Cross Modal Content Based Objective for Learning Adequate Multimodal Representations

Adhiguna Kuncoro* AKUNCORO@CS.CMU.EDU
Akash Bharadwaj* AKASHB@CS.CMU.EDU
Seungwhan Moon* SEUNGWHM@CS.CMU.EDU
Volkan Cirik* VCIRIK@CS.CMU.EDU
Louis-Philippe Morency MORENCY@CS.CMU.EDU
Chris Dyer CDYER@CS.CMU.EDU
Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA *: equal contributions

Abstract

We explore the complementary nature of Machine Translation and Image Captioning through a novel task called Multimodal Machine Translation, introduced in the Workshop on Machine Translation 2016 (WMT, 2016). This paper demonstrates a modular approach to elicit a representation from the image that provides grounding for adequacy in a neural machine translation or image captioning system through the use of a content based objective while training. We demonstrate that our proposed image representation is complementary with the image representations obtained from large convolutional neural nets and yields improvements for various multimodal machine translation architectures.

1. Introduction

Multimodal machine translation is a novel task that, given an image and its caption in a source language, generates a caption in a target language. This task seeks to reconcile two fundamentally different problems: machine translation and image captioning. Machine translation deals with two views (languages) of the same modality (text) where one view is a very explicit depiction of the other. Translations to the other view should not just be fluent, but also adequate and faithful to the source view. Image captioning, on the other hand, deals with heterogeneous modalities (image and text) where the image can be thought of as a noisy or abstract depiction that can be adequately described by multiple possible captions that need not be consistent with one another.

Each of these tasks has its unique challenges. In machine translation, polysemous words may not be correctly translated if a sentence is ambiguous in its construction. Moreover, in neural machine translation systems the decoder often behaves like an aggressive language model that generates fluent but inadequate translations by over-fitting to the training set. In image captioning, extracting a representation from the image that is adequate for a generative task like captioning still remains as a significant challenge. For multimodal machine translation, the challenge is thus to extract the appropriate image representation that can provide grounding to a machine translation system so it can produce fluent and adequate translations.

2. Prior Work

For image captioning, a primary concern is encoding the image into a multimodal representation from which the textual caption can be decoded. Some of the proposed approaches include Deep Boltzmann Machines (DBM) (Srivastava & Salakhutdinov, 2012b), auto-encoders (Jiquan Ngiam & Ng, 2011), DCCAE (Weiran Wang & Bilmes, 2013) and encoder-decoder models (Ryan Kiros & Zemel, 2014; Oriol Vinyals, 2014). (Andrej Karpathy & Li, 2014) and (Fang, 2015) look at sub-regions of images to elicit appropriate representations from them. (Kelvin Xu & Bengio, 2015) introduces visual attention in which the model learns to identify sub-regions of an image relevant to a decoding step during caption generation.

Statistical machine translation tools such as (Philipp Koehn, 2007) have been the workhorses of machine translation for years. More recently, encoder-decoder neural models such as (Ilya Sutskever & Le, 2014) have been proposed. (Dzmitry Bahdanau & Bengio, 2015), (Minh-Thang Luong & Manning, 2015) introduce the notion of attention and obtain strong gains in performance. We implement a model similar to (Dzmitry Bahdanau & Bengio, 2015) to demonstrate our results.

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48. Copyright 2016 by the author(s).

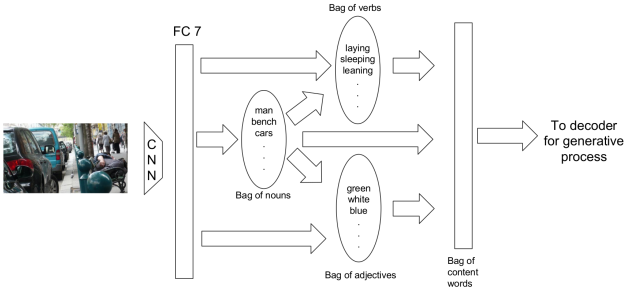

Figure 1. An illustration of the CNN filter approach.

3. Proposed Method

All of the cited prior work except (Fang, 2015) use representations of images and text that abstract away from the surface form of content using vector embeddings. Input image representations are typically obtained from large convolutional neural nets (CNNs) and tuned by backpropagating errors from the end task at hand or to maximize correlation or reconstruction of modalities in the abstract vector space. Similarly to (Fang, 2015), our approach elicit an explicit bag of content words from the image. We differ in that we do not make use of multiple instance learning, and that we use FC7 CNN representation of the entire image instead of sub-regions. In addition, we do not assume content word detectors are independent.

We propose a new cross modal content based objective approach called the **CNN filter**, where the bag of words from the image is supplied along with an encoded representation of the source language sentence (using an LSTM) to the textual generative process of decoding in the target language (illustrated in Figure 1). The objective of the CNN filter is formulated as follows:

$$\begin{aligned}
N =& \sigma\left(fc \cdot W_N + b_N\right)\\
V =& \sigma\left(N \cdot W_{NV} + fc \cdot W_V + b_V\right)\\
A =& \sigma\left(N \cdot W_{NA} + fc \cdot W_A + b_A\right)\\
L =& \sum_{i=1}^{|N|} l(z_i^N, N_i) + \sum_{j=1}^{|V|} l(z_j^V, V_j) + \sum_{k=1}^{|A|} l(z_k^A, A_k)
\end{aligned}$$

where fc is an FC7 image representation, N is a bag of nouns, V is a bag of verbs, A is a bag of adjectives, $l(z, y)$ is a log loss of prediction y given gold label z, L is the CNN filter loss, σ is a sigmoid activation.

