DeeCap: Dynamic Early Exiting for Efficient Image Captioning

Zhengcong Fei [1,2]*, Xu Yan [1,2], Shuhui Wang [1,3]†, Qi Tian [4]
[1] Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3] Peng Cheng Laboratory, Shenzhen, China [4] Huawei Cloud & AI, China
{feizhengcong, yanxu19s, wangshuhui}@ict.ac.cn, tian.qi1@huawei.com

Abstract

Both accuracy and efficiency are crucial for image captioning in real-world scenarios. Although Transformer-based models have gained significant improved captioning performance, their computational cost is very high. A feasible way to reduce the time complexity is to exit the prediction early in internal decoding layers without passing the entire model. However, it is not straightforward to devise early exiting into image captioning due to the following issues. On one hand, the representation in shallow layers lacks high-level semantic and sufficient cross-modal fusion information for accurate prediction. On the other hand, the exiting decisions made by internal classifiers are unreliable sometimes. To solve these issues, we propose DeeCap framework for efficient image captioning, which dynamically selects proper-sized decoding layers from a global perspective to exit early. The key to successful early exiting lies in the specially designed imitation learning mechanism, which predicts the deep layer activation with shallow layer features. By deliberately merging the imitation learning into the whole image captioning architecture, the imitated deep layer representation can mitigate the loss brought by the missing of actual deep layers when early exiting is undertaken, resulting in significant reduction in calculation cost with small sacrifice of accuracy. Experiments on the MS COCO and Flickr30k datasets demonstrate the DeeCap can achieve competitive performances with 4× speed-up. Code is available at: https://github.com/feizc/DeeCap.

1. Introduction

Image captioning aims to generate a textual description for a given image. It requires not only to identify what visual objects the image contains but also to explain their relationship [4]. Recently, encoder-decoder framework has achieved great progress [2, 6, 8, 14, 22, 50, 55, 56, 60, 61], which generates sentence through modeling the next word conditioned on both the image and generated sub-sentence. Moreover, Transformer architecture [48] is introduced to implicitly relate semantic information through dot-product attention and achieves state-of-the-art performances [10, 28, 33, 36, 37, 62]. Although these models achieve more promising results, the low inference speed hinders their application to many real-time applications [20].

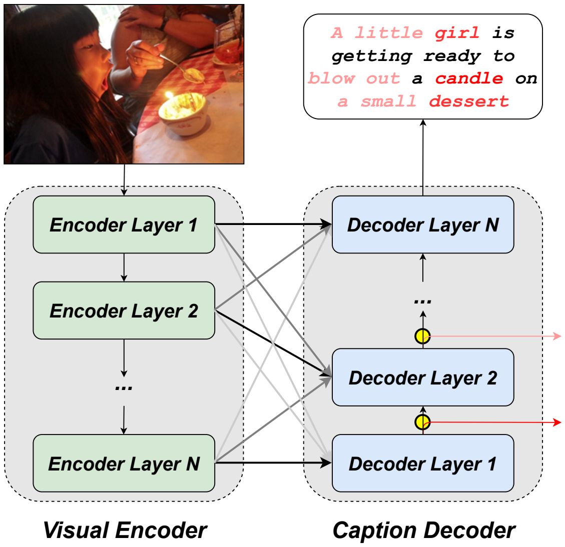


Figure 1. Conceptual workflow of early exiting in image captioning system, which adjusts the number of passed decoder layers, allowing for reduction of computational costs yet facing a performance bottleneck. Yellow circles are internal classification, and red arrows denote tokens exiting early at different decoder layers.

A growing number of studies emerged recently focusing on improving its efficiency. Inspired by the parallelism of Transformer [19], one straight forward strategy is the non-autoregressive decoding [12, 58], which predicts the entire sentence in one shot. Numerous works follow this line, *e.g.*, iterative refinement [13, 16, 23], la-

[*] Intern at Meituan.
[†] The corresponding author.

tent variables [13], multi-agent optimization [20], and semi-autoregressive [15, 57, 64]. Such model-level optimization usually lacks dependencies among words and struggles to produce descriptions with good quality. Another method is proposed for instance-level speed-up, called early exiting [17, 46, 54], which emits output with internal classifiers when the predictions are confident enough. However, its applicability to multimodal context is still largely underexplored. In this paper, we focus on performing early exiting for image captioning, as illustrated in Figure 1.

We first conduct probing experiments to investigate the direct transfer of original early exiting [11, 54] in image captioning, and find that the resulting poor performance lies in: (i) The local shallow representations of caption decoder lack high-level semantic and cross-modal fusion information, it is insufficient to predict accurate tokens. As Transformer-based structure exhibits a hierarchy of representations, *e.g.*, shallow layers extract low-level features like syntactic information while deep layers capture semantic fusion relations [10, 35], we believe that the high-level information usage is usually required even for easy instances. (ii) The internal classifiers in the early exiting cannot provide reliable exiting decisions. In practice, we design an evaluation metric, referred to as false confidence score, to examine the ability and quality of image captioning models to distinguish difficult contexts from easy ones. We discover that the predictions of internal classifiers, *i.e.*, confidence score, cannot truly reflect the difficulty at sometimes, resulting in wrongly generated results and thus hindering the application of early exiting.

