Incorporating Unlikely Negative Cues for Distinctive Image Captioning

Zhengcong Fei , Junshi Huang*
Meituan
{feizhengcong, huangjunshi}@meituan.com

Abstract

While recent neural image captioning models have shown great promise in terms of automatic metrics, they still struggle with generating generic sentences, which limits their use to only a handful of simple scenarios. On the other hand, negative training has been suggested as an effective way to prevent models from producing frequent yet meaningless sentences. However, when applied to image captioning, this approach may overlook low-frequency but generic and vague sentences, which can be problematic when dealing with diverse and changeable visual scenes. In this paper, we introduce a approach to improve image captioning by integrating *negative* knowledge that focuses on preventing the model from producing undesirable generic descriptions while addressing previous limitations. We accomplish this by training a negative teacher model that generates image-wise generic sentences with retrieval entropy-filtered data. Subsequently, the student model is required to maximize the distance with multi-level negative knowledge transferring for optimal guiding. Empirical results evaluated on MS COCO benchmark confirm that our plug-and-play framework incorporating unlikely negative knowledge leads to significant improvements in both accuracy and diversity, surpassing previous state-of-the-art methods for distinctive image captioning.

1 Introduction

Over the past few years, data-driven image captioning has made remarkable progress [Chen *et al.*, 2015; Xu *et al.*, 2015; Vinyals *et al.*, 2015; Anderson *et al.*, 2018; Huang *et al.*, 2019; Cornia *et al.*, 2020; Fei, 2021a; Fei, 2021b; Yan *et al.*, 2021; Zhang *et al.*, 2021; Fang *et al.*, 2022], attracting growing interest from both academic and industrial communities. Traditionally, given a specific input image, neural image captioning models use maximum likelihood estimation (MLE) as a optimization principle to maximize the probability of generating accurate captions that match the ground-truth reference. Unfortunately, the many-to-one phenomenon [Stefanini *et al.*, 2021; Vijayakumar *et al.*, 2016; Fei *et al.*, 2022a], where it is common for an image context to have multiple accurate captions with distinct features, poses a challenge for neural image captioning models. That is, models tend to generate safe but generic captions, which presents an obstacle to the widespread deployment of image captioning systems. To alleviate this issue, some researchers integrate sampling operation in the latent space to meet the requirement of diversity, under the conventional variational autoencoder framework [Wang *et al.*, 2017; Aneja *et al.*, 2019; Mahajan *et al.*, 2019; Mahajan and Roth, 2020; Shen, 2022]. Besides, [Vijayakumar *et al.*, 2016; Holtzman *et al.*, 2019] proposed advanced decoding strategies to alleviate the problem of generic sentences. Indeed, all of the aforementioned methods enhance the diversity of image captions by providing the model with positive guidance on what to generate.

Taking inspiration from negative training techniques [Kim *et al.*, 2019; Ma *et al.*, 2021] that aim to update parameters with unlikely objectives while identifying high-frequency sentences [He and Glass, 2019; Li *et al.*, 2019; Welleck *et al.*, 2019], we argue that it is equally imperative to *instruct image captioning models on what not to generate*. Whilst negative training-based methods have proven effective at promoting sentence diversity, directly applying these techniques to image captioning comes with its own set of drawbacks. Firstly, these methods only consider high-frequency tokens or sentences as negative candidates. However, the high-frequency situation is only a sub-situation. Due to the diversity of visual scenes, descriptions that are low-frequency, yet generic and meaningless may escape punishment, resulting in a failure to address in multi-modal applications. Moreover, we have previously observed that certain generic sub-sentences which evade identification as negative candidates, ultimately compromise the fluency of captions. Secondly, these negative training methods overlook the implicit knowledge within neural networks that identifies negative candidates at multiple levels [Cornia *et al.*, 2020]. We contend that it is more feasible to conduct negative training with richer information, *e.g.*, hierarchical semantic representation.

To address the aforementioned issues and enhance description diversity, we introduce a new Unlikely Negative Knowledge Training training paradigm, referred to as UNKT,

*The corresponding author.

for distinctive image captioning. Inspired by that conventional knowledge distillation [Hinton *et al.*, 2015], where the teacher is typically regarded as a positive role that the student seeks to emulate, instead, we train the teacher model to serve as a negative role model and instruct the student to avoid exhibiting these undesirable behaviors. Specifically, to collect a negative training set, we first employed an entropy filtering strategy that prioritizes retrieving as many many-to-one cases from the raw dataset as possible. It should be noted that usually "one" is a generic sentence and "many" are different images. Next, we trained a standard image captioning model using the aforementioned subset as the negative teacher. The negative teacher generates a set of negative candidates in response to a given image. These negative candidates are captions that seem plausible, yet dull and generic. The purpose of incorporation image-level negative candidates is to encourage the student model to generate more creative and diverse captions by avoiding behaviors that the negative teacher provides, as a result of addressing the previously mentioned drawback. Moreover, to achieve a more all-encompassing training update, we devise two negative objectives, including soft unlikelihood loss on the prediction layer and reverse KL divergence on the intermediate layer. By leveraging multi-level negative knowledge, UNKT effectively encourages the production of more descriptive and contextually vivid captions for student model.