The insight here is that CNNs are competent at a discriminative task that predicts nounal class labels, rather than a generative task. A text generation task requires correct identification of content words such as verbs and adjectives as well. Certain types of adjectives and verbs can be abstract concepts that are hard to detect from an image alone. Leveraging the competency of discriminatively trained CNNs to identify nouns and conditioning identification of other content words on the these nouns allows the model to benefit from distributional semantics, unlike independent word detectors. For example, if the nouns identified include a ‘man’ and a ‘ball’, it is likely that verbs like ‘throw’ or ’kick’ are identified while a verb like ‘cook’ is unlikely. The image can be used as disambiguating context to choose the correct verb from these limited possibilities. The image’s bag of words encompasses words that could have been used in any caption pertaining to it. For our multimodal translation task, this could provide complementary information not present in the particular source caption used for translation. This could help avoid problems in machine translation highlighted in Section 1.

Table 1. METEOR scores for varying architectures with and without CNN Filter. **Source**: English, **Target**: German.

Model	Visual Feature	METEOR
NMMT + Attention	**fc7 + CNN Filter**	**30.28**
NMMT + Attention	fc7	29.54
NMMT	**fc7 + CNN Filter**	**19.32**
NMMT	fc7	18.72
NMT	N\A	18.8

4. Experimental Results

In this section, we report results for the constrained multimodal translation task as described in (WMT, 2016) (English+Image→German). The organizers of the shared task have only made available a training set and a validation set which account for 29,000 and 1,014 images respectively. In our experiments, we only use the validation set for fine tuning hyper parameters and do not use it during training. Reported results are on validation set.

Baselines: NMT refers to a conventional sequence to sequence neural translation model similar to (Ilya Sutskever & Le, 2014). NMMT is a multimodal translation model that uses both encoded text and image representations as an input at each time step to the LSTM decoder. Attentional models only incorporate attention over the text modality. **Main results**: Using simple FC7 representation of the image (NMMT with fc7) actually hurts performance as compared to the NMT model showing that combining the two modalities is not trivial. Using the content based objective to obtain a bag of content words through the CNN filter provides a gain of a little over 0.5 METEOR over simple FC7. We find that this gain increases in the attentional model. **Model Hyperparameters**: In our experiments we find that single hidden layer (256 hidden units) LSTMs work better than deeper models, which could be due to the low data scenario of the task. We keep the decoder and text encoder architecture the same across all models being compared.

Conclusion: We propose a new technique called the CNN filter to elicit useful image representations for a textual generative task that provides gains over FC7 representations. We demonstrate that image and text modalities can be complementary for multimodal machine translation. We will extend this work with more rigorous analyses.

References

Andrej Karpathy, Armand Joulin and Li, Fei Fei F. Deep fragment embeddings for bidirectional image sentence mapping. *Advances in Neural Information Processing Systems*, pp. 1889–1897, 2014.

Dzmitry Bahdanau, Kyunghyun Cho and Bengio, Yoshua. Neural machine translation by jointly learning to align and translate. *In Proceedings of International Conference on Learning Representations*, 2015.

Fang, Hao, Saurabh Gupta Forrest Iandola Rupesh K. Srivastava Li Deng Piotr DollÃąr Jianfeng Gao Xiaodong He Margaret Mitchell John C. Platt C. Lawrence Zitnick Geoffrey Zweig. From captions to visual concepts and back. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1473–1482, 2015.

Ilya Sutskever, Oriol Vinyals and Le, Quoc V. Sequence to sequence learning with neural networks. *In Advances in neural information processing systems*, pp. 3104–3112, 2014.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim Juhan Nam Honglak Lee and Ng, Andrew Y. Multimodal deep learning. *28th International Conference on Machine Learning*, pp. 689–696, 2011.

Kelvin Xu, Jimmy Ba, Ryan Kiros Aaron Courville Ruslan Salakhutdinov Richard Zemel and Bengio, Yoshua. Show, attend and tell: Neural image caption generation with visual attention. *International Conference on Machine Learning*, 2015.

Minh-Thang Luong, Hieu Pham and Manning, Christopher D. Effective approaches to attention-based neural machine translation. *Empirical Methods in Natural Language Processing*, 2015.

Oriol Vinyals, Alexander Toshev, Samy Bengioa dn Dumitru Erhan. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.2539*, 2014.

Philipp Koehn, Hieu Hoang, Alexandra Birch Chris Callison-Burch Marcello Federico Nicola Bertoldi Brooke Cowan et al. Moses: Open source toolkit for statistical machine translation. *In Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pp. 177–180, 2007.

Ryan Kiros, Ruslan Salakhutdinov and Zemel, Richard S. âĂIJunifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

Srivastava, N. and Salakhutdinov, R. R. Multimodal learning with deep boltzmann machines. In Langley, Pat (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 2222–2230, Stanford, CA, 2012a. Morgan Kaufmann.

Srivastava, Nitish and Salakhutdinov, Ruslan. Multimodal learning with deep boltzmann machines. *Advances in neural information processing systems*, pp. 2222–2230, 2012b.

Weiran Wang, Raman Arora, Karen Livescu and Bilmes, J. On deep multi-view representation learning. *32nd International Conference on Machine Learning*, pp. 1083–1092, 2013.

WMT. Shared task: Multimodal machine translation. *http://www.statmt.org/wmt16/multimodal-task.html*, 2016.