Following this premise, we investigate the design of dynamic early exiting method for image captioning and incorporates two key novelties with respect to all previous algorithms: (i) similar to mesh connection [10], all the low-level hidden states are connected for adequate historical information, instead of only one hidden state. (ii) the high-level representation in the uncomputed deep layers is estimated with imitation learning-based [43] lightweight network that only inputs the low-level features. The resulting prediction is also employed for exiting prediction, which in return can compromise the performance degradation brought by the lack of high-level features. By combining both shallow and imitated deep hidden representations, our DeeCap model efficiently generates high-quality sequences with early exiting. Experiments on the MS COCO and Flickr30k benchmarks demonstrate that the proposed dynamic early stopping approach in image captioning can obtain a much better description performance. More importantly, the trained model can be adjusted in real-time without re-training from scratch like previous methods [12, 16, 20]. Further analysis also shows that incorporation of imitated deep layer representation can calibrate the caption generation and proves the effectiveness and generalizability.

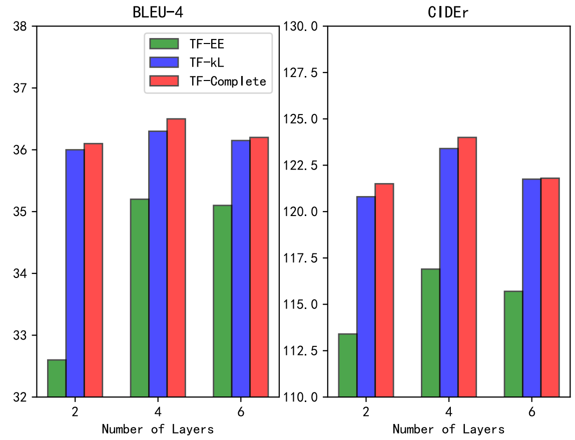


Figure 2. Performance comparison for different models with the same number of computation layers on the MS COCO dataset. Complete image captioning model capable of extracting semantic information clearly outperforms vanilla early exiting model which overlooks the high-level representations.

2. Investigations on Early Exiting

Early exiting strategy accelerates the model inference by stopping forward propagation based on the results of internal classifiers [44]. Specifically, if the internal classifier's prediction based on the current layer of hidden representation is confident enough, then the generation is terminated without passing through the residual layers. However, in image captioning, it is still remain unexplored that whether the local hidden representations could provide sufficient information for word generation and whether the intermediate classifiers are reliable for making exiting decision. Accordingly, we try to in-depth analyze the working mechanism of early exiting in the conventional image captioning model by answering these two questions.

2.1. Are Shallow Representations Sufficient?

It is believed that the Transformer-based model learns a hierarchy of visual and semantic representations [9, 10]. We highlight that high-level fusion features are essential even for easy tokens generation, and the predictions only based on shallow representations are prone to be inaccurate. To examine it, we evaluate the performance of the outputs of different decoder layers, as the representation containing adequate information is necessary for decent task performance. Specifically, we design the following Transformer-based (TF) models:

- **TF-EE**, which is a baseline of early exiting methods. The internal classifiers are appended after each caption decoder layer in the vanilla Transformer image captioning model for exiting generation early.
- **TF-kL**, which only utilizes the fixed first k decoder

layers for sentence generation. A classifier is added directly after the k-th layer. This model could be seen as a static early exiting variant.

- **TF-Complete**, which is a standard Transformer-based model with a classifier after the last decoder layer. The representations of this classifier contain sufficient high-level semantic and cross-modal fusion information, which is complete than the above two models.

We report the evaluation results with a different number of caption decoder layers on the MS COCO dataset [7]. The performance for TF-EE is sentence-level averaged and divided into bins. According to Figure 2, we can find that: (i) The TF-EE performs poorly, especially when the generations are made based on shallow representation. It indicates that the high-level semantic and cross-modal fusion information is important for the image describing. (ii) TF-kL outperforms TF-EE. We attribute it to that the latter several layers can learn more task-specific and comprehensive representation during optimizing. However, since the internal layer representation in TF-EE are restricted in the whole model learning, this fine-tuning effect cannot be fully exploited, resulting in a degrading performance in shallow layers. These findings verify the assumption that the high-level representation are necessary, motivating us to exploit alternative deep information in the uncomputed layers. What's more, the poor results of TF-EE on the generated tokens when it decides to stop early, also forces us further analysis on the quality of exiting decisions.

2.2. Are Internal Classifiers Reliable?

We further analyze whether the early exiting decisions made by internal classifiers in image captioning are reliable by first introducing two concepts referring [29, 39]:

- *Token Difficulty* $d(y_i)$, which denotes whether a token y_i can be generated easily by a learned image captioning model, under current context C_i including given image x and previous generated sub-sentence $y_{<i}$. We define instances that model cannot generate correctly as difficult tokens, *i.e.*, $d(y_i) = 1$, and those can be mastered well as easy ones, *i.e.*, $d(y_i) = 0$.

- *Prediction Confidence* $c(y_i)$, which indicates how confident and determinate the image captioning model is about its prediction for a specific token y_i under current context C_i. Here, we utilize the corresponding probability of y_i in the output vocabulary distribution as the prediction confidence score.

To utilize the prediction confidence as reference for dynamic early exiting decisions, a difficult token under the current context should be predicted with less confidence score than that of an easy one. However, there exists an over-confident problem [5, 52] where the prediction confidence is inconsistent with the token instance difficulty sometimes. To measure the seriousness of consistency phenomenon, we introduce the *False Confidence Score* (FCS). In detail, we first define a false confidence function for the token instance pair (y_i, y_j) to measure the inconsistency degree between prediction confidence and token difficulty as:

