Evidential Interactive Learning for Medical Image Captioning

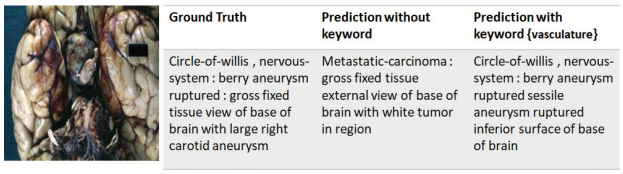
Ervine Zheng [1] **Qi Yu** [1]

Abstract

Medical image captioning alleviates the burden of physicians and possibly reduces medical errors by automatically generating text descriptions to describe image contents and convey findings. It is more challenging than conventional image captioning due to the complexity of medical images and the difficulty of aligning image regions with medical terms. In this paper, we propose an evidential interactive learning framework that leverages evidence-based uncertainty estimation and interactive machine learning to improve image captioning with limited labeled data. The interactive learning process involves three stages: keyword prediction, caption generation, and model updates. First, the model predicts a list of keywords with evidence-based uncertainty estimation and selects the most informative keywords to seek user feedback. Second, user-approved keywords are used as model input to guide the model to generate satisfactory captions. Third, the model is updated based on user-approved keywords and captions, where evidence-based uncertainty is used to allocate different weights to different data instances. Experiments on two medical image datasets illustrate that the proposed framework can effectively learn from human feedback and improve performance in the future.

1. Introduction

Medical image captioning aims to automatically generate text descriptions for medical images to describe image contents and key findings. It typically integrates a computer vision model to extract semantic features from medical images and a language model to generate readable text captions (Pavlopoulos et al., 2022). Compared with conventional captioning tasks on natural images, medical image captioning is usually more challenging because of its highly specialized domain (Li et al., 2019). Meanwhile, the tolerance for error is low because wrong predictions could result in severe consequences.



Ground Truth	Prediction without keyword	Prediction with keyword {vasculature}
Circle-of-willis , nervous-system : berry aneurysm ruptured : gross fixed tissue view of base of brain with large right carotid aneurysm	Metastatic-carcinoma : gross fixed tissue external view of base of brain with white tumor in region	Circle-of-willis , nervous-system : berry aneurysm ruptured sessile aneurysm ruptured inferior surface of base of brain

Figure 1: An illustrative example showing keywords helps improve the quality of medical image caption generation.

Keyword-driven medical image captioning is an approach to address the above challenges (Biswal et al., 2020). Its main idea is to leverage additional keywords provided by the user as side information. According to (Pavlopoulos et al., 2019), keywords commonly exist in the doctors' textual diagnosis records in the early diagnosis process. Keywords are critical for medical image captioning because they are usually highly informative in describing a disease's morphology and potential indications. With correct keywords, the generated captions are more likely to capture the essence of the image and incur fewer mistakes, as shown in the illustrative example in Figure 1. Particularly, if the distribution of the testing image is different from the training data or the quality of the image is low, leveraging only the image data may be insufficient for a model to infer satisfactory captions. Additional keywords would help the model learn the context and retrieve other relevant words for caption generation.

Compared with conventional image captioning with only one input modality (*i.e.,* image), keyword-driven captioning is multi-modal (*i.e.,* image and user-specified keywords), and it can be formulated as an interactive process that involves humans in the loop. While existing keyword-driven models have achieved noticeable success (Huang et al., 2019; Maksoud et al., 2019; Huang et al., 2021), they always require keywords as additional inputs, which may incur an excessive burden on the user. In fact, if the model cannot make a satisfactory prediction and requires additional user inputs as guidance, it indicates that the model is not calibrated well on the current data. In this case, an ideal solution is to make the model learn from users during the interaction process, so that the model can perform better in future cap-

[1]Rochester Institute of Technology. Correspondence to: Qi Yu <qi.yu@rit.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

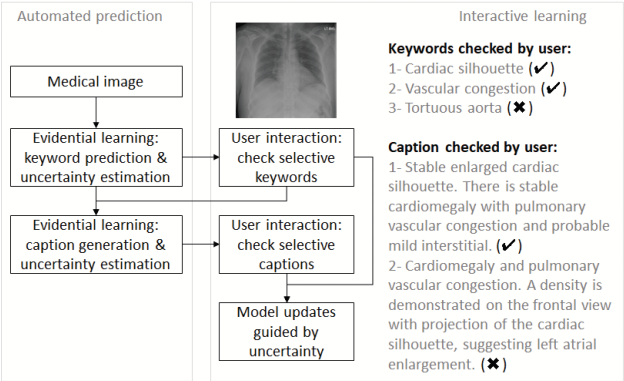


Figure 2: Workflow of the interactive learning framework. During interactive learning, the model predicts a list of keywords with evidence-based uncertainty and selects the most informative keywords to seek user feedback. Based on user-approved keywords, the model generates several candidate captions to seek user feedback. The model is updated using user-approved keywords and captions so that it can perform better in the future.

tioning tasks even without the aid from users. Specifically, during the interaction process, the keywords specified by the users can be considered as ground truth and leveraged as additional supervision to update the model. However, the size of keyword vocabulary could be large. When selecting appropriate keywords, the user needs to evaluate whether each candidate keyword is relevant to the image. This may incur excessive burdens on the user, especially in expert domains such as medicine, where examing an image requires much effort. In addition, the conventional model training process requires a large amount of annotated data, which may also incur annotation burdens.

To address the above challenges, we propose an interactive learning framework that effectively involves user interactions to improve the model's performance. Our framework involves three stages: keyword prediction, caption generation, and model updates. In the first stage, the model evaluates the uncertainty of candidate keywords that are potentially relevant to the image, and queries keywords with the highest uncertainty to the user. The user then determines whether the query keywords are correct, and provide feedback to the model. In the second stage, our model leverages the keywords as the side information to generate captions and uncertainty estimation. The user then selects the captions as feedback to the model. In the last stage, users' feedback is used as weak supervision for model updates, and the updated model is expected to perform better in the future. The workflow is summarized in Figure 2.

Uncertainty plays a critical role in the interactive learning process. First, our model queries keywords from users in a selective way, which provides two major benefits: 1) The user only needs to focus on a few query keywords with the highest uncertainty (rather than all keywords), which greatly reduces the user's burden. 2) User feedback on selected keywords can provide the model with the most helpful information for model updates. Intuitively, an effective model training strategy shall effectively calibrate the model on the data with which the model was initially uncertain. For uncertainty estimation, our framework introduces an evidence-learning paradigm. By leveraging evidence learning under the subjective logic framework (Josang et al., 2018), we focus on two important sources of uncertainty: 1) vacuity, which is caused by lack of evidence; and 2) dissonance, which is caused by the conflict of strong evidence. Evidential learning provides insights into the sources of uncertainty, which is instrumental for model updates.

Second, we provide a theoretical analysis to integrate the two sources of uncertainty systemically. Specifically, in evidential learning, a model typically assigns a low belief mass to the ground-truth class of an unfamiliar data instance. Our analysis unveils important connections between the belief mass and the integrated uncertainty. On top of that, we propose an integrated uncertainty quantification method to effectively select keywords with the highest total uncertainty for user interaction and dynamically balance the two sources of uncertainty during the interactive learning process.