We experimented with our proposed method of unlikely negative knowledge training, evaluating its effectiveness by implementing it into the widely-used Transformer [Vaswani *et al.*, 2017; Cornia *et al.*, 2020] on the MS COCO benchmark. We observe that the model tends to yield clear language patterns in the generated captions, demonstrating that UNKT has successfully learned negative knowledge with explicit supervision. More encouragingly, image captioning models equipped with UNKT perform surprisingly well under diversity evaluation and oracle performance evaluation, achieving new state-of-the-art results. This shows that our learned knowledge is not only distinct but also effectively covers rich semantic understanding. The contributions are summarized as follows:

- We propose a novel unlikely negative knowledge training paradigm for distinctive image captioning. It constructs image-wise generic sentences as the negative candidates to remind the model not to generate. As far as we are concerned, it is the first work to consider unlikely knowledge in diversified image captions;
- With the negative teacher trained with retrieval entropy data, we devise two negative training objectives to progressively calibrate multi-level semantic information to boost the performance of unlikely negative learning;
- We perform extensive experiments and analysis on the MS COCO dataset to verify the effectiveness of the UNKT framework as well as the superiority compared with previous diverse image captioning methods.

2 Background

Image captioning. The objective of the image captioning model is to infer a conditional probability distribution $p(Y|X)$ given a matched image X and its corresponding textual description $Y = \{y_1, \ldots, y_L\}$ from dataset, where the length of the text is denoted by L. Optimizing the parameters of an image captioning model is typically achieved through maximum likelihood estimation (MLE) technique, which can also be expressed as minimizing the factorized negative log-likelihood as:

$$L_{mle} = -\sum_{i=1}^{L} \log p_\theta(y_i|Y_{<i}, X), \tag{1}$$

where $Y_{<i} = \{y_1, \ldots, y_{i-1}\}$ and θ is the model parameters. It is common for multiple different descriptions to be generated for a single visual scene in image captioning task. This "many-to-one" phenomenon occurs frequently in datasets, such as the MS COCO dataset, where each image has five distinct sentences associated with it. However, when using MLE-based training, this phenomenon can lead to the production of overly simplistic and generic sentences by the optimized model.

Unlikelihood training. Unlikelihood training [Welleck *et al.*, 2019] is a proposed solution for neural generation models that struggle with repetitive or overly common tokens in their output. This approach can help to mitigate the problem of undesirable behaviors in generation tasks. By minimizing the likelihood of generating negative candidates during training, the model is encouraged to generate more diverse and interesting content, as:

$$L_{ul} = -\sum_{i=1}^{L} \sum_{y^- \in V_n} \log\left(1 - p_\theta(y^-|Y_{<i}, X)\right), \tag{2}$$

where V_n consists of negative candidates, *e.g.*, overused negative frequent words, that are a subset of the vocabulary.

3 Methodology

Improving the overall quality of image captioning requires guiding the model to identify generic sentences. However, traditional negative training methods that rely on frequency-based analysis have certain limitations. Here, we propose a new negative knowledge transfer that discourages generic sentence generation while promoting distinctive image captioning. Specifically, the approach involves a negative teacher, which helps to equip the student model with enough negative knowledge to prevent undesirable behaviors and ensure smooth information flow.

3.1 Unlikely Negative Teacher Construction

To facilitate negative knowledge transfer, an expert negative teacher must be able to generate precise yet versatile descriptions for a given image. Our approach employs the Transformer [Vaswani *et al.*, 2017] as the foundation for both the negative teacher and student networks. We emphasize the *generality* of this approach, which can be applied to other advanced architecture [Cornia *et al.*, 2020], and have demonstrated its efficacy through experiments.

To construct the negative training dataset, we introduce a retrieval entropy filtering strategy that enables us to identify

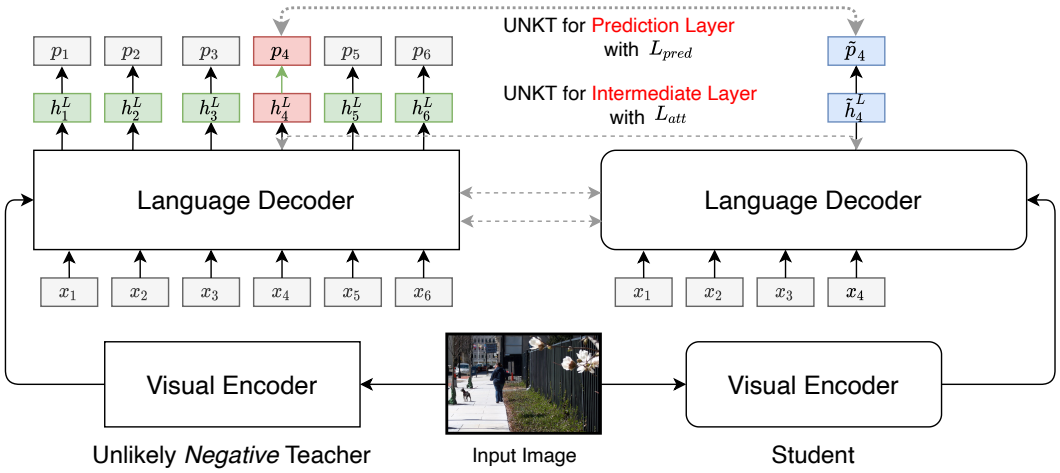

Figure 1: Overview of unlikely negative knowledge transferring for distinctive image captioning, which involves of an unlikely negative teacher model and a student model with an identical network architecture. The negative teacher's output sequence of general examples is used to create a contrast that penalizes the student model, while positive sentences from the dataset are jointly trained with the commonly-used maximum likelihood estimation loss.

