Hierarchical Multi-Source Uncertainty Aggregation for Interactive Video Captioning

Ervine Zheng, Qi Yu

Rochester Institute of Technology
mxz5733@rit.edu, qi.yu@rit.edu

Abstract

Video captioning automatically generates natural language phrases to explain the contents in video frames. When deploying captioning models in specialized domains, active learning can help reduce the high annotation cost. However, the generative nature of the captioning process is more complex than standard supervised learning tasks and introduces several challenges for active learning in video captioning. Entropy-based uncertainty estimation, which is widely used in active learning, may be inflated in captioning tasks and mislead active sampling. Another challenge arises from the rich content of videos, as each video could be described in multiple ways. A single uncertainty score obtained from one possible caption does not capture the diversity induced by the rich content. To fill out this gap, we propose identifying multiple sources of uncertainty and performing hierarchical aggregation to integrate uncertainty from distinct sources. This innovates a holistic uncertainty metric to quantify the overall informativeness of video content for active sampling. The overall uncertainty is built upon conditional vacuity, an extension of the second-order uncertainty introduced along with the evidential learning framework to the captioning setting, leading to more robust uncertainty estimation without inflation. Both theoretical analysis and experimental evaluation are conducted to demonstrate the effectiveness of the proposed framework for complex uncertainty estimation and interactive learning.

Introduction

Video captioning is a challenging vision-language task because videos typically manifest complex spatiotemporal information, and videos can also be significantly variable in terms of content, style, and quality (Li et al. 2019; Islam et al. 2021). It typically leverages a large amount of paired video and annotated text for model training. To collect annotated data, crowdsourcing from the web is a typical solution. However, those approaches may not be applicable in specialized domains, such as public safety and surveillance videos, due to insufficient data available. Another solution is hiring human annotators to watch videos with complex content, which could be time-consuming and expensive, especially if it requires domain expertise for annotation.

Active learning provides a promising solution to reduce the cost of data annotation and model training. It allows the model to learn from a small set of labeled data and then iteratively select the most informative samples to query for annotation (Chan et al. 2020; Das et al. 2022). Conventional active learning methods usually leverage uncertainty-based acquisition functions, where the uncertainty of a video caption is typically quantified through joint likelihood estimation or entropy evaluation from each word in a predicted caption and aggregated through summation or averaging (Ren et al. 2021; Perlitz et al. 2023). However, conventional uncertainty estimation (*e.g.,* entropy or variance) could be unreliable for video captioning due to two reasons. First, unlike standard learning tasks, such as classification and regression, video captioning involves a generative process that produces a sequence of tokens from a large vocabulary. How to objectively quantify the overall uncertainty of the generated content poses a unique challenge. Second, videos contain rich content, which could be described in multiple ways depending on the major concepts being covered by the caption. A single uncertainty score obtained from one possible caption does not capture the diversity induced by the rich content.

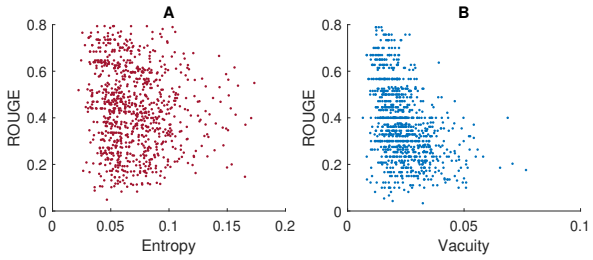

Figure 1: Uncertainty estimation and quality of predicted caption measured by ROUGE score. Entropy and ROUGE may not reveal strong correlation (A). In contrast, a high vacuity typically corresponds to low-quality prediction (B).

To demonstrate those limitations, we performed an analysis to highlight that entropy-based uncertainty estimation may be inflated for video captioning tasks. Specifically, we train a transformer-based captioning model on the MSR-VTT dataset with a 50/50 train-test split and use the trained model for prediction on the test set. For each predicted caption, we plot the mean entropy and the ROUGE score (a widely-used metric for evaluation of captioning) of the prediction against most similar ground-truth caption. A high ROUGE score indicates that the prediction is similar to the ground truth. In

