NATIONAL INSTITUTE OF TECHNOLOGY ANDHRA PRADESH
(An autonomous Institute under the aegis of Ministry of Education, Government of India)
DEPARTMENT OF ELECTRICAL ENGINEERING

सत्यमेव जयते

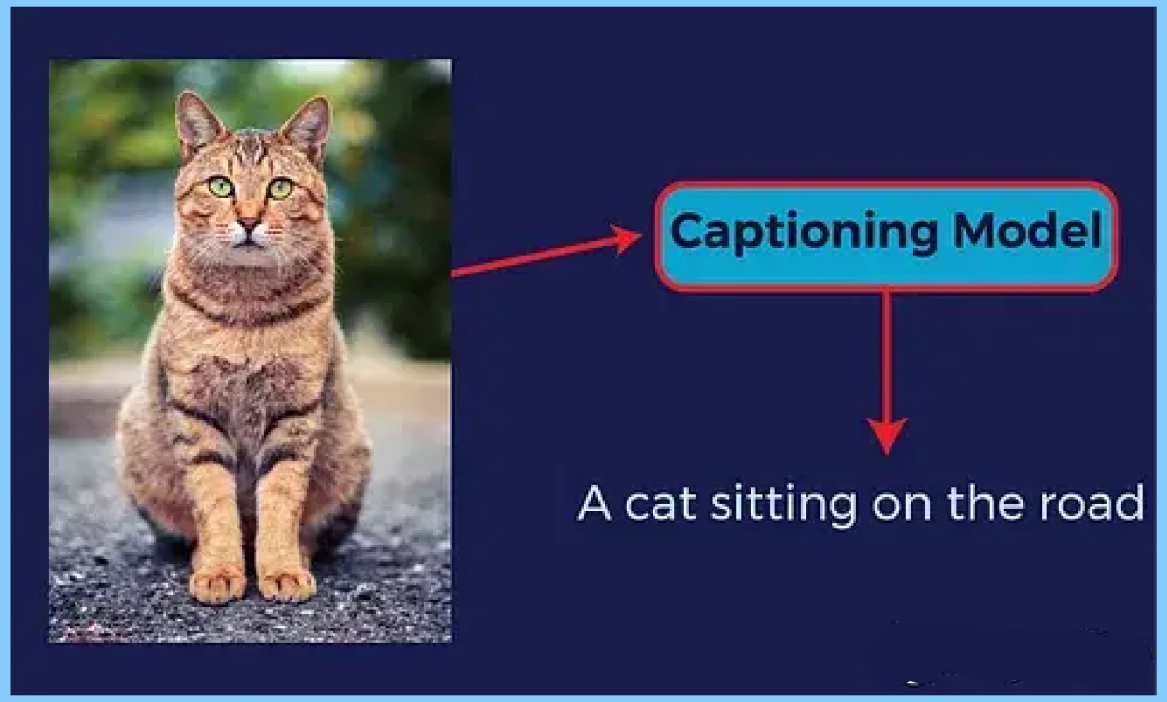
Image Captioning based on CLIP-DIFFUSION LANGUAGE MODEL

Supervisor: Dr. K. Sri Phani Krishna

521214 Joshi P
521144 Manish Reddy
521225 Eswar Chowdary

Abstract

Image captioning task has been extensively researched area. However, limited experiments focus on generating captions based on a non-autoregressive text decoder. Inspired by the recent success of the denoising diffusion model on image synthesis tasks, we apply denoising diffusion probabilistic models to text generation in image captioning tasks. We show that our CLIP-Diffusion-LM is capable of generating image captions using significantly fewer inference steps than autoregressive models. On the Flickr8k dataset, the model achieves 0.1876 BLEU-4 score. By training on the combined Flickr8k and Flickr30k dataset, our model achieves 0.2470 BLEU-4 score.



Introduction

Image captioning has been a focus of research over recent years. Among the previous proposed works, the text encoder used can be classified into 2 general classes, i.e. autoregressive and non-autoregressive class. Most of the state-of-the-art models fall in the autoregressive class. However, autoregressive generation suffer from

1) the slow generation speed due to the generation step is token by token, and

2) not capable of refining prefix of sentences based on the later generated tokens.