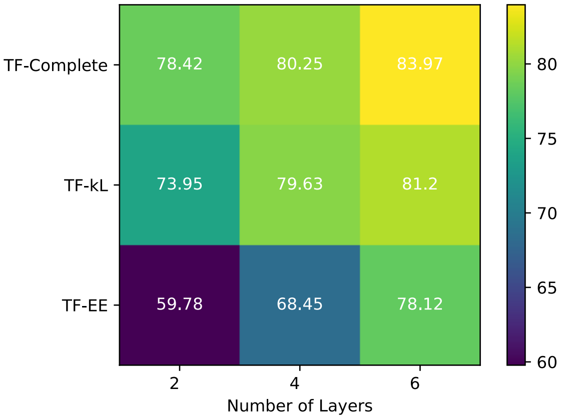


Figure 3. Heat map of evaluated FCS from different models on the MS COCO dataset. FCS of internal classifiers in the TF-EE shallow layers is lower than that of TF-kL and TF-Complete, which leads to more incorrect tokens being generated. The exiting decisions of TF-EE are unreliable sometimes.

$$\mathrm{FC}(y_i, y_j) = \begin{cases} 0 & \text{if } d(y_i) > d(y_j) \text{ and } c(y_i) < c(y_j) \\ 1 & \text{otherwise} \end{cases}. \quad (1)$$

We then sort the context-token pairs according to their confidence scores in an ascending order, *i.e.*, $c(y_i) < c(y_j)$ for any $i < j$. Finally, the dataset-level normalized sum of all false confidence pair can be computed as:

$$\mathrm{FCS} = 1 - \frac{1}{Q}\sum_{i=2}^{L}\sum_{j=1}^{i-1}\mathrm{FC}(y_i, y_j), \quad (2)$$

where L is the total number of context-token instance pairs in the evaluation dataset and Q is a normalizing factor calculated as a half of $L(L-1)$ to restrict the FCS value from 0 to 1. Following the above definition, the FCS metric estimates the ratio of context-token pairs that are correctly prioritized from the internal classifier. Intuitively, classifiers with higher FCS achieve better consistency between the prediction confidence and tokens difficulty, and making more reliable exiting decisions. Therefore, the FCS can be served as an effective approximate for evaluating the quality of early exiting decisions.

Experimentally, we compute the FCS on the MS COCO test set for different baselines discussed in the preceding section, and the results are illustrated in Figure 3. We can

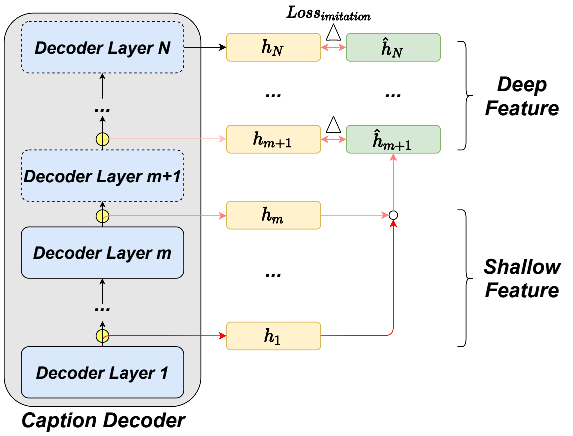


Figure 4. Illustration of high-level deep feature modeling with imitation learning. Early exiting at the m-th layer and the dashed line denotes the uncomputed layers at inference stage.

see that: (i) The FCS of TF-EE model in shallow layers are significantly lower than TF-kL and TF-Complete models. This demonstrates that the exiting decisions in the shallow layers of TF-EE are unreliable, and captioning performance can be accordingly worse when most tokens are predicted incorrectly in early layers. (ii) The capacity of determining the difficult contexts from easy ones is improved as the number of decoder layer increases. An important reason is that the deep representations holds sufficient semantic and cross-modal information, and it is possible for classifiers equipped with deep information to provide more effective exiting decisions. Our analysis demonstrates that directly transferring current early exiting method in image captioning for different decoder layers are not reliable, which inspires us to explore the modeling of deep information and sufficient representation fusion for more robust decisions.

3. Methodology

To remedy the drawbacks of directly executing the early exiting method in image captioning, we improve this idea from the perspective of making a comprehensive early-exit decision, which combines the both shallow and approximated deep representations with a gate mechanism.

3.1. Deep Representations Imitation

Early exiting method aims to stop generation at a shallow Transformer layer and ignore the deep representations captured in the deep layers. However, pilot analysis for image captioning in section 2 highlights that even the prediction of simple tokens relies on not only the surface low-level features but also high-level semantic information. It is actually infeasible to only consider low-level representations, which motivates us to exploit the high-level information. However, directly using deep representation is intractable since the deep states are inaccessible until feed forward the corresponding layers, which is not what we want. To bridge this gap, we introduce a method to approximate the uncomputed hidden states in deep layers referring to ***imitation learning*** [3, 42]. That is, we equip each decoder layer with a lightweight imitation network, which is encouraged to predict the representation of the real state of that layer based on the computed low-level representation. Through a layer-wise imitation process, we can get the deep hidden states with minimum cost. The workflow of the deep representation imitation is shown in Figure 4.

Formally, we denote the output hidden state of the m-th layer as h_m. To get the k-th layer's imitated deep representation when the feed forward propagation is executed at the m-th layer for any $k > m$, the k-th imitation network equipped to the k-th layer inputs the truly hidden state h_m and outputs an approximation $\hat{h}_k^m$ of the real deep representation h_k as $\hat{h}_k^m = \text{MLP}^k(h_m)$, where $\text{MLP}^k(\cdot)$ is a simple Multi-Layer Perceptron (MLP) network. We argue that, despite being limited in learning capacity, the MLP is sufficient for estimating deep representations. Then learning target h_k guides the k-th imitation network in a quick manner. In between, we utilize the cosine similarity as prediction performance measurement between the real deep representation h_k and the generated representation $\hat{h}_k^m$ as:

$$\text{Cos-Sim}(h_k, \hat{h}_k^m) = 1 - \frac{\hat{h}_k^m \cdot h_k}{\|\hat{h}_k^m\| \cdot \|h_k\|}, \tag{3}$$

where $\|\cdot\|$ denotes the L2 norm. Accordingly, we can compute the sum of the similarity difference between predicted hidden representations as $\frac{1}{N-m}\sum_{k=m+1}^{N}\text{Cos-Sim}(h_k, \hat{h}_k^m)$ when the feed forward propagation executes at the m-th layer. Considering that the layer m can be any number between 2 and N, we enumerate all possible layer numbers m, resulting in the overall loss of imitation network as:

$$\mathcal{L}_{imit} = \frac{1}{N-1}\frac{1}{N-m}\sum_{m=2}^{N}\sum_{k=m+1}^{N}\text{Cos-Sim}(h_k, \hat{h}_k^m). \tag{4}$$