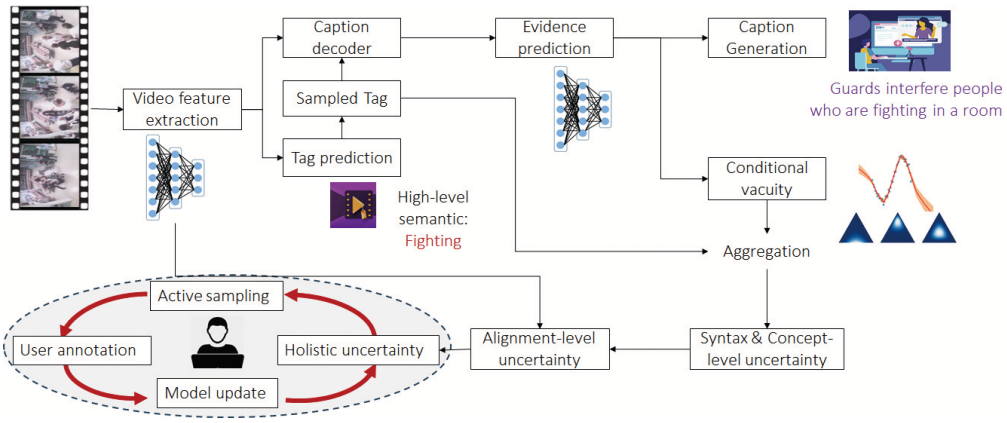

Figure 2: Hierarchical multi-source uncertainty aggregation for interactive video captioning

Caption	Vacuity	Entropy
A road at night.	0.01	0.07
A man is walking on the sidewalk.	0.02	0.13
A man in hoods is walking on the street.	0.04	0.10

Figure 3: Illustrative example of estimated entropy and vacuity on high-quality caption prediction. The predicted captions matche the video contents. Estimated entropy has inflation issue, while vacuity (proposed approach) is close to zero.

Figure 1(left), we observe a significant portion of predictions with high ROUGE scores and high entropy. It indicates the model performs relatively well on those data samples, but entropy-based methods may still select them for annotation and model training. In Figure 3, we show examples where the model generates high-quality captions that match the ground truth. However, the estimated entropy is still not close to 0, indicating that entropy-based uncertainty estimation may be inflated and ineffective for active video captioning.

Some other active learning approaches propose to sample diverse captions to enhance uncertainty estimation by leveraging ensemble strategies and Monte-Carlo dropout to generate a diverse set of captions (Chan et al. 2020), or integrating clustering mechanism to capture the diversity of predicted caption (Das et al. 2022). However, these approaches are still primarily built upon entropy-based uncertainty estimation. In addition, a model generating diverse predictions does not necessarily indicate the model is uncertain about the video. Consequently, those approaches fall short of providing a holistic view of uncertainty that captures the richness of contents in each video.

Given the limitations as outlined, a more principled uncertainty estimation methodology needs to be developed and integrated for effective active video captioning. In this paper, we make extensions of evidential learning (Sensoy, Kaplan, and Kandemir 2018; Amini et al. 2020) to captioning tasks. In the evidential learning framework, vacuity is considered a second-order uncertainty that captures the model's lack of knowledge on a data instance (Shi et al. 2020). We extend vacuity to captioning tasks by defining *conditional vacuity*, which can be evaluated at the word level and then aggregated at the caption level. Unlike entropy-based uncertainty estimation, conditional vacuity does not suffer from inflated estimation, which is demonstrated in our theoretical analysis.

To further capture the rich contents of videos and integrate them into uncertainty estimation, we identify multiple sources of uncertainty during video caption generation and systematically aggregate them to offer a *holistic view* of the overall video uncertainty. Given the inherent complexity of videos, multiple types of uncertainty may be encountered and propagated across different levels of information sources, including syntax (at the *word level*), concept (at the *tag level*), and alignment (at the *semantic level*). At the syntax level, the model needs to generate readable text with appropriate syntax. At the concept level, the model should focus on limited entities in order to condense the rich contents into short descriptions. At the alignment level, the predicted caption should be semantically aligned with the video features. All levels may introduce uncertainty during caption generation. To this end, we propose a hierarchical caption generation and uncertainty quantification framework, encouraging the model to generate diverse captions and aggregate uncertainty with different granularities into a holistic uncertainty metric to support sampling of the most informative videos for annotation.

