Uncertainty quantification in image captioning models

Shashank Bujimalla*
Intel Corporation
shashankbvs@gmail.com

Mahesh Subedar*
Intel Labs
mahesh.subedar@intel.com

Omesh Tickoo
Intel Labs
omesh.tickoo@intel.com

Abstract

Bayesian deep neural networks (DNN) provide a mathematically grounded framework to quantify uncertainty in their predictions. Image captioning is still an active area of research and DNN models can generate incorrect description of the image. Hence, it is important to study the inherent ambiguity or uncertainty estimates from the generated captions. We provide a detailed study of uncertainty quantification for the predicted captions, and demonstrate that it correlates well with the CIDEr-D scores. To our knowledge, this is the first such analysis, and it can pave way to more practical image captioning solutions with interpretable model outputs.

Introduction:

Deep neural networks (DNNs) provide state-of-the-art results for a multitude of applications, including image captioning approaches [4] that generate natural language (NL) descriptions by transforming the image features into a sequence of output words from a predefined vocabulary. However, DNNs have been shown to fail [2] in the case of noisy or out-of-distribution data leading to overly confident SoftMax probability scores. Probabilistic Bayesian models, on the other hand, provide a principled way to gain insight into the data and capture reliable uncertainty estimates in their predictions, hence providing interpretable models. Bayesian DNNs provide a convenient way to combine the two approaches to develop more robust models that can be scaled to large datasets and real-world applications. Bayesian modeling with Monte Carlo (MC) dropout [2] approximate inference is shown as a practical approach to implement Bayesian DNNs in order to obtain principled confidence and quantify predictive uncertainty.

Our main contribution in this work is to present detailed uncertainty quantification of the captions generated using state of the art models [4], and demonstrate a good correlation between CIDEr-D scores (popular NL metric) and predictive entropy.

Uncertainty quantification: According to Bayes rule, the posterior distribution of model parameters w is given by the equation:

$$p(w|x,y) = \frac{p(y|x,w)p(w)}{p(y|x)} \tag{1}$$

where, (x, y) are input-output pairs, $p(y|x, w)$ is the model likelihood and $p(w)$ is the prior over the model parameters. Computing the posterior distribution $p(w|x, y)$ is often intractable, and hence Bayesian approximate inference techniques have been proposed. Predictive distribution for Bayesian DNNs is obtained using multiple stochastic forward passes through the network during the prediction phase, while sampling from the posterior distribution of network parameters through MC estimators. Equation 2 shows predictive distribution of the output y^* given new input x^*:

$$p(y^*|x^*,x,y) \approx \frac{1}{M}\sum_{i=1}^{M} p(y^*|x^*,w_i) \ , \quad w_i \sim q_\theta(w) \tag{2}$$

where, M is number of Monte Carlo samples.

In this study, we perform MC dropout during inference by enabling dropout in the final fully connected layer, and obtain the MC samples for our Bayesian analysis. We evaluate the model uncertainty using Mutual Information (MI) between parameter posterior distribution and predictive distribution [5].

$$MI := H(y^*|x^*,x,y) - \mathbb{E}_{p(w|x,y)}[H(y^*|x^*,w)] \tag{3}$$

where, the first term $H(y^*|x^*, x, y)$ is the Predictive entropy that captures a combination of both input data and model uncertainty. The second term $\mathbb{E}_{p(w|x,y)}[H(y^*|x^*, w)]$ in Equaton 3 is the mean entropy, where $H(y^*|x^*, w)$ is the entropy of each MC sample. For each MC sample here, we obtain the SoftMax probability of an image caption by concatenating