Third, evidence-based uncertainty estimation can also be leveraged for model updates. Once the model generates candidate captions, the user can rank the generated captions to provide feedback to the model. During model updates, the top-ranked caption can be treated as the ground truth. However, using only the top-ranked caption for updates may incur overfitting, because an image can usually be described in multiple ways. To address this issue, we can treat other candidate captions predicted by the model as noisy data, which can be used for model updates with a lower weight. The uncertainty estimation can be used as the weighting mechanism. As a result, this training scheme can effectively mitigate overfitting while being robust to the noise.

Our contributions are summarized as follows:

- An interactive medical image captioning framework that effectively involves human users in the loop to predict accurate keywords and captions;
- A theoretical analysis of evidence-based uncertainty to integrate two sources of uncertainty;
- An evidential uncertainty-guided keyword sampling strategy that queries most informative keywords for user interaction to reduce users' burden;
- An evidential uncertainty-guided update strategy that trains the model in an annotation-efficient way based on sparse user feedback.

2. Related Works

Medical Image Captioning: Medical image captioning is a specialized domain that applies image captioning methods to analyze medical images (Pavlopoulos et al., 2022). It can generally be grouped into automatic generation approaches and template-based retrieval approaches. For example, (Jing et al., 2017) presented a basic co-attention model to implement automatic medical report generation. (Li et al., 2018) integrates a template-based method with the generation framework in a reinforcement learning fashion. (Li et al., 2019) proposes a knowledge-driven encoding with a graph transformer to capture the relationship among abnormality concepts. (Zhang et al., 2020) proposes pre-constructed graph embedding based on multiple disease terms. For keyword-driven medical image captioning, (Huang et al., 2019) encodes keyword information by a multi-layer perception, adding to the topic representation for caption generation. (Biswal et al., 2020) develops a template-based retrieval model that accepts keywords as additional inputs. (Huang et al., 2021) encodes multiple keywords through a contextualized encoder as the loose guidance for sentence generation. (Alfarghaly et al., 2021) leverages a conditioned transformer-based captioning model to integrate image features and keyword embeddings. (You et al., 2021) proposes to match visual regions from medical images and candidate keywords to enhance caption generation. (Wu et al., 2023) proposes an attention-based strategy to match expert-defined keywords with local image patches. Our work advances the approaches mentioned above by learning from user feedback, and using uncertainty estimation for effective user interaction.

Uncertainty Estimation: Uncertainty quantifies the degree to which a machine learning model is uncertain about its predictions and implies whether users can trust the results. For deep learning, Bayesian neural network with Monte Carlo dropout (Gal & Ghahramani, 2016), Bayes-by Backprop (Blundell et al., 2015), and deep ensembles (Lakshminarayanan et al., 2017) are representative approaches to evaluate uncertainty. In natural language processing domains, (Xiao & Wang, 2019) explores uncertainty estimation via Bayesian neural network on sentiment analysis tasks, named entity recognition and language modeling, and (Wang et al., 2019) explores uncertainty estimation via Bayesian neural network on machine translation with back-translation technique. (Siddhant & Lipton, 2018) leverages uncertainty estimates provided by dropout and Bayes-by Backprop for active learning. (Ott et al., 2019) and (Xu et al., 2020) investigate prediction entropy for uncertainty estimation on neural language generation tasks. (Xiao & Wang, 2021) quantifies epistemic and aleatoric uncertainty in natural language generation tasks to address the hallucination issues. In summary, most existing works leverage Monte-Carlo dropout or Bayes-by Backprop approaches, which require stochastic sampling. In contrast, evidence-based uncertainty estimation predicts model uncertainty in a deterministic way, which is more efficient and accurate.

Additional discussion about interactive machine learning are provided in the Appendix.

3. Preliminaries

We discuss the preliminaries of evidential theory, which is the building block for uncertainty estimation in our framework. The evidential theory is a generalization of Bayesian theory to subjective logic. For classification tasks with K mutually exclusive classes, subjective logic assigns a belief mass b_k to each possible class k for a data instance and introduces an overall uncertainty mass u. The belief mass values and uncertainty mass sum up to one,

$$u + \sum_{k=1}^{K} b_k = 1 \tag{1}$$

The belief mass is calculated using the evidence e_k where

$$b_k = \frac{e_k}{S}, \quad u = \frac{K}{S}, \quad S = \sum_{k=1}^{K} (e_k + 1) \tag{2}$$

Evidence e_k measures the amount of information that supports a data instance to be classified into class k. The belief mass assignment corresponds to a K-dimensional Dirichlet distribution $\text{Dir}(p|\mathbf{a})$ where $\mathbf{a} = (a_1, ..., a_K)^\top$ quantifies the strength over K classes and $a_k = e_k + 1$. The expected probability assigned to class k is the mean of the Dirichlet:

$$\mathbb{E}[p_k] = \frac{a_k}{S} \tag{3}$$

where a_k is usually referred to as the opinion for class k.

According to the subjective logic, there are two primary sources of uncertainty (Josang et al., 2018): vacuity and dissonance, that are applicable in classification tasks for our research problem. They are defined as

$$vac = u, \quad diss = \sum_k b_k \frac{\sum_{j \neq k} b_j (1 - \frac{|b_j - b_k|}{b_j + b_k})}{\sum_{j \neq k} b_j} \tag{4}$$

Intuitively, vacuity measures the lack of evidence in the data instance, and dissonance measures the contradictory evidence for different classes.

4. Methodology

The proposed framework formulates interactive medical image captioning in three stages: keyword prediction, caption generation, and model updates. In the first stage, the model leverages the image information to predict the evidence of candidate keywords and quantifies the uncertainty. Based on uncertainty estimation, the model selectively queries potential keywords to the user. In the second stage, the model

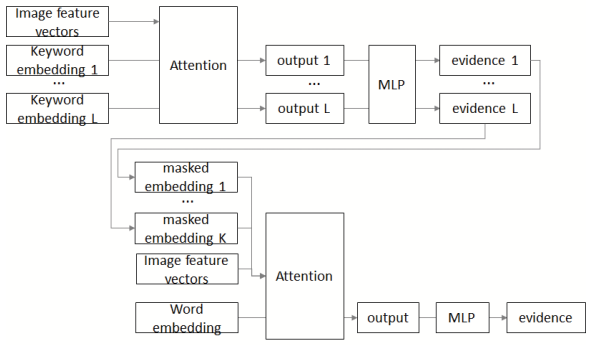


Figure 3: Architecture of the proposed framework. Given a medical image, the model outputs the evidence of keywords and captions, which serves as the foundation for keyword prediction, caption generation and uncertainty estimation.

leverages users' feedback and uncertainty estimation to generate masked embedding of keywords. The image feature vectors and masked embedding are integrated to predict the evidence of the words in captions. In the last stage, users' feedback on keywords and captions is used as weak supervision for model updates. The overall architecture is illustrated in Figure 3. The mathematical notations are summarized in the Appendix.

4.1. Keyword prediction

Keyword prediction is formulated as a multi-label classification problem with the goal of predicting a set of relevant keywords (*i.e.,* labels) for an input image. Let X be an image, and Y be a ground truth set of L binary labels $Y = \{y_l\}_{l=1}^{L}, y_l \in \{0,1\}$. Conventional multi-label classification aims to construct a classifier f to predict the probability of each label l given an image so that: $(p_1, ..., p_L)^\top = f(X)$. However, in the propsoed setting, the model predicts the evidence $\mathbf{e}_l$ of each potential keyword, which is the building block for uncertainty estimation.