As shown in Figure 2, hierarchical caption generation includes 1) predicting tags that capture the main concept of the video, 2) generating captions word-by-word conditioned on the video and the tags. We sample the tags based on the visual features of video frames. Since the candidate tags for describing a video can be diverse, caption generation conditioned on sampled tags is encouraged to be more diverse. An illustrative example is provided in Figure 4, where the generated captions conditioned on three sampled tags focus on different aspects of the video. We then leverage a caption generation module by using both video features and tags as input to predict the next word in the caption. The uncertainty is estimated along with the prediction, where the syntax-level uncertainty is quantified by using conditional vacuity to estimate uncertainty per word and aggregate over sentences. The concept-level uncertainty is quantified by further aggregating over tags. Once a caption is generated, the alignment-level

Tag	Probability	Caption	Concept Uncertainty	Alignment Uncertainty	Holistic Uncertainty
market	0.18	A man is robbing a mini market	0.05	0.19	
stealing	0.23	People with masks are stealing a shop	0.09	0.22	0.259
desk	0.11	A man stands inside the cash desk of a store	0.02	0.14	

Figure 4: Illustrative example of the proposed uncertainty aggregation: Holistic uncertainty is estimated by sampling tags and captions and integrating alignment uncertainty between visual and text features

uncertainty can be evaluated via the distance between the caption and visual features in the embedding space. The holistic uncertainty aggregated over the multi-source uncertainty can be used to select a subset of videos for annotation. Intuitively, a high holistic uncertainty indicates the model is not well-calibrated and needs additional training on the corresponding data samples.

In summary, our main contributions include:

- We formally analyze the problem of inflated uncertainty estimation for entropy-based active learning and propose conditional vacuity as an alternative uncertainty quantification for captioning tasks.
- We identify the multiple sources of uncertainty during caption generation and propose a hierarchical aggregation approach to quantify holistic uncertainty and select informative videos for annotation.
- We provide a theoretical justification to demonstrate the advantage of vacuity-based uncertainty over entropy-based uncertainty when handling videos with rich content.

Related Works

We discuss existing works that are most relevant to active video captioning and uncertainty estimation in this section.

Active Video Captioning. In an active learning setting, the model quantify the informativeness of data samples to query for annotations actively. For text generation and captioning, the common approaches include the lowest joint log-likelihood and the highest sequential entropy. In the paper (Chan et al. 2020), researchers further propose a cluster-regularized ensemble strategy that explores various active learning approaches for automatic video captioning, where multiple captions are sampled, and the per-word divergence across different captions are measured for active selection. In (Das et al. 2022), researchers propose a semantics-aware sequential entropy-based method that integrates semantic similarity and uncertainty of both visual and language dimensions. The semantic similarity is measured by the similarity of different sampled captions after BERT encoding, while the multi-dimension uncertainty is estimated by adding perturbation to visual features. In comparison, the proposed method integrates uncertainty from multiple resources and mitigates inflated uncertainty due to multiple correct captions and video visual features.

Uncertainty Quantification. 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 and Ghahramani 2016), Bayes-by Backprop (Blundell et al. 2015), and deep ensembles (Lakshminarayanan, Pritzel, and Blundell 2017) are representative approaches to evaluate uncertainty. In natural language processing domains, (Xiao and 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 and Lipton 2018) leverages uncertainty estimates provided by dropout and Bayes-by Backprop for active learning. (Ott et al. 2019) and (Xu, Desai, and Durrett 2020) investigate prediction entropy for uncertainty estimation on neural language generation tasks. (Xiao and Wang 2021) quantifies epistemic and aleatoric uncertainty in natural language generation tasks to address the hallucination issues. In summary, most existing works leverage dropout and require stochastic sampling. In contrast, our evidence-based uncertainty estimation predicts model uncertainty in a deterministic way, which reduces computational cost and randomness.

Preliminaries

The evidential learning framework is leveraged for uncertainty estimation in this work. Evidential learning is a generalization of Bayesian theory. In classification setting with K mutually exclusive classes, it assigns a belief mass b_k to each possible class k for a data sample and introduces an overall uncertainty mass u. The belief mass values and uncertainty mass are normalized,

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

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 $\mathrm{Dir}(p|\mathbf{a})$ 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.

Methodology

We propose a multi-source uncertainty aggregation framework to quantify the holistic uncertainty in video captioning

tasks and facilitate interactive learning. As shown in Figure 2, we first introduce the caption generation and uncertainty estimation process, and then provide our theoretical analysis. After that, we discuss the uncertainty aggregation.