Multiple attempts have experimented using a non-auto regressive model in the text generation steps. The closest work to ours is Masked Non-Autoregressive Image Captioning which uses a BERT model as the generator. In particular, we use pre-trained CLIP model for extracting image and text features, and DistilBert model based on Diffusion-LM for text sequence generation.

Background- DIFFUSION MODELS

Diffusion model aims at training a model that denoise Gaussial noise incrementally to reproduce original features. The Denoising Diffusion Probabilistic Model (DDPM) is to simplify the loss function by only letting models predict the noise in generation steps, and proposed an alternative loss function by removing the weight coefficients. In the following explanation, we refer to diffusion model as DDPM for simplicity. Several improvements based on DDPM, including setting variance to be learn-able parameters, apply cosine instead of linear noise schedule, and speed up forward process by reducing forward steps. Their work explored various techniques to improve the performance of continuous diffusion model on text generation.

Background- DIFFUSION MODELS

The classifier guidance for improving generated image FID score. In a classifier-guided diffusion model, a classifier model is pretrained to predict noised images' object class. During training, the classifier provides gradient on which direction to optimise the generated image, so that the generate image resembles an object closer to the target class. To avoid training classifier for guiding model, a classifier-free guidance model is proposed. In classifier-free guidance, the difference between outputs of generative model when provided with either guided and unguided context information is used as implicit guidance. By using classifier-free diffusion model as text-to-image generator, DALL-E2 and High-Resolution Image Synthesis With Latent Diffusion Models model achieves significant image generation performance. In particular, DALL-E2 use CLIP model for extracting feature from text, predict the corresponding image CLIP feature through prior network, then use predicted image CLIP feature for final image generation. The model achieves significant novelty in generated images and also inspired us to train a image-to-text model with diffusion model in generation step.

Mathematical Formulas

$$q(x_t|x_{t-1}) = N(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

$$E_q[-\log(p_\theta(x_0))] = E_q[-\log(\int p_\theta(x_{0:T})d(x_{1:T}))]\,.$$

$$\begin{aligned} E_q[-\log(p_\theta(x_0))] &\leq E_q[\log(\frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)})] \\ &= E_q[\log(p(x_T)) + \sum_{t=1}^{T}\log(\frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})})] \end{aligned}$$

$$L_{simple} = \sum_{t=1}^{T} E_{q(x_t|x_0)}\|\mu_\theta(x_t,t) - \hat{\mu}(x_t,x_0)\|^2$$