Given an input image X, we leverage a convolutional neural network to extract feature maps $V^{\text{img}} \in \mathbb{R}^{h\times w\times d}$, where h, w, and d are the output height, width, and channel, respectively. We can then consider each vector $\mathbf{v}_i^{\text{img}} \in \mathbb{R}^d$ with $i \in [1, h \times w]$ to be representative of a sub-region that maps back to patches in the original image space. In addition, we consider keywords as a set of embeddings $V^{\text{key}} = \{\mathbf{v}_1^{\text{key}}, ..., \mathbf{v}_L^{\text{key}}\}$ with $\mathbf{v}_l^{\text{key}} \in \mathbb{R}^d$, which are the outputs of an embedding layer.

The keyword embeddings and image feature vectors are then fed to transformer blocks (Vaswani et al., 2017) for attention. Let $V = \{\mathbf{v}_1^{\text{img}}, ..., \mathbf{v}_{h\times w}^{\text{img}}, \mathbf{v}_1^{\text{key}}, ..., \mathbf{v}_L^{\text{key}}\}$ denote the image feature vectors and keyword embeddings. They are fed to the transformer block, where each member in H is transformed into query, key, and value vectors. A multi-head self-attention is used to integrate the semantic information of a keyword with other keywords and image feature vectors. The output of multi-head self-attention is passed to a feed-forward layer and layer normalization to generate the output $\mathbf{o}_l$.

$$\{\mathbf{o}_l^{\text{key}}\}_{l=1}^{L}, \{\mathbf{o}_i^{\text{img}}\}_{i=1}^{w\times h} = \text{selfAttn}(V) \tag{5}$$

Lastly, a feed-forward layer makes the final keyword predictions. Existing deep learning-based models typically use a sigmoid layer on top of the transformer blocks to predict the probability of each label. However, the sigmoid-based predictions may not provide uncertainty information because the sigmoid score is essentially a point estimation of the predictive distribution, and the sigmoid outputs may be over-confident in false prediction.

Evidential deep learning is a solution that overcomes the limitations of sigmoid-based predictions by providing a principled way to formulate the classification and uncertainty modeling jointly. Given a candidate keyword indexed by l for classification, a Dirichlet prior parameterized by $\boldsymbol{\alpha}_l$ is introduced to model the class probability, where

$$\boldsymbol{\alpha}_l = \mathbf{e}_l + \mathbf{1} = g(\mathbf{o}_l^{\text{key}}) + \mathbf{1} \tag{6}$$

with $\mathbf{o}_l$ being the output of the transformer block and g being the evidence function to keep evidence $\mathbf{e}_l$ non-negative. We use a feed-forward layer with ReLU activation for g. Given α_l, the predictive probability can be estimated using Eq (3), and the uncertainty can be estimated using Eq (4).

Model pre-training for keyword prediction. The model needs to be pre-trained to make keyword predictions given an image. With the setting of evidential learning, we apply the negative log-likelihood (nll) as the loss, which is minimized for learning evidence $\mathbf{e}_l$ by integrating out the predictive probability p (Sensoy et al., 2018):

$$\begin{aligned} \text{nll}_{\text{key}_l} &= -\log\left(\int \prod_{k=1}^{K} p_k \frac{1}{B(\boldsymbol{\alpha}_l)} \prod_{k=1}^{K} p_k^{a_{l,k}-1} dp\right) \\ &= \sum\nolimits_{k=1}^{K} y_{l,k}(\ln S_l - \ln(\alpha_{l,k})) \end{aligned} \tag{7}$$

where $y_{l,k}$ is an one-hot K-dimensional label for candidate keyword l ($K = 2$ for binary classification) S_l is the total strength of the Dirichlet distribution $\text{Dir}(p|\boldsymbol{\alpha}_l)$, which is parameterized by $\boldsymbol{\alpha}_l \in \mathbb{R}^K$, and S_l is defined as

$$S_l = \sum\nolimits_{k=1}^{K} \alpha_{l,k}, \quad \alpha_{l,k} = e_{l,k} + 1 \tag{8}$$

Based on the evidential theory, the $\alpha_{l,k}$ is determined by the predictive evidence $e_{l,k}$. During the inference, the predicted probability of the k-th class is $\hat{p}_k = \alpha_k / S$, and the predictive vacuity can be computed accordingly.

Intuitively, the predicted evidence should shrink to zero for a keyword if it cannot be correctly classified. Note that

a Dirichlet distribution with zero evidence, *i.e.,* $S = K$, corresponds to the uniform distribution and indicates total uncertainty, *i.e.,* $u = 1$. It is achieved by incorporating a Kullback-Leibler (KL) divergence term into the loss function that regularizes the predictive distribution and penalizes those divergences that do not contribute to data fit. With the regularization term, the loss function is modified as

$$L_{\text{key}_l} = \text{nll}_{\text{key}_l} + \lambda KL[\text{Dir}(\tilde{\boldsymbol{\alpha}}_l)||\text{Dir}(\mathbf{1})] \tag{9}$$

where $\text{Dir}(\mathbf{1})$ is the non-informative uniform Dirichlet distribution and $\mathbf{1}$ denotes a vector with all entries equal to 1. $\tilde{\boldsymbol{\alpha}}_l = y_l + (1 - y_l)\boldsymbol{\alpha}_l$ is the Dirichlet parameters after removal of the non-misleading evidence from predicted parameters $\boldsymbol{\alpha}_l$. λ is a hyper-parameter.

When trained with Eq (9), the model is encouraged to generate correct and strong evidence for in-distribution data by reducing the first term, and predict weak evidence for out-of-distribution data by reducing the second term. Therefore, the model can be calibrated so that it is confident in its accurate predictions on familiar data instances, and is uncertain on unfamiliar data instances.

4.2. Keyword Selection for User Interaction

Given a medical image, the conventional keyword annotation process requires the annotator to consider all candidate keywords, determine whether each is relevant to the image, and mark some keywords as positive and the rest as negative. However, such an annotation process incurs the burden of providing annotation. To address this issue, we propose to conduct an active keyword query to select a small number of candidate keywords that the model is mostly uncertain about to query the user. The user only needs to consider whether the query keywords are relevant or not.

Intuitively, user feedback on initially uncertain keywords will improve the model performance after updates. According to the evidential theory, vacuity and dissonance provide fine-grained quantification of two distinct sources of uncertainty. A straightforward way to design a query function for keyword selection is to aggregate both vacuity and dissonance by manually assigning weights. However, a challenge lies in how to properly balance these two sources of uncertainty in a principled way. Furthermore, as the model learns from user feedback and improves, the emphasis on different sources of uncertainty may need to be dynamically changed to adjust the learning focus.

To address the above challenge, we propose a keyword query function (`KQF`) that integrates vacuity and dissonance. We theoretically establish the connection between the proposed `KQF` and the two different sources of uncertainty. The theoretical analysis reveals that automatically assigned weights dynamically adjust the contribution of vacuity and dissonance according to a principled learning schedule that varies based on the model's accuracy. In addition, we find that for images with low quality, the model may predict some keywords correctly while missing some other keywords or predicting a few irrelevant keywords. In this case, it is helpful to collect human feedback on candidate keywords that the model is confused about, because the model can be updated to learn from humans about those keywords.