Uncertainty-Aware Caption Generation

The input video is represented as a sequence of frames $[V_1^{raw}, V_2^{raw}, ...V_T^{raw}]$. Based on the visual features of the video frames, the model first predicts tags that capture the overall video contents. Denote the tag as z, then the tag prediction is formulated as a multi-label classification problem as $z = g(V^{raw})$. With predicted tags, the interactive captioning model takes those tags and visual features as inputs and generates a complete caption $\hat{Y} = f(V^{raw}, z)$. Using such a hierarchical prediction framework, we sample diverse captions and get a more holistic uncertainty estimation that helps select the informative videos for annotation. Given a pre-trained model and a batch of unseen videos, we leverage the uncertainty estimation framework to evaluate the holistic uncertainty, and select the most uncertain videos to query the user for additional annotation. Once the annotations are collected, the model is updated to improve its predictive performance.

We consider the encoder-decoder architecture for video captioning models with an encoder to extract per-frame features and a decoder to generate textual descriptions. For the encoder, we leverage a pre-trained network as the visual feature extractor (Alayrac et al. 2022) to generate feature vectors V_t for each frame indexed by t. To incorporate temporal information, we also augment the visual features with temporal embedding τ_t. For the prediction of tags, we use the resampling mechanism similar to Perceiver (Alayrac et al. 2022) to generate a fixed number of visual tokens. The tokens are then passed through a fully connected classification layer $h(\cdot)$ to predict whether each candidate tag is relevant to the image or not. Each tag corresponds to a binary classification task.

$$z = h(\texttt{Resampler}((V_t + \tau_t)|_{t=1:T}) \tag{4}$$

For caption prediction, we leverage pre-trained transformer-based language models as backbone (Radford et al. 2019). To leverage the visual and tag information, we embed the tag and concatenate it with the visual tokens as soft prompts, which can be fed into the language model. In this way, we leverage the pre-trained language models and encourage the model to consider the visual and tag information when generating captions. Conditioned on tag z and visual features, text token $x_{1:i}$ is passed through the backbone to predict the output vector for the next token

$$\mathbf{o}_i = \texttt{Backbone}(x_{1:i}, [z, \texttt{Resampler}((V_t + \tau_t)|_{t=1:T})]) \tag{5}$$

Finally, the next word is predicted along with uncertainty estimation as discussed in the next section.

For classification tasks, the softmax operation is widely used to convert the continuous activations of the output layer to class probabilities. However, the softmax-based prediction has a natural interpretation of Maximum Likelihood Estimation, which does not capture the predictive distribution variance. In addition, softmax is prone to inflating the probability of the predicted class. Evidential learning overcomes the limitations of softmax-based predictions by formulating the classification and uncertainty modeling jointly (Sensoy, Kaplan, and Kandemir 2018; Shi et al. 2020). In particular, given position i and a candidate word indexed by k for classification, we place a Dirichlet prior parameterized by $\boldsymbol{\alpha}_{i,k}$ to model the class probability, where

$$\alpha_{i,k} = e_{i,k} + 1 = [g(\mathbf{o}_i)]_k + 1 \tag{6}$$

with $\mathbf{o}$ being the output of the transformer block and g being the evidence function to keep evidence $\mathbf{e}_{i,k}$ non-negative. We use a feed-forward layer with ReLU activation for g. Given $\alpha_{i,k}$, the predictive probability for the next word can be estimated using the definition of probability, and the uncertainty can be estimated using vacuity. The model training needs to consider the setting of evidential learning. Specifically, we apply the negative log-likelihood (nll) as the loss, which is minimized to learn evidence $\mathbf{e}$ by integrating out the predictive probability p (Sensoy, Kaplan, and Kandemir 2018):

$$\begin{aligned} \text{nll} &= -\sum_i \log\left(\int \prod_{k=1}^{K} p_{i,k}^{y_{i,k}} \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} p_{i,k}^{a_{i,k}-1} dp\right) \\ &= \sum\nolimits_{k=1}^{K} y_{i,k}(\ln S_i - \ln(\alpha_{i,k})) \end{aligned} \tag{7}$$

where y_i is a one-hot label indicating the ground truth of the next word. $B(\cdot)$ is Beta function. S is the total Dirichlet strength of $\text{Dir}(p|\boldsymbol{\alpha})$, which is parameterized by $\boldsymbol{\alpha} \in \mathbb{R}^K$. S_i is defined as