The feed forward propagation is computed in all layers, and all imitation networks are encouraged to generate representations close to the real deep representations. Note that we pass through the entire N-layer image captioning model, but we simulate the situation that the feed forward propagation ends up at the m-th layer for any $m < N$.

3.2. Multi-Level Representations Fusion

After obtaining the real low-level and imitated high-level representation, we investigate how to aggregate all the hidden states into one, respectively. Formally, when the feed forward propagation proceeds to the m-th layer, all the previously generated shallow hidden states is $\{h_1, \ldots, h_m\}$, the subsequent imitation networks take the m-th real state as input to generate the approximations of deep representations from the (m+1)-th layer to the N-th layer as $\{\hat{h}^m_{m+1}, \ldots, \hat{h}^m_N\}$. Hence, the fusion of multi-level of shall and deep representation can be computed as:

$$h_{shallow} = g(\{h_1, \ldots, h_m\}), \tag{5}$$

$$h_{deep} = g(\{\hat{h}^m_{m+1}, \ldots, \hat{h}^m_N\}), \tag{6}$$

where $g(\cdot)$ refers to the feature fusion strategy.

For the fusion of a variable number of multi-level feature, we explore the following four strategies to aggregate multi-level representations into one as:

- **Average.** The average strategy sums and averages all hidden representations in different layers directly.
- **Concatenation.** All the hidden representations are concatenated in the sequence dimension and then fed into a linear transformation layer to obtain a final compressed representation.
- **Attention-pooling.** The attention-pooling strategy utilizes the weighted projection of all hidden representation as the integrated information. The attention weights are computed with the last hidden representation as the query and hold certain robustness to noise.
- **Sequential Network.** All multi-level representations are sequentially fed into a LSTM network, and the output hidden state of the last time-step is regarded as the fusion representation.

3.3. Gate Decision Mechanism

We finally explore how to merge the shallow and deep hidden representation for early exiting decision. Intuitively, the shallow representation $h_{shallow}$ and the deep representation h_{deep} are of different confidence since the truly generated low-level representations are more reliable than predicted deep representations. In addition, different token difficulty requires high-level representation differently. Therefore, it is necessary to develop a decision mechanism to combine the low-level and high-level representation dynamically. In practice, we design a gate network to incorporate the both representation into decision. When comes to the i-th layer, we compute the fusion of different level representation, and the merged inputting is a trade-off between these two as:

$$\alpha = \sigma(\text{MLP}([h_{shallow}, h_{deep}])), \tag{7}$$

$$z_m = \alpha h_{shallow} + (1-\alpha) h_{deep}, \tag{8}$$

where z_m represents the merged information for the input of internal classifier and MLP is a multi-layer perceptron network for the merging gate.

Under the DeeCap framework, each decoder layer can produce imitated deep representations and a final merged representation z_m which is used for early exiting decision. Then the entire model will be updated with the layer-wise cross-entropy loss following the provided ground-truth token y_i. The gate decision mechanism dynamically learns to adjust the balance of low-level and high-level information under the supervision signal from ground-truth tokens and the corresponding loss can be formalized as:

$$p_m = \text{softmax}(z_m), \tag{9}$$

$$\mathcal{L}_{ce} = -\sum_{m=1}^{N} \sum_{y_i \in V} [y_i \log(p_m(y_i))]. \tag{10}$$

The final training objective can be combined as:

$$\mathcal{L} = \lambda \mathcal{L}_{ce} + (1-\lambda)\mathcal{L}_{imit}, \tag{11}$$

where λ denotes a balancing factor to adjust the impact of imitation networks and internal classifiers learning.

3.4. Training and Inference

We train the model according to the loss in Equation 11 with the following two-fold improvements: (1) The shallow decoder layers will be updated more frequently as they receive more updating signals with the original layer-equal objectives. Therefore, we heuristically re-weight the cross-entropy loss of each decoder layer depending on its depth m as: $w_m = \frac{m}{\sum_{k=1}^{N} k}$. (2) Directly updating all parameters of image captioning model at each step may damage the well-trained features learned in the preceding stage. Therefore, we try to balance the requirements of maintaining previous learned parameters and adapting to new domain at fine-tuning gradually. To be specific, the parameters of a layer may be frozen with a probability p, and the probability p linearly decreases from the first decoder layer to the last decoder layer in a range of 1 to 0.

During inference, we model the prediction confidence e of current token with the calculated entropy $\mathcal{H}$ of the output distribution p_m of the m-th layer as $e(p_m) = \mathcal{H}(p_m)$. The inference stops once the confidence $e(p_m)$ is lower than a predefined threshold τ. The hyper-parameter τ can be adjusted according to the required speed-up ratios. Note that if the exiting condition is never reached, our model degrades into the conventional case of inference that the complete computation in decoding layers is executed.