and gather instances of many-to-one cases. The retrieval entropy can be defined as:

$$H_{ent}(y, D) = -\sum_{(x_i, y) \in D} p_r(x_i|y) \log p_r(x_i|y), \quad (3)$$

where $p_r(x_i|y)$ is the conditional probability calculated based on the retrieval similarity of image-text pairs from CLIP model [Radford *et al.*, 2021], x_i is the searched image for the query text y, and D represents the raw training set. A higher retrieval entropy indicates that the query text y corresponds to more related images, *i.e.*, the many-to-one problem is more serious. Then, we select the top 50% image-text pairs with a high retrieval entropy and designate them as the negative training set D^-, which contains a considerably larger proportion of generic descriptions than the raw training set. After that, we can train negative teacher p_n on the negative training set D^- with MLE loss in Equation 1, which will naturally produce generic sentences for any input image.

3.2 Unlikely Negative Knowledge Transferring

This section presents the Unlikely Negative Knowledge Transferring (UNKT) framework, which involves transferring knowledge between negative teacher and student models. This transfer of knowledge is achieved through the use of a multi-level approach, as illustrated in Figure 1.

UNKT for prediction layer. It is believed that the softened logits in the prediction layer contain more information than the hard ground-truth labels, such as the similarity between labels [Hinton *et al.*, 2015]. Therefore, conventional KD transfers knowledge by narrowing the gap between the output probability distributions of the teacher model p_t and the student model p_s as:

$$L_{kd} = -\sum_{i=1}^{L}\sum_{k=1}^{V} p_t(y_i = k|Y_{<i}, X) \log p_s(y_i = k|Y_{<i}, X). \quad (4)$$

With regards to Unlikely Negative Distillation, the additional knowledge contained in the softened logits generated by the negative teacher p_n reflects how to generate generic sentences based on the input image. Therefore, we instead introduce a soft unlikelihood loss to maximize the distance between the predictions of the negative teacher p_n and the student p_s as:

$$L_{pred} = -\sum_{i=1}^{L}\sum_{k=1}^{V} p_n(y_i = k|Y_{<i}, X) \log\left(1 - p_s(y_i = k|Y_{<i}, X)\right), \quad (5)$$

where model distribution p_n and p_s can be computed as $p^i = \text{softmax}(\frac{h_i^L}{t})$, where h_i^L is the i-th hidden state from the last L layer and t is a temperature coefficient. It is important to note that previous negative training methods have only used high-frequency words with one-hot representation as targets, which disregards the rich semantic information contained in the softened logits. For instance, generic words often have similar probabilities.

UNKT for intermediate layer. Apart from the knowledge produced by the prediction layer, there is also implicit knowledge contained within intermediate layers, such as hidden states and attention matrices within the network architecture. To enhance the efficacy of deterring unwanted behaviors, such as the generation of generic sentences, in the student model, we further consider the above knowledge into unlikely negative knowledge transferring. Specifically, the distance between features of the negative teacher and student models should also be increased. In this work, we propose a new measurement function, called mean Reverse KL divergence (RKL), to calculate the information distance as:

$$L_{rkl}(P, Q) = \frac{1}{M}\sum_{i=1}^{M} \exp^{-KL(P_i, Q_i)}, \quad (6)$$

Algorithm 1 Unlikely Negative Distillation algorithm

Input: Training dataset D, text-image retrieval model p_r, negative teacher model p_n and student model p_s
Output: Student model p_s

▷ Collection of the negative training set
Compute retrieval entropy $[H_{ent}(y_i, D)]$;
Select top-50% $[H_{ent}(y_i, D)]$ as D^-;
▷ Training of negative teacher
while convergence **do**
    Optimize teacher p_n with L_{mle} on D^-;
end while
▷ Negative distillation training
while convergence **do**
    Optimize student p_s by minimizing L on D ;
end while
return trained diversified student p_s;

where P and Q are the corresponding feature matrices of the negative teacher and the student models, respectively, and M is the number of elements. $KL(\cdot)$ is a standard KL divergence computation function. Considering that semantic modeling mainly focuses on language decoder as well as operation simplicity, we only conduct unlikely negative knowledge transferring on the intermediate layers of the language decoder. The negative transferring objective of hidden states in each decoder can be defined as:

$$L^l_{hid} = L_{rkl}(h^l_n, h^l_s), \tag{7}$$

where h^l_n and h^l_s are the output hidden states of the l-th language decode layer from negative teacher model p_n and student model p_s, respectively.

On the other hand, as the attention weights can learn substantial semantic knowledge [Vig and Belinkov, 2019; Kobayashi *et al.*, 2020], it is beneficial for the student model to further conduct UNKT on the attention matrices. Formally, we set Q, K, and V as the matrices of queries, keys, and values, respectively, and d_k is a model dimension for scaling. We choose $a = soft(\frac{QK^T}{\sqrt{d_k}})$ to calculate the distance. Similar to Equation 7, the unlikely negative transferring objective of matrices is formulated as:

$$L^l_{att} = L_{rkl}(a^l_n, a^l_s), \tag{8}$$

where a_n and a^l_s denote the attention matrices of the l-th language decoder layer of negative teacher and student models.

3.3 Progressive Hyper-Parameter Adjustment

The overall training loss, which combines the above multi-level unlikely negative knowledge objectives and the MLE objective, is used to train the student model end-to-end as:

$$L = (1-\lambda)L_{mle} + \lambda(L_{pred} + \sum_l L^l_{hid} + \sum_l L^l_{att}), \tag{9}$$

where λ is a hyper-parameter that balances the importance of supervised learning as well as unlikely negative knowledge transferring. Note that MLE loss is optimized with corresponding positive image-text pairs.