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

Based on evidential learning, $\alpha_{i,k}$ is determined by the predictive evidence $e_{i,k}$. During the inference, the predicted probability is $\hat{p}_i = \alpha_i / S_i$ and the predictive vacuity can be computed accordingly. Intuitively, we prefer the predicted evidence to shrink to zero if the model is totally unsure about the next word. Note that a Dirichlet distribution with zero evidence, i.e., $S = K$, corresponds to the uniform distribution and indicates total uncertainty, *i.e.,* $u = 1$. We achieve this by incorporating a regularization term into our loss function that regularizes our predictive distribution and penalizes those divergences that do not contribute to data fit. With the regularization term, the loss function becomes

$$L = \text{nll} + \lambda \sum_{i,k} (1 - y_{i,k}) e_{i,k} \tag{9}$$

where the regularization term penalize high evidence $e_{i,k}$ to be assigned to incorrect candidate word (i.e., $y_{i,k} = 0$). λ is a hyper-parameter that controls the regularization.

In summary, we introduce an evidence prediction layer to replace conventional token prediction via probabilistic logits. When trained with Eq (7) and (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. Quantitatively, the uncertainty can be captured by *conditional vacuity*, which corresponds to the normalized uncertainty mass

$$vac_i = u_i = K / S_i \tag{10}$$

Intuitively, vacuity captures a model's lack of knowledge of its prediction. A high vacuity corresponds to the model predicting weak evidence for all candidate words in the vocabulary, and it usually happens on data samples far from training data.

Theoretical Analysis

We further illustrate why the proposed conditional vacuity provides a more robust uncertainty estimation. Caption generation is usually formulated as a sequential prediction problem where the next word is generated given the previous words. For simplification, we consider the task of predicting the next word as a multi-class classification problem. Typically, there are multiple correct captions to describe a video. Given an incomplete sentence fragment, the next word may have multiple appropriate candidates. Intuitively, a model should be confident with training samples. However, when the next word corresponds to multiple appropriate candidates, the conventional entropy-based uncertainty estimation is inflated, while the proposed vacuity-based uncertainty estimation is still small. Formally, we introduce the following lemma.

Lemma 1. *Assume there exist multiple appropriate candidates as the next word for caption generation, and a sufficiently large model is trained with such samples. Denote the predicted probability for word k as $\hat{p}_k$, and the predicted evidence for word k as $\hat{e}_k$, the entropy-based uncertainty estimation on training sample is inflated as*

$$H[\hat{\mathbf{p}}] = -\sum_k \hat{p}_k \ln \hat{p}_k > 0 \tag{11}$$

while the vacuity-based uncertainty estimation is not inflated

$$vac = \frac{K}{K + \sum_{k \in A} \hat{e}_k} \to 0 \tag{12}$$

Detailed proof is provided in the Appendix.

Hierarchical Multi-Source Uncertainty Aggregation

The above approach only evaluates the vacuity per word for uncertainty estimation. To select videos to query annotation, we need to aggregate the uncertainty by multiple sources (*e.g.,* syntax, content, alignment). Denote the maximum length of the generated caption as T. For an instantiated caption $[w_1, w_2, ...w_I]$ where w_i is the predicted word at position n, we consider the syntax uncertainty as the sum of *conditional vacuity*

$$u_{\text{syntax}} = vac(w_1|V, z) + \sum_{i>1} vac\left(w_i|V, z, w_{1:i-1}\right) \tag{13}$$

The predicted caption is unlikely to have exactly I words. In practice, a special token 'eos' is introduced to indicate the end of the sentence. The caption generation is stopped after 'eos'. However, summing up per-word vacuity until 'eos' token may introduce bias to favor short sentences. In other words, short sentences tend to have lower overall vacuity because per-word vacuity ranges in $(0, 1)$, and any additional words in a sentence will incur a non-zero uncertainty contribution. To mitigate this problem, we leverage the concept of evidence and introduce a flexible solution: any word after 'eos' token is predicted as a special 'empty' token with evidence m. Although this prediction has a 100% probability of generating the 'empty' token, it introduces a non-zero vacuity $1/(1+m)$ where m is a hyperparameter. For short sentences, the remaining 'empty' token still incurs a non-zero uncertainty, and it helps offset the bias towards short sentences. With this additional rule, the syntax-level uncertainty with variable sentence length can be expressed in a universal form as Eq (13).

