

We chose 20 object categories from ShapeNetCore Models [3]. We eliminated categories that have less than 300 examples and did not select categories that contain non part-whole objects (e.g., flowers on flower pots). Most categories have clear part-whole relations and some categories have overlapping parts/subcomponents. We selected the first 300 distinct 3D models for each category and rotate them with 8 evenly spaced angles before taking screen-shot, resulting in a dataset containing 48,000 labeled images. The image size is 128*128 with 3 color channels. One third of the dataset is used as the test set and the rest as our training set. We keep all hyper-parameters same as in [19].

We compared our results to that of the original capsule layers [19]. Below we list several capsule network models and show the classification accuracy in the table. All the networks are built upon one convolutional layer with kernel stride equals 1 and two convolutional layers with kernel stride 2. The

Table 1: Accuracy of models on the test set.

Model	Accuarcy
CapsuleNet	75.52%
CapsuleNet_2layer	76.93%
CapsuleNet_wordconcat	76.22%
CapsuleNet_attention	**78.12%**

kernel size is 5*5. We use Adam stochastic gradient descent [9] as the optimizer. All models use the same hyper parameters and we train 100 epoch before reporting the best test accuracies.

- **CapsuleNet**: the original capsule network with dynamic routing.
- **CapsuleNet_2layer**: CapsuleNet with one more capsule layer with dynamic routing.
- **CapsuleNet_wordconcat**: CapsuleNet with 2 capsule layer, and the transformed word embedding vector is concatenated to middle layers before generate voting.
- **CapsuleNet_attention**: Our CapsuleNet with scaled dot attention and word embedding is used as attention query.

Table 1 shows that the proposed model achieves better accuracy on the test set. Adding more capsule layers with dynamic routing can improve classification accuracy while the naive approach of concatenating transformed word embedding to lower layer capsules does not help. We due this to the fact that the network cannot automatically bridge between concepts learned from text and that potentially learned from the image feature space. Our proposed model uses scaled dot product and achieved better performance without adding too many parameters.

5 Conclusion and Discussion

Our preliminary exploration shows that incorporating into capsule networks the external part-whole knowledge is promising, improving the performance on the ShapeNet screenshot image classification. The attention mechanism is capable of absorbing the external knowledge into the network. This is a preliminary but the first attempt, according to our knowledge, to study visual and language information in capsule nets with the aim to better leverage the networks' potentials in modelling part-whole relationship.

References

[1] Parnian Afshar, Arash Mohammadi, and Konstantinos N Plataniotis. Brain tumor type classification via capsule networks. *arXiv preprint arXiv:1802.10200*, 2018.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.

[3] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[4] Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. Visual grounding via accumulated attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7746–7755, 2018.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[6] Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. Transforming auto-encoders. In *International Conference on Artificial Neural Networks*, pages 44–51. Springer, 2011.

[7] Geoffrey E Hinton, Sara Sabour, and Nicholas Frosst. Matrix capsules with em routing. *The International Conference on Learning Representations (ICLR).*

[8] Syed Ashar Javed, Shreyas Saxena, and Vineet Gandhi. Learning unsupervised visual grounding through semantic self-supervision. *arXiv preprint arXiv:1803.06506*, 2018.

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations*, 2014.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[11] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[12] Yann LeCun, Fu Jie Huang, and Leon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–104. IEEE, 2004.

[13] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, pages 852–869. Springer, 2016.

[14] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–9, 2015.

[15] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[16] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[17] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[18] Zhou Ren, Hailin Jin, Zhe Lin, Chen Fang, and Alan Yuille. Joint image-text representation by gaussian visual-semantic embedding. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 207–211. ACM, 2016.

[19] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, pages 3856–3866, 2017.

[20] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[22] Edgar Xi, Selina Bing, and Yang Jin. Capsule network performance on complex data. *arXiv preprint arXiv:1712.03480*, 2017.

[23] Deunsol Yoon, Dongbok Lee, and SangKeun Lee. Dynamic self-attention: Computing attention over words dynamically for sentence embedding. *arXiv preprint arXiv:1808.07383*, 2018.

[24] Wei Zhao, Jianbo Ye, Min Yang, Zeyang Lei, Suofei Zhang, and Zhou Zhao. Investigating capsule networks with dynamic routing for text classification. *arXiv preprint arXiv:1804.00538*, 2018.