When it comes to unlikely negative knowledge parts, it is preferable for the student model to have the ability to generate captions before being reminded of what not to say. Thus, we adopt a *progressive* optimization approach that initially increases the negative knowledge ratio linearly and then gradually decreases it. The latter stage closely resembles to the cosine learning rate decay and can be determined by:

$$\lambda_{step} = \frac{1}{2}(1 + cos(\frac{step}{T} * \frac{\pi}{2})) \tag{10}$$

where $step$ corresponds to the training step, and T is total training steps without warmup.

4 Experiments

4.1 Experimental Settings

Dataset. We train and evaluate the UNKT method on the MS COCO dataset [Chen *et al.*, 2015] that contains 123,287 images and each image is equipped with at least 5 captions. For a fair comparison, we follow the previous works [Mahajan and Roth, 2020] in the area of diverse and controllable image captioning to use the m-RNN split [Mao *et al.*, 2014] of the COCO dataset, which divides the data into 118,287 for training, 4K and 1K for validation and testing, respectively.

Quality metrics. To assess the quality of the generated captions, we use five widely used evaluation metrics, *i.e.*, BLEU-n [Papineni *et al.*, 2002], METEOR [Lavie and Agarwal, 2007], ROUGE-L [Lin, 2004], CIDE [Vedantam *et al.*, 2015], and SPICE [Anderson *et al.*, 2016]. In between, CIDEr focuses on semantic analysis and has a higher correlation with human judgment, and other metrics favor frequent n-grams and measure the overall fluency.

Diversity metrics. To investigate the diversity of the generated captions, we use SelfCIDEr [Wang and Chan, 2019], mBLEU, and n-gram diversity denoted as Div-n [Li *et al.*, 2015]. All of these metrics evaluate the diversity by comparing the n-gram differences among the generated captions that belong to the same image.

Implementation details. The proposed UNKT is a general training paradigm and we expect it can be easily applied to many existing image captioning models and improve their diversity. In this paper, we choose widely-used and representative Transformer architectures [Vaswani *et al.*, 2017] as our base models, to show the generalization ability of UNKT. Since most current SoTA image captioning models, like M2Transformer [Cornia *et al.*, 2020], COSNet [Li *et al.*, 2022a], ViTCAP [Fang *et al.*, 2022], and many vision-language pre-training models, are based on the Transformer architecture. Both the negative teacher and the student networks hold the same setting in terms of network architecture and hyper-parameters. Specifically, following the settings of Transformer in [Vaswani *et al.*, 2017], both visual encoder and language decoder contain 6 layers, in which the self-attention module has 8 attention heads and the number of feed-forward units is 2046. The size of hidden states is set to 512 and the dimension is 64 for query, key, and value, with a dropout ratio of 0.1. We use AdamW [Kingma and Ba, 2014] optimizer and employ 3000 steps warm-up trick to adjust the

Method	#Samples	BLEU-4	BLEU-3	BLEU-2	BLEU-1	CIDEr	ROUGE	METEOR	SPICE
Div-BS		38.3	53.8	68.7	83.7	140.5	65.3	35.7	26.9
POS		44.9	59.3	73.7	87.4	146.8	67.8	36.5	27.7
AG-CVAE		47.1	57.3	69.8	83.4	125.9	63.8	30.9	24.4
Seq-CVAE	20	44.5	59.1	72.7	87.0	144.8	67.1	35.6	27.9
COS-CVAE		50.0	64.0	77.1	90.3	162.4	70.6	38.7	29.5
Transformer-DML		52.6	66.3	78.8	91.5	170.4	72.6	41.7	32.5
UNKT		54.5	68.5	80.0	92.0	174.2	73.8	41.9	33.0
UNKT *w/* memory		**55.8**	**69.1**	**80.4**	**92.3**	**176.4**	**74.2**	**43.9**	**33.5**
Div-BS		40.2	55.5	69.8	84.6	144.8	66.6	37.2	29.0
POS		55.0	67.2	78.7	90.9	166.1	72.5	40.9	31.1
AG-CVAE		55.7	65.4	76.7	88.3	151.7	69.0	34.5	27.7
Seq-CVAE		57.5	69.1	80.3	92.2	169.5	73.3	41.0	32.0
LNFNN	100	59.6	69.5	80.2	92.0	170.5	72.9	40.2	31.6
COS-CVAE		63.3	73.9	84.2	94.2	189.3	77.0	45.0	33.9
Transformer-DML		64.9	75.0	84.9	94.6	195.3	78.0	47.4	35.4
UNKT		65.6	75.5	85.3	94.9	196.8	78.4	48.1	35.7
UNKT *w/* memory		**66.1**	**75.8**	**85.5**	**95.0**	**197.5**	**78.7**	**48.7**	**35.9**

Table 1: Oracle performance comparisons, *i.e.*, best-1 quality, in the MS COCO dataset. "#Samples" refers to the number of generated captions for each image. The first places for diverse image captioning are marked with the bold font.