A video may be described in multiple ways, and therefore, relying only on a single sentence during decoding may incur unreliable uncertainty estimation. To address this issue, we consider concept-level uncertainty as the expectation of per-word vacuity aggregated by sentences and tags.

$$\begin{aligned}\mathbb{E}[u_{\text{concept}}] = \sum_z p(z)[\sum_{w_1} p(w_1) vac(w_1|V, z) \\ + \sum_{w_{2:i}} p(w_{2:i}) vac(w_i|V, z, w_{1:i-1})]\end{aligned} \tag{14}$$

Since the candidate tags are diverse, the generated captions conditioned on tags are of high diversity and help capture holistic uncertainty.

To evaluate alignment uncertainty, we can leverage multimodal models such as CLIP (Radford et al. 2021) to generate representations of video frames, tags, and predicted captions in embedding space. Intuitively, if the video features and the generated captions are not well-aligned, it indicates that the model is poorly calibrated and unfamiliar with the unlabeled test samples. In that case, additional annotation on these samples and the corresponding model tuning will likely boost the performance. Empirically, the alignment uncertainty can be calculated by the sum of negative similarity between prediction and visual features, as well as between prediction and tags.

$$u_{\text{align}} = \min_t [1 - h(w_{1:i}) h(V_t)] + [1 - h(w_{1:i}) h(z)] \tag{15}$$

where $h(.)$ denotes the multimodal encoder. V_t denotes the frame at timestamp t, $w_{1:i})$ denotes the prediction, and z denotes the tag. Overall, the holistic uncertainty is aggregated via

$$u_{\text{holistic}} = \mathbb{E}[u_{\text{concept}}] + u_{\text{align}} \tag{16}$$

which can be used to rank data samples for annotation. It should be noted that it is possible that the framework predicts good-quality captions with high uncertainty. However, the overall chance is low. Empirically, we can apply Monte-Carlo sampling to sample multiple captions and estimate the expectation. The sampling mechanism incurs additional computational overhead. However, batch decoding can be implemented to leverage the available computational resources fully. In batch decoding, one batch corresponds to a video: each (incomplete) caption corresponds to a sample in the batch, and the model predicts the next words for each sample in an auto-regressive way.

Ranking Unseen Videos. Given unseen videos with unknown ground-truth captions, the proposed framework quantifies the uncertainty based on Eq (16) and selects the most uncertain ones to query annotation.

Model	UCFCAP			MSRVTT		
	BLEU-4 ↑	ROUGE ↑	CIDEr ↑	BLEU-4 ↑	ROUGE ↑	CIDEr ↑
JLL (Ren et al. 2021)	0.124	0.327	0.384	0.227	0.453	0.408
SE (Perlitz et al. 2023)	0.140	0.336	0.395	0.235	0.458	0.425
CRE (Chan et al. 2020)	0.137	0.334	0.399	0.231	0.466	0.420
MSSE (Das et al. 2022)	0.143	0.342	0.406	0.239	0.470	0.417
Proposed	0.156	0.359	0.425	0.252	0.493	0.431

Table 1: Comparison on video captioning performance

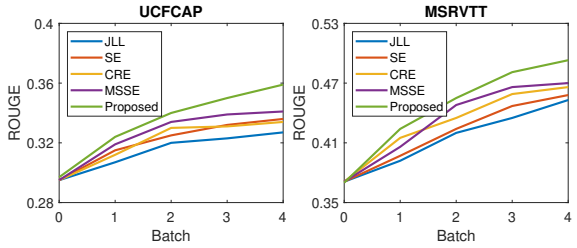

Figure 5: Comparison on active video captioning

We summarize the overall process in Algorithm 1.

Algorithm 1: Training Process

for data sample V in unlabeled set **do**
  Sample tag using Eq (4)
  Calculate evidence and sample word using Eq (6)
  Aggregate uncertainty via Eq (14)
  Estimate alignment uncertainty via Eq (15)
  Calculate holistic uncertainty via Eq (16)
end for
Rank video via holistic uncertainty and query annotation
Model update using Eq (9).

Experiments

We conduct experiments on two captioning datasets. The UCFCAP dataset (Chatzikonstantinou et al. 2022) is an extension of the UCF dataset and contains 2k surveillance videos with descriptions of crimes. The MSR-VTT dataset (Xu et al. 2016) is a diverse dataset for video captioning that consists of over 10k open-domain video clips from 20 categories, and each clip is associated with 20 sentences. The appendix and source code are presented in (Zheng and Yu 2023).