We first introduce a concept referred to as expected correct belief (`ECB`). Given a data sample, an evidential-learning model predicts the evidence for each class, which can be used to calculate the belief mass b_k for each class and the uncertainty mass u. Ideally, a well-calibrated model should assign a high belief to the ground-truth class, which indicates the model is making a confident and correct prediction. However, in an interactive learning setting, the ground truth of a new data sample is not known beforehand. In this case, we propose to use the expectation of belief assigned to the correct class to evaluate a model's prediction. Assuming the model's prediction accuracy is p, *i.e.,* the probability of the model making correct predictions is p, which can be estimated using a hold-out validation set. The expectation of correct belief is

$$\texttt{ECB} = p \max(b_k) + (1 - p) \sum_{j \neq \arg\max b_k} p_j b_j \tag{10}$$

where p_j is the probability that the ground-truth class being j if the model's prediction is incorrect, and $\sum_{j \neq \arg\max b_k} p_j = 1$. Without further information, we assume a no-informative prior for $\{p_j\}$, and thus the expectation can be written as

$$\begin{aligned} \texttt{ECB} &= p \max(b_k) + \frac{1-p}{K-1} \sum_{j \neq \arg\max b_k} b_j \\ &= \frac{1-p}{K-1}(1-u) + (p - \frac{1-p}{K-1}) \max(b_k) \end{aligned} \tag{11}$$

Intuitively, if `ECB` is low, the model is likely to be unfamiliar (*i.e.,* uncertain) with a candidate keyword for a given image. In this case, it is suggested to query the user for this keyword. Practically, we can rank the keywords based on `ECB`, and choose the keyword with the minimum `ECB` for the query. This is equivalent to the largest expected wrong belief (`EWB`),

$$\texttt{EWB} = 1 - \texttt{ECB} \tag{12}$$

In order to further explain why `EWB` provides a principled keyword function, we theoretically demonstrate its connection to both types of uncertainty. First, we present Theorem 1, which shows that `EWB` is upper bounded by a weighted sum of vacuity and dissonance in a multi-class setting.

Theorem 1. *For a classification problem with K classes, the expectation of wrong belief (`EWB`) is upper bounded by* $\texttt{EWB} \leq (1 - p) + (w_v \times vac + w_d \times diss)$*, where* $w_v = p, w_d = \frac{1}{2}(p - \frac{1-p}{K-1})$.

Given the multi-label classification problem for keyword prediction, each label corresponds to a binary task with $K = 2$ classes. We further introduce the following theorem:

Theorem 2. *For a classification problem with $K = 2$ classes, the upper bound of* EWB *is tight as* $\texttt{EWB} = (1-p) + (w_v \times vac + w_d \times diss)$*, where* $w_v = p, w_d = (p - 0.5)$.

The proofs of both theorems are provided in the Appendix. Theorem 2 provides a selection criterion to select the most uncertain keywords to collect user annotation. The model is likely to be unfamiliar (*i.e.,* uncertain) with a candidate keyword if EWB is high. Collecting user feedback on those keywords and updating the model may effectively improve model performance.

Note that given the current model, p is fixed for all keywords. So keyword selection is essentially based on $w_v \times vac + w_d \times diss$. Furthermore, we can normalize the weights (w_v, w_s) and introduce a query function KQF as

$$\texttt{KQF} = \bar{w}_v \times vac + \bar{w}_d \times diss \tag{13}$$

where $\bar{w}_v = \frac{p}{2p-0.5}$ and $\bar{w}_d = \frac{p-0.5}{2p-0.5}$ are normalized.

Remark: We provide a discussion about the meaning of $\bar{w}_v$ and $\bar{w}_d$ and how they dynamically change as interactive learning continues. At the beginning of the interactive learning process, the model is trained with limited data. Therefore the prediction accuracy p may be low, and $\bar{w}_v$ may be high. After the model is trained with more data, p increases, $\bar{w}_v$ decreases, and $\bar{w}_d$ increases. Intuitively, the interactive learning process should rely more on vacuity in the early phase, which can effectively shape the decision boundary. As the learning process goes on, dissonance should gradually gain a higher weight. It allows the model to fine-tune the decision boundary with the right shape but is less accurate, aiming to maximize the discriminate power of the model.

4.3. Caption Generation

The uncertainty estimation and user feedback can be integrated to guide the model to generate the captions. Specifically, we introduce an uncertainty-aware weighting mechanism to control how the transformer can attend to different keywords when generating captions. Using (13) for uncertainty estimation, the weighting factor for keyword l is defined as

$$m_l = \begin{cases} 1 & \text{if approved by user} \\ 0 & \text{if rejected by user} \\ \delta(p_l > 0.5)(1 - unc_l) & \text{otherwise} \end{cases} \tag{14}$$

Intuitively, suppose the user specifies that a keyword is relevant to the image. In that case, it should receive a higher weight $m_l = 1$ for masking. For a rejected keyword, the weight is set to $m_i = 0$, and the corresponding keyword has no contribution to downstream caption generation. For other keywords, if a keyword is relevant to the image, the predicted probability must be greater than 0.5. However, if the prediction is highly uncertain, then the model shall pay little attention. To this end, a high uncertainty makes the multiplicative term $(1 - unc_l)$ close to zero. The weighted embeddings of keywords are concatenated with image feature vectors and then used for cross attention during caption generation, where $H = \{\mathbf{v}_1, ..., \mathbf{v}_{h\times w}, m_1\mathbf{q}_1, ..., m_L\mathbf{q}_L\}$.

The downstream caption generation also leverages transformer blocks to predict words in the caption in an autoregressive way. The setting of the transformer blocks is similar to that used for keyword prediction. First, the embedding of words in the current incomplete sentence (plus positional embedding) is fed into a transformer block with attention. For self-attention, each word in the current incomplete sentence attends to other words. After that, the output is fed into another transformer block with cross-attention. For cross-attention, each word in the current incomplete sentence attends to keywords and image feature vectors. Finally, the model predicts the evidence of the next word in the caption.

Practically, we apply nucleus sampling and select the candidate captions with high joint probability to present to the user. The user can select the best caption and pass it back to the model as feedback. User feedback on keywords and captions is leveraged as weak supervision to update the model for interactive learning, which is discussed below.

Model pre-training for caption generation. The model needs to be pre-trained for caption generation, and the evidential learning technique is applied for uncertainty estimation. Recall that caption generation can be considered a sequential multi-class classification problem where the model predicts the next word from the vocabulary. Therefore, the evidential loss for caption prediction is

$$\begin{aligned} L_{\text{cap}_t} =& \sum\nolimits_{v=1}^{V} y_{t,v}(\ln S_t - \ln(\alpha_{t,v})) \\ &+ \lambda KL[\text{Dir}(\widetilde{\boldsymbol{\alpha}}_t)||\text{Dir}(\mathbf{1})] \\ \alpha_{t,v} =& e_{t,v} + 1 \end{aligned} \tag{15}$$

where V is the size of vocabulary, $e_{t,v}$ is the predicted evidence of word v at position t of the caption, and $y_{t,v}$ is the ground-truth.

4.4. Interactive Learning from Users

It would be ideal if the model could learn from the user so that the model's performance is improved for tasks in the future. User feedback about candidate keywords and captions can be leveraged as additional supervision to update the model. However, there are several critical issues to be addressed. 1) Users only provide feedback on queried keywords (*i.e.,* only a subset of the keyword vocabulary). Simple model updates using those keywords may cause the

model to overfit the limited data. 2) Although the user-provided caption can be considered the ground truth (*i.e.,* gold standard), an image may be described in multiple ways. Simple model updates using those captions may limit the model's capacity to generate diverse captions.