Models	Div-1	Div-2	SelfCIDEr	mB.(↓)
LNFMM	0.37	0.50	-	0.64
COS-CVAE	0.39	0.57	0.79	0.53
Seq-CVAE	0.33	0.48	-	0.64
Transformer-BS	0.21	0.29	0.57	0.78
Transformer-DML	0.43	0.59	0.83	0.54
UNKT	**0.46**	**0.63**	**0.85**	**0.50**
UNKT w/ memory	0.45	0.60	0.83	0.53

Table 2: Diversity evaluation on the best-5 sentences obtained from consensus re-ranking in the MS COCO dataset.

learning rate of 3e-4 during training. We linearly increase λ to 1 in first 5 epochs and then gradually decay after. For the temperature coefficient t, we simply set it to 1.

4.2 Evaluation on Quality

To evaluate the effectiveness of unlikely negative knowledge in generating accurate descriptions, we first calculate the quality evaluation metrics in the oracle setting consistent with prior works [Mahajan and Roth, 2020; Mahajan *et al.*, 2020], *i.e.*, taking the maximum score for each automatic metric over all the candidate captions for each image. Specifically, we infer from the Transformer-based UNKT model with sampling sizes of 20 and 100, and evaluate the oracle results of the generated captions. The results are listed in Table 1. As we can see that the standard Transformer equipped with UNKT obtain a significant improvement compared with the results of the best competitors on all the quality metrics w.r.t. both 20 and 100 samples. More encouraging, when employed with the advanced meshed memory structure in [Cornia *et al.*, 2020], the captioning performance obtains further gains. It demonstrates that UNKT method is generalized and can be easy to transfer to other advanced models.

Moreover, we compare our UNKT paradigm with the previous SoTA baselines Transformer-DML [Chen *et al.*, 2022] and COS-CVAE [Mahajan and Roth, 2020] by calculating the oracle scores of CIDEr with different numbers of samples. The evaluated results are shown in Figure 2, and we can observe that the proposed UNKT model consistently outperforms baselines in all settings. The high gain in quality metrics demonstrates that the proposed UNKT framework successfully avoids the generic and meaningless negative productions in the training corpus.

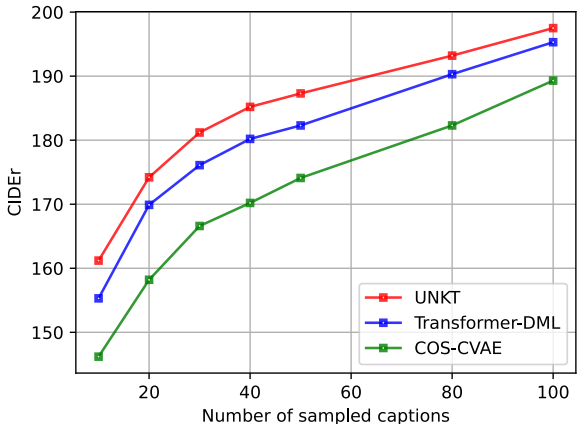

Figure 2: Comparison between our UNKT model and previous SoTA baseline COS-CVAE and Transformer-DML. We show the oracle results of CIDEr score with different numbers of samples.

4.3 Evaluation on Diversity

Quantitative results. In this part, we perform a diversity analysis for the proposed UNKT paradigm based on Transformer and advanced memory strategy. We compare our model with previous SoTA methods as well as Beam Search (BS) and show the results in Table 2. Overall, the proposed

	B-4	M	C	Div-1	SelfCIDEr	mB.
UNKT	**54.5**	**41.9**	**174.2**	**0.46**	**0.85**	**0.50**
w/o L_{pred}	55.3	43.0	175.5	0.45	0.83	0.54
w/o L_{att}	54.9	42.4	174.8	0.43	0.82	0.54
w/o L_{hid}	53.6	40.4	172.5	0.41	0.80	0.57
w/o L_{neg}	44.9	36.4	145.7	0.22	0.58	0.79

Table 3: Ablation studies of different knowledge transferring objectives in unlikely negative training.

	B-4	M	C	Div-1	SelfCIDEr	mB.
p_n^-	39.2	35.3	138.2	0.18	0.55	0.80
p_n^+	**42.3**	**36.5**	**144.2**	**0.32**	**0.63**	**0.64**

Table 4: Effect of retrieval entropy filtering method in unlikely negative training dataset construction.

UNKT model achieves higher performance on most of the diversity evaluation metrics, which demonstrates its effectiveness in expressing distinct semantic content. Besides, we also find that the incorporation of memory will degrade the diverse performance to some degree.

Qualitative results. We further provide several examples of the generated captions of the UNKT model and baseline of standard Transformer with beam search from the MS COCO dataset in Figure 3. Obviously, UNKT model generates more diverse and vivid captions, which further demonstrate the superiority of the improved negative training framework for multi-level information transfer modeling.

4.4 Model Analysis

Ablation study. We conducted an analysis to examine the impact of multi-level unlikely negative knowledge transfer objectives by removing various components: the prediction layer transfer (w/oL_{pred}), the attention transfer (w/oL_{att}), the hidden state transfer (w/oL_{hid}), and the entire negative training (w/oL_{neg}), which corresponds to the standard MLE loss, in the combined loss as shown in Equation 9. The results presented in Table 3 demonstrate that all three knowledge transfer objectives, from distinct perspectives, contribute to enhancing both the diversity and quality of captioning. Notably, the substantial performance decrease observed in the "without L_{hid}" scenario highlights the importance and abundance of negative information in intermediate layers for effective negative information transmission.