Experimental setup. For experiments, the visual feature extractor is the CLIP model (Radford et al. 2021), and the language model is the GPT2 model (Radford et al. 2019). The backbone model is frozen during the training process, and we only train the adapter layer (Yu et al. 2023), the perceiver layer, and the evidential module. Active learning requires uncertainty quantification of predicted captions to query annotation. To this end, we used ten percent of the videos and corresponding captions from each dataset to pretrain those modules. In addition, the tags are nouns extracted from ground-truth captions for model pretraining. After that, the learning process is performed in four rounds, each related to five percent of the data. During one round of the interaction, the model selects five percent of the videos with the highest holistic uncertainty score to collect annotation (i.e., making the ground-truth captions available to the model). Once a batch of data is processed, the model is further trained based on the annotated videos, and evaluated on its performance on the hold-out test set, which includes the remaining data from the dataset. We evaluate the quality of generated captions based on standard metrics (BLEU-4, ROUGE, and CIDEr scores) (Hossain et al. 2019). Generally, 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. CIDEr also considers synonyms and word order when matching words and phrases. Hyperparameter λ is set to 0.2. We use stochastic gradient descent and Adam optimizer with a learning rate set to 0.0001.

Comparison baselines. We compare with relevant video captioning baselines that actively select videos for annotation. It should be noted that uncertainty estimation for video captioning is a relatively under-explored area, and the relevant baselines are limited, as summarized below. Sample selection based on lowest joint log-likelihood (JLL) and highest sequential entropy (SE) are basic methods that extend multi-class classification to sequential predictions of words in captions (Ren et al. 2021; Perlitz et al. 2023). Additionally, CRE (Chan et al. 2020) proposes an ensemble strategy where the per-word divergence across different captions is measured for the active selection of videos. MSSE (Das et al. 2022) proposes to integrate a clustering mechanism to capture the diversity of predicted captions across clusters.

Active Video Captioning Performance Comparison

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 videos, and the proposed framework outperforms other baselines. 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 Table 1. The difference in uncertainty estimation may explain the results. The proposed method quantifies the overall uncertainty of each video via the vacuity of a bunch of candidate captions. In contrast, the joint log-likelihood method and sequential entropy method may suffer from inflated uncertainty estimation, which accrues errors in overall uncertainty estimation. CRE (Chan et al. 2020) focuses on the divergence of captions, but such divergence does not necessarily indicate model uncertainty because a complex video

Alternative	UCFCAP			MSRVTT		
	BLEU-4 ↑	ROUGE ↑	CIDEr ↑	BLEU-4 ↑	ROUGE ↑	CIDEr ↑
No syntax	0.152	0.354	0.418	0.237	0.462	0.418
No concept	0.148	0.349	0.411	0.234	0.454	0.404
No alignment	0.143	0.340	0.402	0.229	0.451	0.412
Proposed	0.156	0.359	0.425	0.252	0.493	0.431

Table 2: Comparison on alternative aggregation approaches

Approach	UCFCAP			MSRVTT		
	BLEU-4 ↑	ROUGE ↑	CIDEr ↑	BLEU-4 ↑	ROUGE ↑	CIDEr ↑
Greedy	0.144	0.350	0.411	0.241	0.464	0.417
Beam	0.151	0.352	0.416	0.249	0.485	0.422
Proposed	0.156	0.359	0.425	0.252	0.493	0.431

Table 3: Comparison on decoding strategies

Tag	Probability	Caption	Concept Uncertainty	Alignment Uncertainty	Holistic Uncertainty
car	0.37	A car is on fire on sideways	0.02	0.11	
explosion	0.23	People are helping others after an explosion	0.06	0.18	0.173
parking	0.16	A car in the parking lot is on fire	0.04	0.13	
vehicle	0.26	A group of people are walking while two vehicles crashes close to them	0.11	0.17	
people	0.14	People are walking on the sidewalk	0.04	0.19	0.282
motorcycle	0.31	A man is riding a motorcycle then a car crashes another car	0.08	0.21	

Figure 6: Illustrative example of uncertainty estimation in video captioning

may be described in multiple ways. MSSE (Das et al. 2022) focuses on diversity and alignment, but the confidence in text generation may be ignored.

We also provide illustrative examples in Figure 6 to explain different uncertainty components in the proposed framework. In the first example, the model generates reasonably good captions, indicating that the model is familiar with this video. The proposed framework quantifies the uncertainty from multiple sources and uses Monte-Carlo sampling to integrate multiple candidate captions. In the second example, possibly due to the limited training data, the model is unfamiliar with the video, and some of the predicted captions have high uncertainty, which contributes to the high holistic uncertainty after uncertainty aggregation over multiple possible candidates.