*Equal Contribution

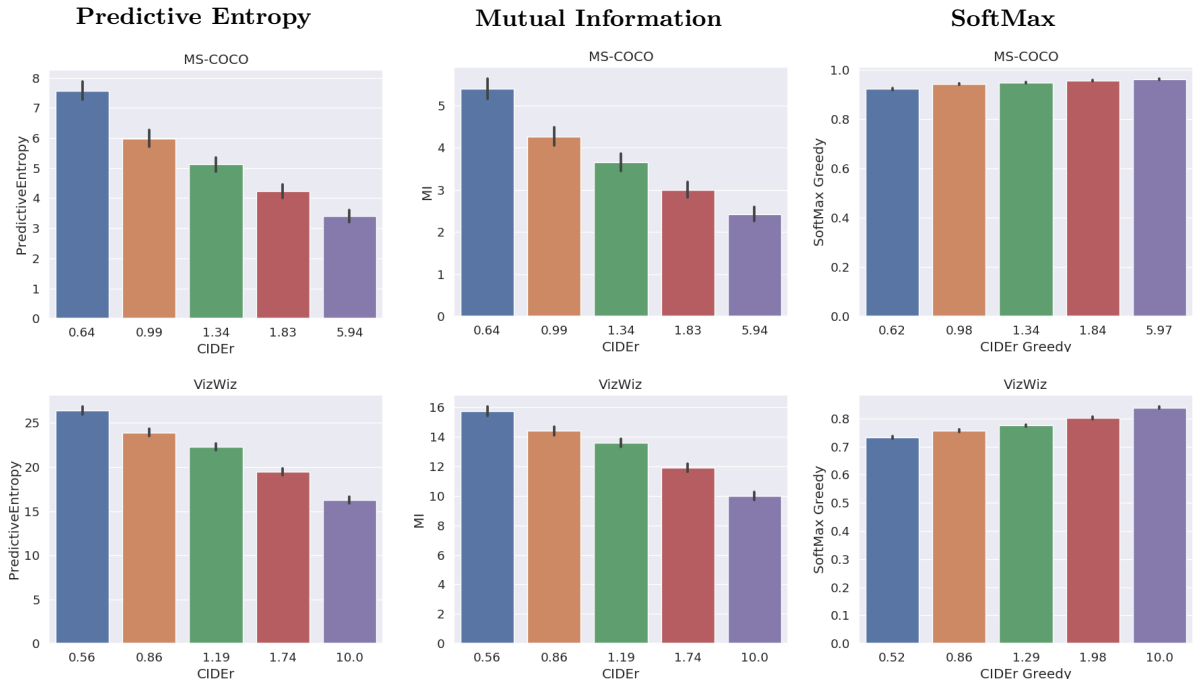


Figure 1: Comparison of Uncertainty vs predictive mean CIDEr-D scores using Bayesian inference (columns 1 & 2) and SoftMax vs CIDEr-D scores using standard DNN inference (column 3). It demonstrates that the uncertainty estimates obtained from Bayesian inference are well correlated with the predictive mean CIDEr-D scores, where lower uncertainty (higher confidence) scores are observed for higher CIDEr-D scores. On the contrary, SoftMax probabilities give high scores for different levels of CIDEr-D scores.

the SoftMax probabilities of all the words in the caption, where each word's Softmax probability vector contains all the output classes (i.e., vocabulary). The predictive distribution of the captions, obtained using 30 MC dropout forward passes, is used to estimate MI (Equation 3) and predictive entropy uncertainty scores [1].

In first two columns of Figure 1, we plot the uncertainty estimates (Predictive Entropy and Mutual information) against predictive mean CIDEr-D scores across MC dropout forward passes for MS COCO and ViwWiz [3] datasets, obtained using AoANet model [4]. We map CIDEr-D scores into five quantiles and plot the average uncertainty score for each quantile. We observe that lower CIDEr-D scores indicate higher uncertainty in the predictions, where as higher CIDEr-D scores indicate lower uncertainty. Both these uncertainty measures show good correlation with the CIDEr-D scores, which is critical for the interpretability of the captions generated by the model. In the last column of Figure 1, we plot the mean SoftMax probability per word in the caption against the caption's CIDEr-D scores using standard DNN inference, i.e. greedily choosing the word with highest SoftMax at each timestep to generate the caption. We observe that SoftMax probabilities are uniformly distributed for different levels of CIDEr-D scores, further validating that the SoftMax probabilities could be overly confident in predicting the CIDEr-D scores, and not a good measure of predictive confidence of the model.

In summary, this uncertainty quantification analysis demonstrates that Bayesian approaches provide robust predictive confidence scores compared to SoftMax probabilities obtained from standard DNN image captioning models.

References

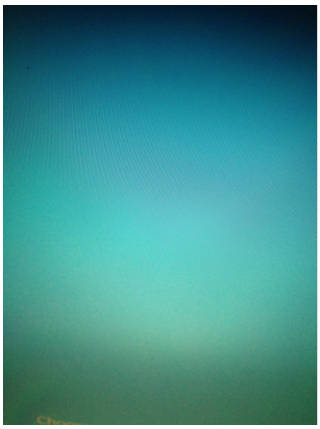
[1] Shashank Bujimalla, Mahesh Subedar, and Omesh Tickoo. B-SCST: Bayesian self-critical sequence training for image captioning. *arXiv preprint arXiv:2004.02435*, 2020.

[2] Yarin Gal. Uncertainty in deep learning. *University of Cambridge*, 2016.

[3] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. *arXiv preprint arXiv:2002.08565*, 2020.

[4] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *IEEE International Conference on Computer Vision*, pages 4634–4643, 2019.

[5] Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. *arXiv preprint arXiv:1803.08533*, 2018.



Ground-truth Caption
a turquoise computer screen contains partial text at the bottom

CIDEr-D: Predicted Caption
0.000: quality issues are too severe to recognize visual content

Predicted Captions using MC Dropout (showing 10/30)
a turquoise desktop screen that is darker at the bottom
quality issues are too severe to recognize visual content
a turquoise computer screen with no on it
teal turquoise ombre screen in the dark room
plain greenish blue ombre image in dark view
a turquoise computer screen with a little bit of shadow
a turquoise computer screen contains some UNK at the bottom
quality blue computer monitor with no images on it
a turquoise grainy screen with no design on it
a greenish blue textured background in the dark

SoftMax Probability (per word mean): 0.9234
Predictive entropy: 20.99
Mutual information (MI): 11.35

Figure 2: The image from VizWiz val split dataset. Softmax score (0.92) for the standard DNN inference is high (indicating higher confidence) even if the CIDEr score is low (0.000). On the other hand, Bayesian uncertainty scores (Predictive Entropy = 20.99, MI = 11.35) are high indicating appropriate lower confidence.