Effect of retrieve entropy in negative dataset construction. To evaluate the effectiveness of the retrieval entropy filtering method in gathering generic text from image-text datasets, we divide the sorted training set into two equal parts: the top 50% (D^-) and the bottom 50% (D^+). We then train standard Transformer models, p^- and p^+, on their respective subsets. The results in Table 4 reveal that the p^+ model significantly surpasses the p^- model in all diversity-related metrics, thereby confirming the efficacy of retrieval entropy filtering in data selection.

Furthermore, we utilize p_n^- and p_n^+ as negative teachers for the student models, p_s^- and p_s^+, and perform negative knowledge transfer on both models. The outcomes presented in Table 5 reveal that p_s^- exhibits greater improvements in diversity than p_s^+, suggesting that p_s^+ discards more valuable negative knowledge. This observation is consistent with the previous analysis, which demonstrated that p_t^- possesses more negative knowledge than p_t^+.

	B-4	M	C	Div-1	SelfCIDEr	mB.
p_s^-	**54.5**	**41.9**	**174.2**	**0.46**	**0.85**	**0.50**
p_s^+	54.2	41.5	173.7	0.31	0.63	0.65

Table 5: Effect of unlikely negative knowledge incorporation.

	B-4	M	C	Div-1	SelfCIDEr	mB.
Random	52.5	39.2	168.3	0.38	0.77	0.61
Hard	54.4	41.7	174.0	0.45	0.83	0.52
Soft	**54.5**	**41.9**	**174.2**	**0.46**	**0.85**	**0.50**

Table 6: Comparison of soft targets, hard targets, and random targets for unlikely negative knowledge transferring.

Effect of soft targets in knowledge transferring. To assess the effectiveness of soft targets for negative knowledge transfer, we compare them with hard targets obtained by sampling sentences using greedy search on the predictions of negative teachers, given the image. The results presented in Table 6 indicate that UNKT with soft targets can significantly enhance caption diversity, highlighting the benefits of incorporating abundant unlikely negative information, such as label similarity, in soft targets. Additionally, we randomly select sentences from the negative training set, D^-, as negative targets. The substantial drop in image captioning performance confirms that the negative teacher model can generate high-quality, yet hard, conditional generic captions.

Effect of progressive hyper-parameter adjustment. To evaluate the effectiveness of the progressive hyper-parameter adjustment method presented in Section 3.3, we perform unlikely negative training with a fixed λ value, which is obtained by set λ in Equation 9 to 0.5 across the convergence steps. The outcomes presented in Table 8 indicate that the progressive optimization policy enables the student model to leverage negative knowledge more effectively and comprehensively.

4.5 Human Evaluation

Apart from automatic evaluations, we conducted human evaluations to further validate the effectiveness of the UNKT method compared to previous diverse image captioning techniques. We randomly selected 300 samples from the MS COCO test set and invited three well-educated annotators to judge which of the sentences generated by UNKT and the baselines were better in terms of three aspects: informativeness, relevance, and fluency. Informativeness reflects the amount of information related to the image contained in the generated caption, relevance reflects the coherence of the generated caption with its image, and fluency reflects the likelihood of the generated sentence being produced by humans. The results of the human evaluation are summarized in Table 7. We can see that UNKT approach is overall better than all

Transformer:
a dog and a person are watching television together
a dog and a person are watching TV together
a dog is watching television with a person
a dog is watching television together with person
a dog and a person are watching television

UNKT:
a person is watching television with a dog
a dog between two feet watching television
a dog and a person are looking at TV together
a man is watching TV at home with his dog
a dog sitting with a person are watching television

Transformer:
there are two sinks in the bathroom
there are two sinks in the clean bathroom
there are two sinks with mirrors in the room
there are two sinks with mirrors in the bathroom
two sinks and two mirrors in the bathroom

UNKT:
a bathroom with two sinks and two mirrors
two sinks with mirrors above sinks
a double sink with mirrors in the bath room
there are two sinks with mirrors in the bathroom
two mirrors and lights over the sink in the room

Figure 3: Qualitative comparison of image captions generated by our UNKT framework as well as standard Transformer with beam search in the MS COCO testing set. We can see that UNKT model produces more diverse and vivid descriptions.

	Informativeness			Relevance			Fluency		
vs. UNKT	Win (%)	Tie (%)	Lose (%)	Win (%)	Tie (%)	Lose (%)	Win (%)	Tie (%)	Lose (%)
Transformer	22.0	26.3	**51.7**	30.3	**35.7**	34.0	32.7	**37.7**	29.6
CVAE	29.7	30.7	**39.6**	28.3	29.7	**42.0**	26.0	35.0	**39.0**
Transformer-DML	32.7	29.3	**38.0**	30.7	32.7	**36.6**	29.3	33.7	**37.0**

Table 7: Results of human evaluations on the MS COCO benchmark. The proposed UNKT framework has a higher win rate than baselines.

	B-4	M	C	Div-1	SelfCIDEr	mB.
Fixed λ	54.3	41.6	173.9	0.43	0.82	0.55
Progressive	**54.5**	**41.9**	**174.2**	**0.46**	**0.85**	**0.50**

Table 8: Effect of progressive hyper-parameter adjustment.

baselines. Specifically, UNKT achieved better performance than the standard Transformer in terms of informativeness and remained competitive in fluency and relevance. Compared to both Transformer-DML and CVAE, our approach demonstrated significant advantages, particularly in fluency. These results suggest that incorporating negative knowledge can enhance the distinctive image captioning capacity.