4. Experiments

4.1. Experimental Preparation

Dataset. MS COCO [7] and Flickr30k [40] image captioning datasets are used for evaluation. They contain 123,287 images and 31,783 images, respectively. There are

Models	BLEU-1	BLEU-4	METEOR	ROUGE	CIDEr	SPICE	SpeedUp
Autoregressive Image Captioning models							
NIC-v2 [50]	-	32.1	25.7	-	99.8	-	-
Up-Down [2]	79.8	36.3	27.7	56.9	120.1	21.4	-
AoANet [22]	80.2	38.9	29.2	58.8	129.8	22.4	-
M2-T [10]	80.8	39.1	29.2	58.6	131.2	22.6	-
TF-Complete	80.2	38.8	29.0	58.3	129.5	22.7	1.00×
Non-Autoregressive Image Captioning models							
MNIC [16]	75.4	30.9	27.5	55.6	108.1	21.0	2.80×
FNIC [12]	-	36.2	27.1	55.3	115.7	20.2	8.15×
MIR [23]	-	32.5	27.2	55.4	109.5	20.6	1.56×
CMAL [20]	80.3	37.3	28.1	58.0	124.0	21.8	13.90×
IBM [13]	77.2	36.6	27.8	56.2	113.2	20.9	3.06×
SAIC [57]	80.3	38.4	29.0	58.1	127.1	21.9	3.42×
Early Exiting-based Image Captioning models							
TF-EE	79.8	37.2	28.2	57.7	126.3	21.8	**4.54**×
DeeCap	80.1	**38.7**	29.1	58.1	**129.0**	22.5	**4.35**×

Table 1. Performance comparison of different captioning models using different evaluation metrics on the MS COCO Karpathy test set. All values except SpeedUp are reported as a percentage (%).

	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE-L		CIDEr	
Models	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
Up-Down* [2]	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
AoANet* [22]	81.0	95.0	65.8	89.6	51.4	81.3	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6
M2-T* [10]	81.6	96.0	66.4	90.8	51.8	82.7	39.7	72.8	29.4	39.0	59.2	74.8	129.3	132.1
CMAL [20]	79.8	94.3	63.8	87.2	48.8	77.2	36.8	66.1	27.9	36.4	57.6	72.0	119.3	121.2
DeeCap	80.5	95.1	65.2	89.1	50.3	80.0	38.1	69.5	28.0	37.0	58.4	73.5	121.4	124.4

Table 2. Leaderboard of different image captioning models on the online MS COCO test server. * denotes the ensemble model.

5 human-annotated descriptions per image. To be consistent with previous works [10,22], we convert all the descriptions to lower case and omit words which occur less than 5 times. Image features are pre-extracted following [2].

Evaluation Metrics. For performance evaluation, five metrics are utilized: BLEU@N [38], METEOR [26], ROUGE-L [32], CIDEr-D [49], SPICE [1]. For efficiency estimation, as the measurement of runtime might not be stable [54], we manually adjust the exiting threshold τ and calculate the speed-up ratio by comparing the actually executed layers in forward propagation with the complete layers. For an N-layer model, the SpeedUp ratio is calculated as: $\frac{\sum_{m=1}^{N} N \times w^m}{\sum_{m=1}^{N} m \times w^m}$, where w^m is the number of words that exit at the m-th layer of caption decoder.

Implementation Details. The proposed DeeCap model closely follows the same network architecture and hyper-parameters settings as Transformer basic model [48]. Specifically, the number of stacked blocks for the visual encoder is 6, and for caption decoder is 6, hidden size is 512, and feed-forward network size is 2048. We train the model for 25 epochs with an initial learning rate of 3e-5, and it decays by 0.9 every five epochs [41]. Adam [25] optimizer is employed. We perform a grid search for frozen layer number during fine-tuning over $\{0, 1, 2, 3\}$. We find that small models need more time to converge. The best model is selected based on the validation performance. The decoding time is measured on a single NVIDIA GeForce GTX 1080 Ti as prior works reported [13,20]. All speeds are measured by running three times and reporting the average value.

4.2. Overall Results

Comparison with State-of-the-Arts. For a fair comparison, we adopt the Karpathy split [24] for the MS COCO dataset, for which ground-truth annotations are not publicly available. The performance comparison with vanilla early exiting methods and other top-performing non-autoregressive accelerating models in MS COCO offline test set is presented in Table 1. In addition, the evaluation results for Flickr30k are provided in the appendix. We can find that early exiting-based methods can achieve more than 4x acceleration. In terms of performance, DeeCap

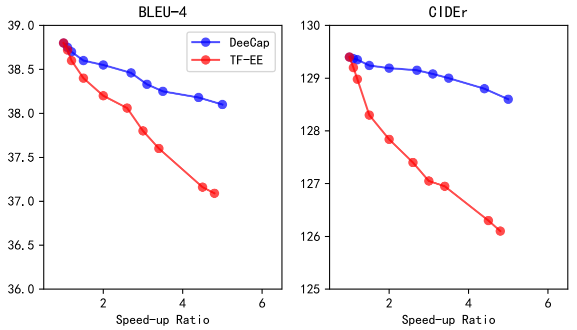


Figure 5. Performance and efficiency trade-off line for early exiting models. Our method outperforms original early exiting method by a large margin especially with high speed-up ratio.

outperforms all the fast baselines, especially in BLEU@4 and CIDEr. Even compared with the basic autoregressive baselines without considering the acceleration, it has competitive performances. This phenomenon demonstrates that DeeCap can break the performance bottleneck with a high speed-up ratio by utilizing early exiting and comprehensive representations from both high-level and low-level information. Moreover, DeeCap outperforms TF-EE in all performance metrics, validating the effectiveness of our proposal. We also report the performance of our DeeCap on the online MS COCO test server. Results are reported in Table 2. It can be seen that, compared with the ensemble model, our approach has little performance loss when accelerating generation and achieves an advancement of 2.1 CIDEr points with respect to the best accelerating performer CMAL [20].

