Captions generated from a single image are possibly different from each others as for representations (e.g. attention points or sentence expressions). However, a vast amount of image captioning datasets in the world have few or no annotations of latent variables. Learning latent variables of captions with no supervision is an important from perspectives of scalability and interpretability of conditional image captioning models. In this research, we propose a deep generative model to learn and leverage latent variables of image captions. In experiments, we used the task of image classification with several MNIST images and ground truth labels as down-scaled setting of image captioning, and we show that our proposed model acquired latent variables which represent sub-groups of labels.