To address those issues, other keywords and the captions predicted by the model are incorporated as weak supervision. However, such predictions are not the ground truth, and the model is not equally confident in predicting different keywords and captions. To this end, the uncertainty estimation can be leveraged to down-weigh the predictions with high uncertainty during updates.

Evidential learning provides a natural way to quantify uncertainty. At inference, we can record the uncertainty estimation of all keywords. During updates, both user feedback on query keywords and the model's prediction on other keywords are treated as supervised labels but assigned different weights. And the objective function is modified as

$$\begin{aligned} L_{\text{key}_l} &= \sum\nolimits_{k=1}^{K} \tilde{y}_{l,k}(\ln S_l - \ln(\alpha_{l,k})) \\ &\quad + \lambda KL[\text{Dir}(\tilde{\boldsymbol{\alpha}}_l)||\text{Dir}(\mathbf{1})] \\ \tilde{y}_{l,k} &= \begin{cases} y_{l,k} & \text{for keywords with user feedback} \\ \hat{y}_{l,k}(1 - unc_l) & \text{otherwise} \end{cases} \end{aligned} \tag{16}$$

where $\hat{y}_{l,k}$ is the model's prediction of keywords not queried to the user. The weighting factor with $(1 - unc_l) \rightarrow 1$ for confident predictions and $\rightarrow 0$ for uncertain predictions.

In addition, an image can usually be described in multiple ways. During inference, the model predicts multiple caption candidates via the nucleus sampling, which can be used for model updates along with the user-approved caption. Specifically, we book keep multiple predicted captions with high probability as $\{\hat{Y}_1, \hat{Y}_2, ...\hat{Y}_B\}$ where B is the size of the candidate set. For each word in the candidate captions, we record the uncertainty estimation. During updates, the captions predicted by the model receive different weights based on uncertainty estimation, and the objective function is similar to Eq (16).

5. Experiments

We evaluate the proposed method on medical image captioning datasets. The IU X-RAY (Young et al., 2014) dataset includes a collection of radiology examinations, including images and narrative reports by radiologists. The PEIR Gross (Library, 2022) dataset is released by the Pathology Education Informational Resource digital library and includes teaching images of gross lesions along with their associated captions. Each image in the two datasets is associated with a number of tags, which are considered ground-truth keywords.

5.1. Experimental Settings

The model needs to be pre-trained to predict candidate keywords and captions for user feedback. To this end, we randomly split the two datasets and used forty percent of the images and corresponding captions and keywords for updates. After that, we involve user interactions for four batches of data, each corresponding to ten percent of the data. During the interaction, the model selects eight candidate keywords with the highest uncertainty score to collect user feedback. After receiving user feedback, the model leverages the keywords to generate captions via nucleus sampling and present the candidate captions to the user. In addition, the user can select the best caption from the candidate set, and the feedback is saved for model updates. Since involving actual users may be costly and time-consuming, simulated user interaction can be used for interactive learning (Wu et al., 2022). In our experiment, we simulate user interactions by assuming the user's opinion is the same as the ground truth keywords and captions from the dataset. Given an image, the simulated user approves query keywords matching the ground truth, rejects other keywords, and selects the caption with the highest METEOR score. Once a batch of data is processed, the model is updated based on users' feedback, and the model is evaluated on its performance on the hold-out test set, which includes the remaining data. The evaluation is conducted on both the keyword prediction task and the caption generation task, where we make the updated model generate automated keyword and caption predictions and compare them with the ground truth. For keyword prediction, we evaluate the mean average precision and F1 scores. For caption generation, we assess the quality of generated captions based on BLEU, ROUGE, and METEOR scores. BLEU is a precision-based metric that evaluates the matching of n-grams in texts. ROUGE is a recall-based metric that focuses on important words and phrases. METEOR also considers synonyms and word order when matching words and phrases. The details of those metrics can be found at (Hossain et al., 2019).

For experiments, the proposed method and baselines are trained with Intel Core i7-3820 CPU and NVIDIA GeForce RTX2070 GPU. We use five-fold cross-validation for hyper-parameter tuning. We use the architecture of EfficientNet (Tan & Le, 2019) for the image feature extractor. The embedding dimension is tuned via grid search and set to 512, and the number of attention heads is tuned and set to 2. The number of stacked transformer blocks for keyword prediction and caption generation is tuned and set to 4. λ is set to 1. We use stochastic gradient descent and Adam optimizer with a learning rate scheduled from 5e-5 to 1e-5.

5.2. Comparison Baselines

We compare with representative keyword-driven medical image captioning baselines that jointly perform keyword

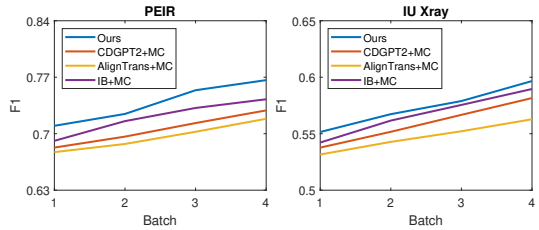


Figure 4: Quantitative comparison for keyword prediction

prediction and caption generation. CDGPT (Alfarghaly et al., 2021) is a conditioned transformer-based model that integrates image feature embeddings, predicted keyword embeddings, and token embeddings for self-attention and caption prediction. AlignTrans (You et al., 2021) aligns visual regions from medical images and predicted keywords to provide better representation learning and enhance caption generation. Image captioning with Interpretability Boosters leverages an attention-based strategy in the caption generation process to match expert-defined keywords with local image patches (Wu et al., 2023). For interactive learning settings, the above baselines are integrated with alternative uncertainty estimation methods, including the Bayesian neural network with Monte-Carlo (MC) dropout for uncertainty estimation (Xiao & Wang, 2019), and Bayes By Backprop (Siddhant & Lipton, 2018). Those methods are used to select keywords for user feedback and assign weights to data samples for model updates. For MC dropout, we add a dropout layer after each convolutional block and set the dropout rate to 0.2, a widely used setting. For Bayes By Backprop, we add Gaussian noise to the feed-forward layers in the transformer blocks and reparameterize the standard deviation of the noise as learnable parameters. The proposed method and baselines are trained with the same data split.

5.3. Comparison Results

We first present the results for model performance on keyword prediction. Quantitative comparisons are provided in Figure 4. We observe an upward trend of precision which indicates that all methods learn from the user feedback in multiple batches to improve their performance on the test data. The proposed framework outperforms other baselines. A possible reason is that our framework effectively leverages evidence-based uncertainty estimation to select the most informative keywords for user feedback. In addition, the model can be effectively updated to learn from users. The keyword prediction performance on the test set after the models are updated after four batches of interactive learning are reported in Table 1. In general, the proposed method outperforms baselines in most cases. Empirically, we observe that the model performs reasonably well on keyword prediction. For some images with low quality, the vision model typically predicts some keywords correctly while missing some other keywords or predicting a few irrelevant keywords. In this case, it is helpful to collect human

Table 1: Quantitative comparison for keyword prediction

Model	PEIR		IU-Xray	
	mAP	F1	mAP	F1
CDGPT+MC	0.823	0.729	0.660	0.574
CDGPT+Bayes	0.815	0.720	0.657	0.565
AlignTrans+MC	0.797	0.717	0.648	0.557
AlignTrans+Bayes	0.806	0.712	0.640	0.538
IB+MC	0.841	0.742	0.685	0.584
IB+Bayes	0.835	0.736	0.679	0.566
Proposed	0.851	0.766	0.698	0.591