Performance and Efficiency Trade-off. To verify the robustness and efficiency of our proposed DeeCap, we visualize the performance and efficiency trade-off curves in Figure 5 on the MS COCO test set. The competitive baseline is the original early exiting method TF-EE. As can be seen, the performance of early exiting model drops dramatically when the speed-up ratio increases. This reflects the shortcomings of TF-EE, unstable performance can not meet the needs of real-time applications, while our DeeCap demonstrates more tolerance of speed-up ratio. At the same speed-up ratio, the performance loss of DeeCap is less than one-third compared with TF-EE, indicating that early exiting with multi-level feature fusion is effective. In addition, the model adjusts the speed-up ratio on this curve without retraining, which is suitable for engineering applications.

Methods	FCS (↑)	B-4	C
DeeCap (× 2)	80.12	38.9	129.5
-w/o Deep Info.	78.67	38.5	128.3
DeeCap (× 4)	82.40	38.7	129.0
-w/o Deep Info.	79.55	38.2	127.8

Table 3. Effect of the incorporation of approximated deep representation under various speed-up ratios.

Methods	B-4	M	R	C	S
Average	38.3	28.8	57.7	127.3	21.9
Concatenation	38.7	29.1	58.1	129.0	22.5
Attention-Pooling	38.6	29.0	57.9	128.7	22.3
SeqNN	38.5	29.0	58.0	129.0	22.3

Table 4. Performance comparison of different fusion strategies for multi-level hidden representations.

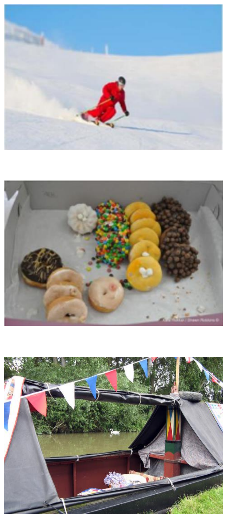
4.3. Model Analysis

Effects of Deep Information Imitation. To assess whether and how imitated deep representation from deep layers contributes to the current word decision, we first evaluate the performance changes of our DeeCap method on the MS COCO offline test set. The results shown in Table 3 demonstrate the impact of deep information incorporation. We can observe that the global fusion mechanism brings improvement on most metrics for both 2× speed-up ratio and 4× speed-up ratio, which confirms that the approximations of deep representations help enhance the model ability in prediction. Beyond that, the deep representation can be especially advantageous for the models with a higher speed-up ratio. Recall that approximations of deep representation complement the high-level information, and the exiting at shallow layers loses more semantic representation compared with the exiting at deep layers. Therefore, the benefit of deep information is more significant for the exiting at shallow layers, which is validated by the larger improvement gap with a 4× speed-up ratio.

Effects of Representation Fusion Strategies. The results of different shallow representation incorporation strategies on the MS COCO offline test set are shown in Table 4. The naive average strategies perform poorly, which reflects that focusing on local strategy does not perform well. On the contrary, three simple yet effective global strategies designed to combine all of the past hidden states bring significant improvement compared to baselines. Within them, we empirically find that the concatenation strategy works best from an overall point of view. We assume that such a strategy allows interaction among different states, yielding a better captioning performance.

4.4. Case Study

For more intuitive understanding, we present several examples of generated image captions from vanilla early exiting (TF-EE) and the proposed DeeCap models, which hold the same model architecture, coupled with human-



GT: a person in a red suit down the mountain
TF-EE: a person riding skis on a mountain
DeeCap: a person in red clothes down a mountain

GT: a box that contains multiple kinds of doughnuts
TF-EE: a box with donuts in in
DeeCap: a box filled with lots of different doughnuts

GT: a boat with flags and tents is docked next to a grassy bank
TF-EE: a boat sitting on the water
DeeCap: a boat with different flags sitting on the water

Figure 6. Case studies of original early exiting (TF-EE) and the proposed DeeCap model, coupled with the corresponding ground-truth sentences (GT) for image caption generation.

annotated ground-truth sentences (GT) in Figure 6. As we can be seen, in general, both models hold the capability to reflect the content of the image accurately. Meantime, some semantic problems, including repeated words and incomplete content, is severe in the sentence generated by pure early exiting, while it can be effectively alleviated by DeeCap, *i.e.*, two "in" terms in the second sample with nothing behind. This confirms our approach can guide the model to reduce word prediction errors effectively.

4.5. Human Evaluation

Following previous works [22,60], we also conduct a human evaluation test to compare DeeCap model against the original early exiting method. To be specific, we randomly selected 300 samples from the MS COCO testing set and recruited eight workers to evaluate model performances. Each time, we show only one sentence paired with a corresponding image generated by different models or human annotation and ask: can you determine whether the given sentence has been generated by a system or a person? We then calculate the captions that pass the Turing test. The results of Human, DeeCap, and original early exiting are 91.7%, 82.0%, and 61.3%, separately. It demonstrates the superiority of DeeCap fused with high-level and low-level information in providing human-like captions.

5. Related Works

Efficient Image Captioning. Current image captioning systems mainly follow an autoregressive manner [4], meaning that the model generates captions word by word and is not suitable for parallel execution. Several recent works attempt to accelerate generation by using a non-autoregressive framework [19, 51], which produces the entire sentences simultaneously. Fei *et al.* [12] reorders words detected in the image to form better latent variables before decoding. Cho *et al.* [23] and Gao *et al.* [16] introduce an iterative mask refinement strategy to learn the position matching information. Lu *et al.* [20] addresses the inconsistency problem with a multi-agent learning paradigm. The biggest difference lies that our DeeCap adopts sample-level speed-up acceleration for inference via adapting the computation according to the sample complexity while all previous methods focus on model structure adjustment.