5 Related Works

Diverse Image Captioning. Recent studies have shown that MLE-optimized models tend to exhibit a bias towards safe and average versions that only contain common words and phrases from the training corpus [Mao *et al.*, 2022; Fei, 2019; Fei *et al.*, 2022b]. In contrast, diverse image captioning aims to train a model that can generate a variety of captions for the same image. CVAE-based models [Wang *et al.*, 2017; Aneja *et al.*, 2019; Mahajan *et al.*, 2019; Mahajan and Roth, 2020; Shen, 2022; Chen *et al.*, 2022] learn a latent space during training and then generate diverse captions by sampling different priors from the latent space. GAN-based models [Dai *et al.*, 2017; Zhang *et al.*, 2022] predict diverse captions by using different random noises as inputs accompanied by the given images. Diverse decoding methods [Vijayakumar *et al.*, 2016; Holtzman *et al.*, 2019] incorporate token-level constraints during inference. Although these diverse image captioning models are able to produce different captions during inference, all of them handle the generic response problem only from the angle of negative, thus can not capture all the features of generic sentences.

Unlikely Negative Training. Generally, unlikelihood training moderates MLE by imposing an explicit penalty on the decoding of next target words that are predefined in a candidate set of tokens. This set typically consists of undesirable words and n-grams that contradict or have already appeared in the previous context. Unlikely negative training has been widely explored and shown promising results in text generation [Welleck *et al.*, 2019], language modeling [Son *et al.*, 2022], dialogue [Nugmanova *et al.*, 2019; He and Glass, 2020; Li *et al.*, 2020; Li *et al.*, 2022b], machine translation [Ott *et al.*, 2018], motion forecasting [Zhu *et al.*, 2022], and interpretable embedding [Wu *et al.*, 2022]. However, its influence in improving the diversity of image captioning is under-explored. To this end, our work proposes a unlikely negative training paradigm to cover the visual element pluralistic, avoiding the problem of previous work.

6 Conclusion

In this paper, we introduce an unlikely negative knowledge training paradigm to improve the diversity of image captioning. It formulates negative training as a knowledge transferring process in a multi-level perspective from a trained negative teacher learned with retrieval-entropy filtering data. By generating generic and uninteresting sentences for any given image, the negative teacher method circumvents obstacles that have previously impeded frequency-based approaches. Extensive experiments on the MS COCO dataset validate the superiority of proposed UNKT method compared with previous diverse image captioning works.

References

[Anderson *et al.*, 2016] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Proc. ECCV*, pages 382–398, 2016.

[Anderson *et al.*, 2018] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proc. IEEE CVPR*, pages 6077–6080, 2018.

[Aneja *et al.*, 2019] Jyoti Aneja, Harsh Agrawal, Dhruv Batra, and Alexander Schwing. Sequential latent spaces for modeling the intention during diverse image captioning. In *Proc. IEEE CVPR*, pages 4261–4270, 2019.

[Chen *et al.*, 2015] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[Chen *et al.*, 2022] Qi Chen, Chaorui Deng, and Qi Wu. Learning distinct and representative modes for image captioning. *arXiv preprint arXiv:2209.08231*, 2022.

[Cornia *et al.*, 2020] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proc. IEEE CVPR*, pages 10578–10587, 2020.

[Dai *et al.*, 2017] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional gan. In *Proc. IEEE CVPR*, pages 2970–2979, 2017.

[Fang *et al.*, 2022] Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lin Liang, Zhe Gan, Lijuan Wang, Yezhou Yang, and Zicheng Liu. Injecting semantic concepts into end-to-end image captioning. In *Proc. IEEE CVPR*, pages 18009–18019, 2022.

[Fei *et al.*, 2022a] Zhengcong Fei, Mingyuan Fan, Li Zhu, Junshi Huang, Xiaoming Wei, and Xiaolin Wei. Uncertainty-aware image captioning. *arXiv preprint arXiv:2211.16769*, 2022.

[Fei *et al.*, 2022b] Zhengcong Fei, Xu Yan, Shuhui Wang, and Qi Tian. Deecap: Dynamic early exiting for efficient image captioning. In *Proc. IEEE CVPR*, pages 12216–12226, June 2022.

[Fei, 2019] Zheng-cong Fei. Fast image caption generation with position alignment. *arXiv preprint arXiv:1912.06365*, 2019.

[Fei, 2021a] Zhengcong Fei. Memory-augmented image captioning. In *Proc. AAAI*, volume 35, pages 1317–1324, 2021.

[Fei, 2021b] Zhengcong Fei. Partially non-autoregressive image captioning. In *Proc. AAAI*, volume 35, pages 1309–1316, 2021.

[He and Glass, 2019] Tianxing He and James Glass. Negative training for neural dialogue response generation. *arXiv preprint arXiv:1903.02134*, 2019.

[He and Glass, 2020] Tianxing He and James Glass. Negative training for neural dialogue response generation. In *Proc. ACL*, pages 2044–2058, 2020.

[Hinton *et al.*, 2015] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[Holtzman *et al.*, 2019] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *Proc. ICLR*, 2019.

[Huang *et al.*, 2019] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proc. IEEE ICCV*, pages 4634–4643, 2019.

[Kim *et al.*, 2019] Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. Nlnl: Negative learning for noisy labels. In *Proc.IEEE ICCV*, pages 101–110, 2019.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Kobayashi *et al.*, 2020] Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. *arXiv preprint arXiv:2004.10102*, 2020.

[Lavie and Agarwal, 2007] Alon Lavie and Abhaya Agarwal. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proc. ACL Workshop*, pages 228–231, 2007.