Table 2: Quantitative comparison on PEIR

Model	PEIR		
	BLEU	ROUGE	METEOR
CDGPT+MC	0.134	0.315	0.141
CDGPT+Bayes	0.129	0.312	0.138
AlignTrans+MC	0.119	0.295	0.135
AlignTrans+Bayes	0.125	0.288	0.131
IB+MC	0.135	0.318	0.145
IB+Bayes	0.130	0.309	0.138
Proposed	0.142	0.330	0.156

Table 3: Quantitative comparison on IU-Xray

Model	IU-Xray		
	BLEU	ROUGE	METEOR
CDGPT+MC	0.146	0.341	0.150
CDGPT+Bayes	0.142	0.336	0.148
AlignTrans+MC	0.135	0.328	0.142
AlignTrans+Bayes	0.138	0.332	0.144
IB+MC	0.145	0.342	0.153
IB+Bayes	0.149	0.345	0.154
Proposed	0.157	0.356	0.162

feedback on candidate keywords that the model is confused about. We then present the experiment results for caption generation. Quantitative comparisons on the test set after the models are updated for interactive learning are provided in Tables 2, and 3. The proposed framework outperforms other baselines. A possible explanation is that our framework generates predicted keywords with better quality which benefits downstream caption generation. Generally speaking, the task of predicting keywords is relatively easier than predicting the entire caption, because caption generation requires extracting almost all the semantic information from the image, while keyword prediction requires extracting only the most important features.

Quantitative comparisons for caption generation with respect to interactive learning batches are provided in Figure 5. We observe an upward trend in the scores of generated captions for testing images, and the proposed framework outperforms other baselines. The results can be intuitively explained by the difference in uncertainty estimation. The proposed method quantifies the uncertainty of each word in the captions via vacuity and dissonance, which can be used to weigh data samples during model updates effectively. In contrast, MC dropout and Bayes By Backprop require

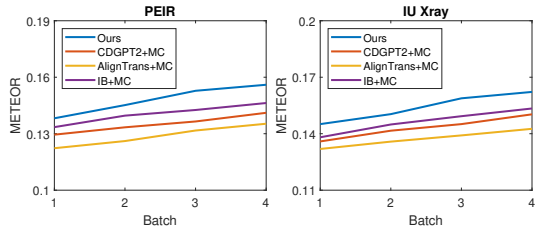


Figure 5: Quantitative comparison for caption prediction

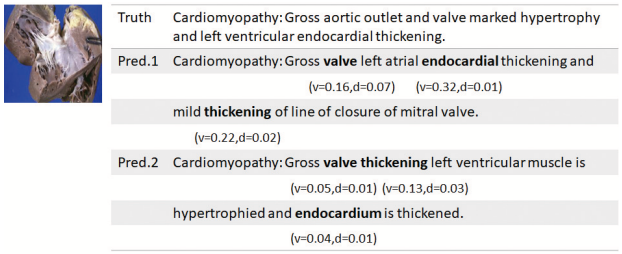


Figure 6: Illustrative example of caption prediction with the change of uncertainty due to interactive learning (Pred.1 and Pred.2 are the predictions before and after model updates)

stochastic sampling, which may be prone to errors.

One notable benefit of the proposed interactive learning is improving the model's performance and confidence in its predictions. We provide illustrative examples in Figure 6. Before user interaction, we make the model generate captions given an image (Pred.1), and report the uncertainty score of important medical terms in the generated captions. After interactive learning for four batches, we make the model generate captions again and report the corresponding uncertainty (Pred.2). The examples indicate that the updated model generates better captions. The vacuity of all key medical terms is decreasing. This is consistent with the intuition: After model updates based on user feedback, the model is better calibrated and familiar with the data. On the other hand, we observe that the dissonance is not necessarily decreasing. It is because dissonance measures the contradictory evidence from the data samples. In other words, dissonance essentially captures the uncertainty from the data, which may not decrease after model updates.

5.4. Ablation Study

Our model leverages evidential learning as the foundation of uncertainty estimation and query keywords for user feedback. We conduct an ablation study to compare with alternative methods of uncertainty estimation, including the Bayesian neural network with Monte-Carlo dropout and Bayes By Backprop to evaluate the contribution of the proposed keyword selection method. In addition, we also compare with the vanilla evidential deep learning method (EDL) for uncertainty estimation (Sensoy et al., 2018). Note that the proposed method differs from vanilla evidential learning

Table 4: Ablation study for keyword prediction

Model	PEIR		IU-Xray	
	mAP	F1	mAP	F1
MC	0.835	0.746	0.689	0.575
Bayes By-backprop	0.827	0.739	0.683	0.568
EDL	0.846	0.757	0.692	0.577
Proposed	0.851	0.766	0.698	0.591

Table 5: Ablation study for caption generation

PEIR	BLEU	ROUGE	METEOR
MC	0.135	0.318	0.149
Bayes By-backprop	0.130	0.311	0.147
EDL	0.139	0.325	0.152
Proposed	0.142	0.330	0.156
IU X-ray	BLEU	ROUGE	METEOR
MC	0.145	0.343	0.155
Bayes By-backprop	0.149	0.347	0.157
EDL	0.156	0.353	0.163
Proposed	0.157	0.356	0.162

because the former quantifies the uncertainty using vacuity and dissonance and dynamically balances their corresponding weights during interactive learning. Quantitative results are reported in Tables 4 and 5, which indicate that the proposed keyword selection method achieves good performance. In contrast, vanilla evidential learning does not distinguish between vacuity and dissonance. MC dropout and Bayes By Backprop require stochastic sampling, which is less efficient and prone to errors. In the Appendix, we also provide additional experiment results on sub-categories of medical images, and ablation studies with the number of keywords.

6. Conclusion

In this paper, we propose an interactive learning framework that involves human users in the loop to improve model performance on medical image captioning tasks. The framework deploys an evidence-based uncertainty estimation to select the most informative keywords to query users for feedback, which are then used as weak supervision to update the model and encode users' knowledge. In addition, uncertainty estimations are leveraged as weighting factors to guide the self-supervision process during model updates to mitigate the overfitting issue. The framework can be potentially applied to medical domains for interactive learning.

Acknowledgements

This research was partially supported by NSF IIS award IIS-1814450 and ONR award N00014-18-1-2875. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the official views of any funding agency. We thank the anonymous reviewers for reviewing the manuscript.

References

Alfarghaly, O., Khaled, R., Elkorany, A., Helal, M., and Fahmy, A. Automated radiology report generation using conditioned transformers. *Informatics in Medicine Unlocked*, 24:100557, 2021.

Biswal, S., Xiao, C., Glass, L. M., Westover, B., and Sun, J. Clara: clinical report auto-completion. In *Proceedings of The Web Conference 2020*, pp. 541–550, 2020.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. Weight uncertainty in neural network. In *International conference on machine learning*, pp. 1613–1622. PMLR, 2015.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Hossain, M. Z., Sohel, F., Shiratuddin, M. F., and Laga, H. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36, 2019.

Huang, J.-H., Wu, T.-W., Yang, C.-H. H., and Worring, M. Deep context-encoding network for retinal image captioning. In *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 3762–3766. IEEE, 2021.