Early Exiting Strategy. A representative acceleration framework for sample difficulty [18] is early exiting [54]. Prior works have mainly been used for image classification. Deeply-supervised nets [27] and BranchyNet [47] propose architectures that are composed of a cascade of intermediate classifiers. This allows simpler examples to exit early via an intermediate classifier while more difficult samples proceed deeper in the network for more accurate predictions. Multi-scale dense networks [21] and adaptive resolution networks [59] focus on spatial redundancy of input samples and use a multi-scale dense connection architecture for stopping. Early exiting has also been verified in natural language understanding [31, 34, 44, 63], sequence labeling [30], text classification [29], question answering [45], and document ranking [53]. Following these, we study an working mechanism of early exiting in image captioning, and try to deal with the performance bottleneck with multi-level representation fusion.

6. Conclusion

In this paper, we point out that applying vanilla early exiting strategy in image captioning faces the performance bottleneck, due to insufficient cross-modal representations and poor decisions of the internal classifiers. To alleviate this problem, We propose a dynamic early exiting method for efficient image captioning from a multi-level perspective. Unlike previous works only utilizing local hidden representation, DeeCap model employs a novel approach to approximate and engage the multi-level representation from different layers, which are originally inaccessible for prediction. Experiments illustrate that our approach achieves significant improvement over the original early stopping baseline with a high speed-up ratio, suggesting the superiority in application prospects.

Acknowledgment

This work was supported in part by the National Key R&D Program of China under Grant 2018AAA0102000, and in part by National Natural Science Foundation of China: 62022083, 61836002 and 61931008.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Proc. ECCV*, pages 382–398, 2016. 6

[2] 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. 1, 6

[3] Alexandre Attia and Sharone Dayan. Global overview of imitation learning. *arXiv preprint arXiv:1801.06503*, 2018. 4

[4] Shuang Bai and Shan An. A survey on automatic image caption generation. *Neurocomputing*, 311:291–304, 2018. 1, 8

[5] Ali Furkan Biten, Lluis Gomez, and Dimosthenis Karatzas. Let there be a clock on the beach: Reducing object hallucination in image captioning. *arXiv preprint arXiv:2110.01705*, 2021. 3

[6] Long Chen, Zhihong Jiang, Jun Xiao, and Wei Liu. Human-like controllable image captioning with verb-specific semantic roles. In *Proc. IEEE CVPR*, pages 16846–16856, 2021. 1

[7] 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. 3, 5

[8] Xinlei Chen and C Lawrence Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *Proc. IEEE CVPR*, pages 2422–2431, 2015. 1

[9] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019. 2

[10] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proc. IEEE CVPR*, pages 10578–10587, 2020. 1, 2, 6

[11] Maha Elbayad, Jiatao Gu, Edouard Grave, and Michael Auli. Depth-adaptive transformer. In *Proc. ICLR*, pages 1–14, 2020. 2

[12] Zhengcong Fei. Fast image caption generation with position alignment. *arXiv preprint arXiv:1912.06365*, 2019. 1, 2, 6, 8

[13] Zhengcong Fei. Iterative back modification for faster image captioning. In *Proc. ACM MM*, pages 3182–3190, 2020. 1, 2, 6

[14] Zhengcong Fei. Memory-augmented image captioning. In *Proc. AAAI*, pages 2–9, 2021. 1

[15] Zhengcong Fei. Partially non-autoregressive image captioning. In *Proc. AAAI*, volume 35, pages 1309–1316, 2021. 2

[16] Junlong Gao, Xi Meng, Shiqi Wang, Xia Li, Shanshe Wang, Siwei Ma, and Wen Gao. Masked non-autoregressive image captioning. *arXiv preprint arXiv:1906.00717*, 2019. 1, 2, 6, 8

[17] Amir Ghodrati, Babak Ehteshami Bejnordi, and Amirhossein Habibian. Frameexit: Conditional early exiting for efficient video recognition. In *Proc. IEEE CVPR*, pages 15608–15618, 2021. 2

[18] Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*, 2016. 8

[19] Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. Non-autoregressive neural machine translation. In *Proc. ICLR*, 2018. 1, 8

[20] Longteng Guo, Jing Liu, Xinxin Zhu, Xingjian He, Jie Jiang, and Hanqing Lu. Non-autoregressive image captioning with counterfactuals-critical multi-agent learning. *arXiv preprint arXiv:2005.04690*, 2020. 1, 2, 6, 7, 8

[21] Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, Laurens van der Maaten, and Kilian Q Weinberger. Multi-scale dense networks for resource efficient image classification. *arXiv preprint arXiv:1703.09844*, 2017. 8

[22] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proc. IEEE ICCV*, pages 4634–4643, 2019. 1, 6, 8

[23] Lee Jason, Mansimov Elman, Graham Neubig, and Cho Kyunghyun. Deterministic non-autoregressive neural sequence modeling by iterative refinement. In *Proc. EMNLP*, pages 1138–1149, 2018. 1, 6, 8

[24] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proc. IEEE CVPR*, pages 3128–3137, 2015. 6

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[26] 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. 6

[27] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial intelligence and statistics*, pages 562–570. PMLR, 2015. 8

[28] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. Entangled transformer for image captioning. In *Proc. IEEE ICCV*, pages 8928–8937, 2019. 1