Ablation Study Our model leverages evidence as the foundation of uncertainty estimation, and we introduce the MC sampling method to estimate the holistic uncertainty and encourage diverse caption. We conduct an ablation study to compare with alternative settings, including the decoding strategies for caption generation, and varied evidence-based temperatures, to evaluate the contribution of the proposed selection method. Quantitative results are reported in Table 3. A greedy approach for caption generation always chooses the next word with the highest predicted likelihood and estimates the uncertainty accordingly. It only generates a single caption given the input video, and it achieves the fastest decoding speed. However, it may incur errors in overall uncertainty estimation, especially for complex videos. In beam search, we keep the top candidates with the highest cumulative probability up to a threshold number (the beam size is set to 3), but the joint probability may be dominated by a few words. In addition, we evaluate the contribution of using uncertainty aggregation for active learning. Alternative approaches include using negative likelihood rather than conditional vacuity (no syntax uncertainty), ignoring aggregation over tags (no concept uncertainty), and ignoring the alignment between visual and text embeddings (no alignment uncertainty). The results evaluated on the test set are summarized in the following Table 2, which indicates that the aggregation of multi-source uncertainty contributes to better performance.

Conclusion

In this work, we propose a hierarchical uncertainty aggregation framework for video captioning to provide holistic uncertainty estimation. We identify multiple sources of uncertainty during caption generation at syntax, concept, and alignment levels and integrate conditional vacuity into hierarchical uncertainty aggregation to encourage better uncertainty quantification and aggregation even with limited training data. It facilitates interactive learning and selects the most informative videos to query human annotation. The proposed framework can be potentially applied to specialized domains for interactive learning (e.g., security), where data annotation requires domain knowledge and is challenging to crowdsource at low cost.

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

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Amini, A.; Schwarting, W.; Soleimany, A.; and Rus, D. 2020. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33: 14927–14937.

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

Chan, D. M.; Vijayanarasimhan, S.; Ross, D. A.; and Canny, J. F. 2020. Active learning for video description with cluster-regularized ensemble ranking. In *Proceedings of the Asian Conference on Computer Vision*.

Chatzikonstantinou, C.; Valasidis, G. G.; Stavridis, K.; Malogiannis, G.; Axenopoulos, A.; and Daras, P. 2022. UCF-CAP, Video Captioning in the Wild. In *2022 IEEE International Conference on Image Processing (ICIP)*, 1386–1390. IEEE.

Das, G.; Thomas, X.; Raj, A.; and Gupta, V. 2022. MAViC: Multimodal Active Learning for Video Captioning. *arXiv preprint arXiv:2212.11109*.

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

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

Islam, S.; Dash, A.; Seum, A.; Raj, A. H.; Hossain, T.; and Shah, F. M. 2021. Exploring video captioning techniques: A comprehensive survey on deep learning methods. *SN Computer Science*, 2(2): 1–28.

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

Li, S.; Tao, Z.; Li, K.; and Fu, Y. 2019. Visual to text: Survey of image and video captioning. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 3(4): 297–312.

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

Perlitz, Y.; Gera, A.; Shmueli-Scheuer, M.; Sheinwald, D.; Slonim, N.; and Ein-Dor, L. 2023. Active Learning for Natural Language Generation. *arXiv preprint arXiv:2305.15040*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Ren, P.; Xiao, Y.; Chang, X.; Huang, P.-Y.; Li, Z.; Gupta, B. B.; Chen, X.; and Wang, X. 2021. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9): 1–40.

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

Shi, W.; Zhao, X.; Chen, F.; and Yu, Q. 2020. Multifaceted uncertainty estimation for label-efficient deep learning. *Advances in neural information processing systems*, 33: 17247–17257.

Siddhant, A.; and Lipton, Z. C. 2018. 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*, 2904–2909.

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

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

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

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

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5288–5296.

Yu, Y.; Yang, C.-H. H.; Kolehmainen, J.; Shivakumar, P. G.; Gu, Y.; Ren, S. R. R.; Luo, Q.; Gourav, A.; Chen, I.-F.; Liu, Y.-C.; et al. 2023. Low-rank adaptation of large language model rescoring for parameter-efficient speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 1–8. IEEE.

Zheng, E.; and Yu, Q. 2023. Appendix: Hierarchical Multi-Source Uncertainty Aggregation for Interactive Video Captioning. https://github.com/ritminiglab/HMSUA/.