Huang, X., Yan, F., Xu, W., and Li, M. Multi-attention and incorporating background information model for chest x-ray image report generation. *IEEE Access*, 7:154808–154817, 2019.

Jing, B., Xie, P., and Xing, E. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*, 2017.

Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Mark, R. G., and Horng, S. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

Josang, A., Cho, J.-H., and Chen, F. Uncertainty characteristics of subjective opinions. In *2018 21st International Conference on Information Fusion (FUSION)*, pp. 1998–2005. IEEE, 2018.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Li, C. Y., Liang, X., Hu, Z., and Xing, E. P. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 6666–6673, 2019.

Li, Y., Liang, X., Hu, Z., and Xing, E. P. Hybrid retrieval-generation reinforced agent for medical image report generation. *Advances in neural information processing systems*, 31, 2018.

Library, P. D. Peir digital library, 2022. URL https://peir.path.uab.edu/library/.

Maksoud, S., Wiliem, A., Zhao, K., Zhang, T., Wu, L., and Lovell, B. Coral8: Concurrent object regression for area localization in medical image panels. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 432–441. Springer, 2019.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*, 2019.

Pavlopoulos, J., Kougia, V., and Androutsopoulos, I. A survey on biomedical image captioning. In *Proceedings of the second workshop on shortcomings in vision and language*, pp. 26–36, 2019.

Pavlopoulos, J., Kougia, V., Androutsopoulos, I., and Papamichail, D. Diagnostic captioning: a survey. *Knowledge and Information Systems*, pp. 1–32, 2022.

Saidu, I. C. and Csató, L. Active learning with bayesian unet for efficient semantic image segmentation. *Journal of Imaging*, 7(2):37, 2021.

Sensoy, M., Kaplan, L., and Kandemir, M. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

Siddhant, A. and Lipton, Z. C. Deep bayesian active learning for natural language processing: Results of a large-scale empirical study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2904–2909, 2018.

Tan, M. and Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, J., Yan, Y., Zhang, Y., Cao, G., Yang, M., and Ng, M. K. Deep reinforcement active learning for medical

image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 33–42. Springer, 2020.

Wang, K., Zhang, D., Li, Y., Zhang, R., and Lin, L. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2591–2600, 2017.

Wang, S., Liu, Y., Wang, C., Luan, H., and Sun, M. Improving back-translation with uncertainty-based confidence estimation. *arXiv preprint arXiv:1909.00157*, 2019.

Wu, T.-W., Huang, J.-H., Lin, J., and Worring, M. Expert-defined keywords improve interpretability of retinal image captioning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1859–1868, 2023.

Wu, X., Chen, C., Zhong, M., Wang, J., and Shi, J. Covid-al: The diagnosis of covid-19 with deep active learning. *Medical Image Analysis*, 68:101913, 2021.

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., and He, L. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 2022.

Xiao, Y. and Wang, W. Y. Quantifying uncertainties in natural language processing tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7322–7329, 2019.

Xiao, Y. and Wang, W. Y. On hallucination and predictive uncertainty in conditional language generation. *arXiv preprint arXiv:2103.15025*, 2021.

Xu, J., Desai, S., and Durrett, G. Understanding neural abstractive summarization models via uncertainty. *arXiv preprint arXiv:2010.07882*, 2020.

Yang, L., Zhang, Y., Chen, J., Zhang, S., and Chen, D. Z. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 399–407. Springer, 2017.

You, D., Liu, F., Ge, S., Xie, X., Zhang, J., and Wu, X. Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 72–82. Springer, 2021.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Zhang, Y., Wang, X., Xu, Z., Yu, Q., Yuille, A., and Xu, D. When radiology report generation meets knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 12910–12917, 2020.

Appendix

Organization of Appendix. In this Appendix, we first summarize the main notations used throughout the paper in Table 6. We then provide the mathematical proof to Theorems 1 and 2. After that, we provide additional discussion about interactive machine learning. We provide additional experiment results on sub-categories of medical images and ablation studies with different annotation budgets by varying the number of keywords. We also discuss the limitations and future directions. The link to the source code is provided at the end.

Table 6: Summary of Main Notations

Notation	Description
X	input image
V	total number of keywords
L	size of vocabulary
K	number of classes
v_i^{img}	i-th vector of image feature
v_l^{key}	embedding vector of l-th keyword
o_i^{img}	output of transformer block corresponding to i-th vector of image feature
o_l^{key}	output of transformer block corresponding to embedding vector of l-th keyword
$\alpha_{l,k}$	the subjective opinion of keyword l ($k = \{0,1\}$ indicating positive or negative)
$e_{l,k}$	predicted evidence of keyword l
S_l	total strength of Dirichlet distribution of keyword l
p_l	predicted probability of keyword l
$\alpha_{t,v}$	the subjective opinion of word v at position t
$e_{t,v}$	predicted evidence of word v at position t
$y_{t,v}$	ground truth of word v at position t
b_j	predicted belief of class j
b_0,b_1	predicted belief for binary classification of a keyword
u	uncertainty mass
p	probability of a model making correct prediction
unc_l	uncertainty score for keyword l

A. Proof of Theorems

Proof of Theorem 1. To connect the expectation of belief to the two sources of uncertainty, we consider the definition of vacuity and dissonance as

$$vac = u, \quad diss = \sum_k b_k \frac{\sum_{j\neq k} b_j (1 - \frac{|b_j - b_k|}{b_j + b_k})}{\sum_{j\neq k} b_j} \tag{17}$$

After sorting $\{b_j\}$ in descending order, it can be shown that

$$\begin{aligned}
&\frac{1}{2}diss \\
=&\frac{b_1}{\sum_{k\neq 1} b_k}(\frac{b_2^2}{b_1+b_2} + \frac{b_3^2}{b_1+b_3} + ... + \frac{b_K^2}{b_1+b_K}) \\
&+ \frac{b_2}{\sum_{k\neq 2} b_k}(\frac{b_1 b_2}{b_1+b_2} + \frac{b_3^2}{b_1+b_3} + ... + \frac{b_K^2}{b_1+b_K}) \\
&+ ... \\
&+ \frac{b_K}{\sum_{k\neq K} b_k}(\frac{b_1 b_K}{b_1+b_K} + \frac{b_2 b_K}{b_1+b_K} + ... + \frac{b_{K-1} b_K}{b_{K-1}+b_K}) \\
=&\sum_{k=2}^{K} b_k^2 [\sum_{j=1}^{k-1} \frac{b_j}{b_j+b_k}(\frac{1}{\sum_{l\neq j} b_l} + \frac{1}{\sum_{l\neq k} b_l})] \\
\leq&\sum_{k=2}^{K} b_k [\sum_{j=1}^{k-1} \frac{b_j}{\sum_{l\neq k} b_k}] \\
\leq&\sum_{k=2}^{K} b_k \\
=&1 - u - b_1
\end{aligned} \tag{18}$$

Therefore, the expectation can be expanded as

$$\begin{aligned}
E =& \frac{1-p}{K-1}(1-u) + (p - \frac{1-p}{K-1})\max(b_k) \\
\geq& p - p * vac - \frac{1}{2}(p - \frac{1-p}{K-1}) * diss
\end{aligned} \tag{19}$$

Proof of Theorem 2. Denote the prediction accuracy of the model as p. For classification with number of classes $K = 2$,t he expectation of correct belief is

$$E = p\max(b_k) + (1-p)(1 - u - \max(b_k)) \tag{20}$$

Note that for $K = 2$, the dissonance is reduced to

$$\begin{aligned}
\frac{1}{2}diss &= \frac{1}{2}(b_0 + b_1 - |b_0 - b_1|) \\
&= 1 - u - \max(b_k)
\end{aligned} \tag{21}$$

Therefore, the expectation can be expanded as

$$E = p - p * vac - (p - \frac{1}{2}) * diss \tag{22}$$

It should be noted that Eq 22 is the tight bound of Eq 19 for $K = 2$.