[29] Lei Li, Yankai Lin, Deli Chen, Shuhuai Ren, Peng Li, Jie Zhou, and Xu Sun. Cascadebert: Accelerating inference of pre-trained language models via calibrated complete models cascade. *arXiv preprint arXiv:2012.14682*, 2020. 3, 8

[30] Xiaonan Li, Yunfan Shao, Tianxiang Sun, Hang Yan, Xipeng Qiu, and Xuanjing Huang. Accelerating bert inference for sequence labeling via early-exit. *arXiv preprint arXiv:2105.13878*, 2021. 8

[31] Kaiyuan Liao, Yi Zhang, Xuancheng Ren, Qi Su, Xu Sun, and Bin He. A global past-future early exit method for accelerating inference of pre-trained language models. In *Proc. NAACL*, pages 2013–2023, 2021. 8

[32] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Proc. ACL Workshops*, pages 74–81, 2004. 6

[33] Fenglin Liu, Xuancheng Ren, Xian Wu, Shen Ge, Wei Fan, Yuexian Zou, and Xu Sun. Prophet attention: Predicting attention with future attention for improved image captioning. In *Proc. NIPS*, 2021. 1

[34] Weijie Liu, Peng Zhou, Zhiruo Wang, Zhe Zhao, Haotang Deng, and Qi Ju. Fastbert: a self-distilling bert with adaptive inference time. In *Proc. ACL*, pages 6035–6044, 2020. 8

[35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 2

[36] Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin, and Rongrong Ji. Dual-level collaborative transformer for image captioning. In *Proc. AAAI*, volume 35, pages 2286–2293, 2021. 1

[37] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *Proc. IEEE CVPR*, pages 10971–10980, 2020. 1

[38] 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. 6

[39] Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabás Póczos, and Tom Mitchell. Competence-based curriculum learning for neural machine translation. In *Proc. NACCL*, pages 1162–1172, 2019. 3

[40] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proc. IEEE ICCV*, pages 2641–2649, 2015. 5

[41] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proc. IEEE CVPR*, pages 1179–1195, 2017. 6

[42] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proc. ICAIS*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011. 4

[43] Stefan Schaal. Learning from demonstration. *Proc. NIPS*, 9, 1996. 2

[44] Roy Schwartz, Gabriel Stanovsky, Swabha Swayamdipta, Jesse Dodge, and Noah A Smith. The right tool for the job: Matching model and instance complexities. In *Proc. ACL*, 2020. 2, 8

[45] Luca Soldaini and Alessandro Moschitti. The cascade transformer: an application for efficient answer sentence selection. In *Proc. ACL*, pages 5697–5708, 2020. 8

[46] Tianxiang Sun, Yunhua Zhou, Xiangyang Liu, Xinyu Zhang, Hao Jiang, Zhao Cao, Xuanjing Huang, and Xipeng Qiu. Early exiting with ensemble internal classifiers. *arXiv preprint arXiv:2105.13792*, 2021. 2

[47] Surat Teerapittayanon, Bradley McDanel, and Hsiang-Tsung Kung. Branchynet: Fast inference via early exiting from deep neural networks. In *Proc. ICPR*, pages 2464–2469. IEEE, 2016. 8

[48] 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. 1, 6

[49] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proc. IEEE CVPR*, pages 4566–4575, 2015. 6

[50] 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. 1, 6

[51] Bingzhen Wei, Mingxuan Wang, Hao Zhou, Junyang Lin, Jun Xie, and Xu Sun. Imitation learning for non-autoregressive neural machine translation. *arXiv preprint arXiv:1906.02041*, 2019. 8

[52] Yijun Xiao and William Yang Wang. On hallucination and predictive uncertainty in conditional language generation. *arXiv preprint arXiv:2103.15025*, 2021. 3

[53] Ji Xin, Rodrigo Nogueira, Yaoliang Yu, and Jimmy Lin. Early exiting bert for efficient document ranking. In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 83–88, 2020. 8

[54] Ji Xin, Raphael Tang, Jaejun Lee, Yaoliang Yu, and Jimmy Lin. Deebert: Dynamic early exiting for accelerating bert inference. In *Proc. ACL*, pages 2246–2251, 2020. 2, 6, 8

[55] Guanghui Xu, Shuaicheng Niu, Mingkui Tan, Yucheng Luo, Qing Du, and Qi Wu. Towards accurate text-based image captioning with content diversity exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12637–12646, 2021. 1

[56] 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. 1

[57] Xu Yan, Zhengcong Fei, Zekang Li, Shuhui Wang, Qingming Huang, and Qi Tian. Semi-autoregressive image captioning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2708–2716, 2021. 2, 6

[58] Bang Yang, Fenglin Liu, Can Zhang, and Yuexian Zou. Non-autoregressive coarse-to-fine video captioning. *arXiv preprint arXiv:1911.12018*, 2019. 1

[59] Le Yang, Yizeng Han, Xi Chen, Shiji Song, Jifeng Dai, and Gao Huang. Resolution adaptive networks for efficient inference. In *Proc. IEEE CVPR*, pages 2369–2378, 2020. 8

[60] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proc. ECCV*, pages 684–699, 2018. 1, 8

[61] 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. 1

[62] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. *arXiv preprint arXiv:1909.11059*, 2019. 1

[63] Wangchunshu Zhou, Canwen Xu, Tao Ge, Julian McAuley, Ke Xu, and Furu Wei. Bert loses patience: Fast and robust inference with early exit. *Proc. NIPS*, 33, 2020. 8

[64] Yuanen Zhou, Yong Zhang, Zhenzhen Hu, and Meng Wang. Semi-autoregressive transformer for image captioning. In *Proc. ICCV Workshop*, pages 3139–3143, 2021. 2