[Li *et al.*, 2015] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015.

[Li *et al.*, 2019] Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. *arXiv preprint arXiv:1911.03860*, 2019.

[Li *et al.*, 2020] Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proc. ACL*, pages 4715–4728, 2020.

[Li *et al.*, 2022a] Yehao Li, Yingwei Pan, Ting Yao, and Tao Mei. Comprehending and ordering semantics for image captioning. In *Proc. IEEE CVPR*, pages 17990–17999, 2022.

[Li *et al.*, 2022b] Yiwei Li, Shaoxiong Feng, Bin Sun, and Kan Li. Diversifying neural dialogue generation via negative distillation. *arXiv preprint arXiv:2205.02795*, 2022.

[Lin, 2004] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Proc. ACL Workshops*, pages 74–81, 2004.

[Ma *et al.*, 2021] Ruotian Ma, Tao Gui, Linyang Li, Qi Zhang, Yaqian Zhou, and Xuanjing Huang. Sent: sentence-level distant relation extraction via negative training. *arXiv preprint arXiv:2106.11566*, 2021.

[Mahajan and Roth, 2020] Shweta Mahajan and Stefan Roth. Diverse image captioning with context-object split latent spaces. *arXiv preprint arXiv:2011.00966*, 2020.

[Mahajan *et al.*, 2019] Shweta Mahajan, Iryna Gurevych, and Stefan Roth. Latent normalizing flows for many-to-many cross-domain mappings. In *Proc. ICLR*, 2019.

[Mahajan *et al.*, 2020] Shweta Mahajan, Iryna Gurevych, and Stefan Roth. Latent normalizing flows for many-to-many cross-domain mappings. *arXiv preprint arXiv:2002.06661*, 2020.

[Mao *et al.*, 2014] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014.

[Mao *et al.*, 2022] Yangjun Mao, Long Chen, Zhihong Jiang, Dong Zhang, Zhimeng Zhang, Jian Shao, and Jun Xiao. Rethinking the reference-based distinctive image captioning. In *Proc. ACM MM*, pages 4374–4384, 2022.

[Nugmanova *et al.*, 2019] Aigul Nugmanova, Andrei Smirnov, Galina Lavrentyeva, and Irina Chernykh. Strategy of the negative sampling for training retrieval-based dialogue systems. In *Proc. IEEE PCCW*, pages 844–848. IEEE, 2019.

[Ott *et al.*, 2018] Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. Analyzing uncertainty in neural machine translation. In *Proc. ICML*, pages 3956–3965. PMLR, 2018.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proc. ACL*, pages 311–318, 2002.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. ICML*, pages 8748–8763. PMLR, 2021.

[Shen, 2022] Xiaoyu Shen. Deep latent-variable models for text generation. *arXiv preprint arXiv:2203.02055*, 2022.

[Son *et al.*, 2022] Seonil Son, Jaeseo Lim, Youwon Jang, Jaeyoung Lee, and Byoung-Tak Zhang. Learning to write with coherence from negative examples. *arXiv preprint arXiv:2209.10922*, 2022.

[Stefanini *et al.*, 2021] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on image captioning. *arXiv preprint arXiv:2107.06912*, 2021.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. NIPS*, pages 5998–6008, 2017.

[Vedantam *et al.*, 2015] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proc. IEEE CVPR*, pages 4566–4575, 2015.

[Vig and Belinkov, 2019] Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284*, 2019.

[Vijayakumar *et al.*, 2016] Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*, 2016.

[Vinyals *et al.*, 2015] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proc. IEEE CVPR*, pages 3156–3164, 2015.

[Wang and Chan, 2019] Qingzhong Wang and Antoni B Chan. Describing like humans: on diversity in image captioning. In *Proc. IEEE CVPR*, pages 4195–4203, 2019.

[Wang *et al.*, 2017] Liwei Wang, Alexander G Schwing, Dahua Lin, and Svetlana Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *Proc. NIPS*, pages 5758–5768, 2017.

[Welleck *et al.*, 2019] Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*, 2019.

[Wu *et al.*, 2022] Jiaxin Wu, Chong-Wah Ngo, Wing-Kwong Chan, and Zhijian Hou. (un) likelihood training for interpretable embedding. *arXiv preprint arXiv:2207.00282*, 2022.

[Xu *et al.*, 2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. ICML*, pages 2048–2057, 2015.

[Yan *et al.*, 2021] Xu Yan, Zhengcong Fei, Zekang Li, Shuhui Wang, Qingming Huang, and Qi Tian. Semi-autoregressive image captioning. In *Proc. ACM MM*, pages 2708–2716, 2021.

[Zhang *et al.*, 2021] Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Rstnet: Captioning with adaptive attention on visual and non-visual words. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15465–15474, 2021.

[Zhang *et al.*, 2022] Wenqiao Zhang, Haochen Shi, Jiannan Guo, Shengyu Zhang, Qingpeng Cai, Juncheng Li, Sihui Luo, and Yueting Zhuang. Magic: Multimodal relational graph adversarial inference for diverse and unpaired text-based image captioning. In *Proc. AAAI*, volume 36, pages 3335–3343, 2022.

[Zhu *et al.*, 2022] Deyao Zhu, Mohamed Zahran, Li Erran Li, and Mohamed Elhoseiny. Motion forecasting with unlikelihood training in continuous space. In *Conference on Robot Learning*, pages 1003–1012. PMLR, 2022.