B. Additional Discussion of Interactive Learning

Interactive machine learning aims to integrate human knowledge and experience to train machine learning models effectively. Humans can be involved in the loop in multiple

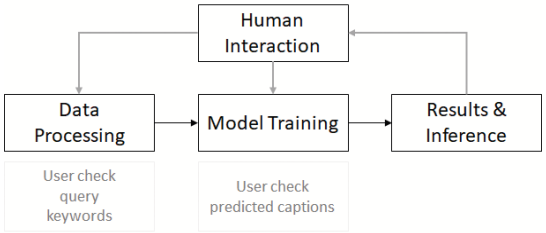


Figure 7: Illustration of general interactive machine learning: humans may involve in data processing and model training. For our framework, the user checks query keywords to provide annotation for those keywords (data processing), and checks predicted captions to provide feedback (model training)

ways, including data processing, interventional model training, and the design of the system (Wu et al., 2022). For interactive learning on medical image understanding tasks, existing works mainly focus on image classification (Wang et al., 2017; 2020; Wu et al., 2021) and segmentation (Yang et al., 2017; Saidu & Csató, 2021), while interactive learning for captioning is an under-explored area.

For data processing, interactive methods emphasize finding important data samples. This is similar to active learning, where the goal is to train an accurate prediction model with the least cost by annotating the data samples that provide the most information. However, interactive learning also emphasizes adding human knowledge to the learning system and facilitating human-machine interaction. For our framework, user feedback on query keywords is essentially annotated labels, which can be used for model updates.

For model training, human participants provide feedback according to specific tasks to boost performance. For instance, an object detection framework may employ individuals to correct a few annotations proposed by a detector, and an online question-answering model may seek human feedback to update the model continuously (Wu et al., 2022). For our framework, the user checks generated captions to provide feedback on model predictions, which can be used for model updates to improve model performance.

In the proposed framework, the model selects the most uncertain keywords for user interaction. The user annotates whether each selected keyword is relevant to the image or not. User annotation is used for model updates to improve performance on keyword prediction. The model also uses nucleus sampling to generate a list of predicted captions for user interaction. User feedback is used for model updates to generate better captions. For uncertainty estimation, evidential learning effectively quantifies uncertainty to assign different weights to different keywords and captions during model updates in order to reduce noise. The user only needs to annotate the selected uncertain keywords (rather than all keywords), which reduces users' burden.

C. Additional Results

It should be noted that some datasets contain images with different pre-existing conditions, which may be considered subcategories of medical images. Based on the subcategories in the IU-Xray dataset (*e.g.,* pleural diffusion, cardiomegaly, nodule), we evaluate model performance of caption prediction, and the results are summarized in Table 7. In general, the difficulty of captioning tasks on some subcategories may be greater than the difficulty on others, and the proposed method usually outperforms competing baselines.

Table 7: Quantitative comparison on different subcategories of medical images

Nodule	BLEU	ROUGE	METEOR
CDGPT+MC	0.137	0.330	0.144
CDGPT+Bayes	0.129	0.324	0.142
AlignTrans+MC	0.131	0.318	0.140
AlignTrans+Bayes	0.125	0.312	0.135
Proposed	0.153	0.352	0.158
Cardiomegaly	BLEU	ROUGE	METEOR
CDGPT+MC	0.144	0.340	0.147
CDGPT+Bayes	0.145	0.351	0.149
AlignTrans+MC	0.137	0.335	0.141
AlignTrans+Bayes	0.142	0.339	0.143
Proposed	0.161	0.362	0.165
Pleural Diffusion	BLEU	ROUGE	METEOR
CDGPT+MC	0.139	0.335	0.145
CDGPT+Bayes	0.142	0.339	0.151
AlignTrans+MC	0.132	0.324	0.140
AlignTrans+Bayes	0.129	0.317	0.139
Proposed	0.151	0.345	0.156

We also explored model performance on an additional radiography image dataset MIMIC-CXR (Johnson et al., 2019) for caption generations. Quantitative results are summarized in Table 8. The performance of the proposed model is better than the competing baselines, which is consistent with the experimental evaluation of other datasets presented in the paper.

Table 8: Quantitative comparison on MIMIC-CXR

Model	BLEU	ROUGE	METEOR
CDGPT+MC	0.101	0.298	0.130
CDGPT+Bayes	0.108	0.302	0.117
AlignTrans+MC	0.095	0.287	0.124
AlignTrans+Bayes	0.104	0.295	0.113
Proposed	0.112	0.316	0.139

In addition, we conduct ablation studies to examine the effect of annotation budget by varying the number of keywords. The number of keywords is changed to 4 and 12. Results on the PEIR dataset are provided below. With four

keywords, the performance dropped significantly, indicating that it is difficult to collect sufficient feedback with only four keywords. With 12 keywords, the performance is slightly improved. The two datasets for experiments include the ground truth keywords for each image, and one image typically corresponds to 3-5 keywords. There are more than 50 keywords in total in the datasets. In real-world applications, when setting the number of candidate keywords, it is suggested to consider the following factors: 1) the number of positive keywords an image typically corresponds to, 2) too few proposed keywords may hurt model performance, and 3) too many proposed keywords may incur annotation burdens.

Table 9: Ablation study on the number of keywords

4 keywords	BLEU	ROUGE	METEOR
CDGPT+MC	0.118	0.297	0.134
CDGPT+Bayes	0.115	0.289	0.128
AlignTrans+MC	0.114	0.285	0.126
AlignTrans+Bayes	0.109	0.277	0.121
Proposed	0.125	0.306	0.141
12 keywords	BLEU	ROUGE	METEOR
CDGPT+MC	0.131	0.317	0.144
CDGPT+Bayes	0.134	0.314	0.139
AlignTrans+MC	0.123	0.304	0.135
AlignTrans+Bayes	0.126	0.298	0.133
Proposed	0.146	0.343	0.158

D. Limitations and Future Directions

Our framework involves humans in the loop for medical image captioning tasks, where the model interacts with the user to collect user-approved keywords and captions. Since our framework is interactive, the model may be misguided and generate inaccurate predictions if the user provides irrelevant or wrong keywords. Therefore, it is suggested that the users understand the interactive process and the captioning tasks before using the proposed framework.

An alternative option to interactive learning is annotating a large amount of data for model training to provide a strong supervision model. A cost-benefit analysis is suggested to compare the cost estimation of training a strong supervision model and having the expert in the loop to provide feedback, and we consider it as a future direction.

Data augmentation is another future direction to generate rich captions for limited data to train a model properly. Some representative data augmentation methods for generating multiple captions include 1) back translation: translating a caption into another language and then translating it back 2) synonyms replacement: replacing certain words from the caption with their synonyms. There are some challenges in applying the two augmentation methods to medical image captioning datasets. Different from texts in general domains, texts in medical domains are specialized, back translation may incur some errors, and it might not be easy to find synonyms.

E. Link to the Source Code

The source code is provided at https://github.com/ritmininglab/EIL-MIC