Architecture

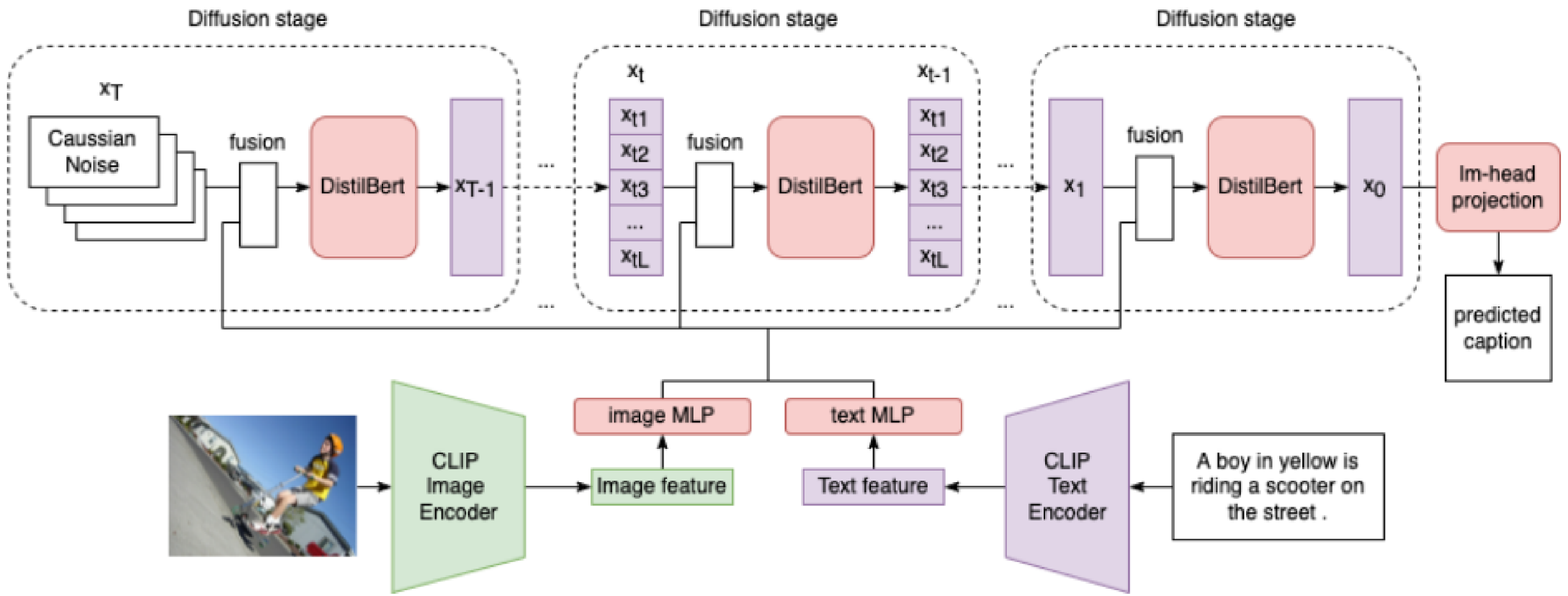
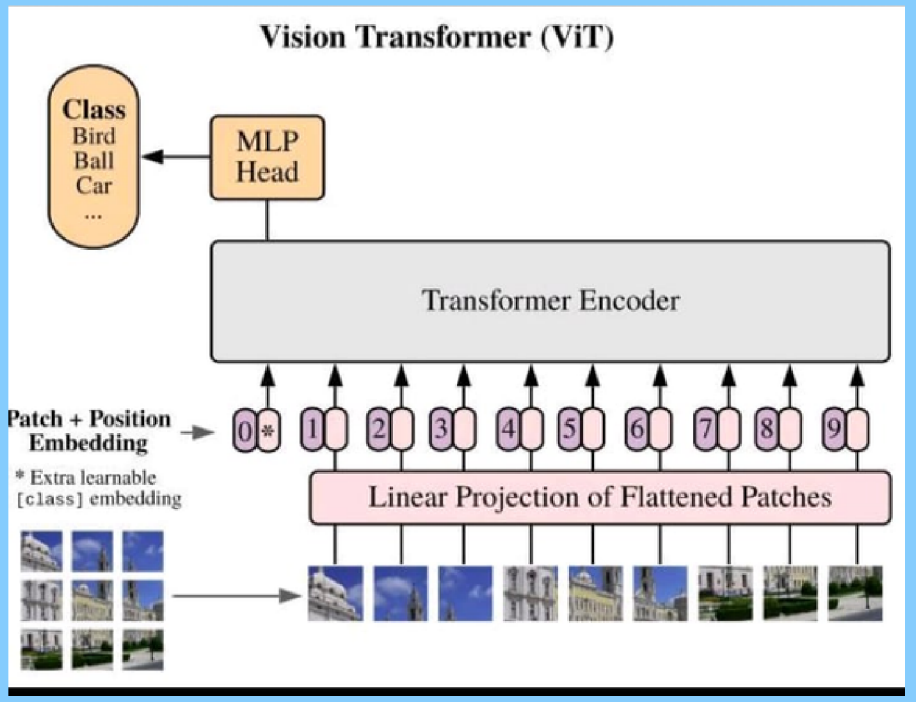


Figure 1: CLIP-Diffusion-LM model

Learning Process

CLIP model is a contrastive-learning-based model trained on WebImageText dataset. The WebImageText consists of 400 million image-text pairs collected from publicly available sources on the Internet. CLIP model demonstrates strong zero-shot performance in their evaluation on ImageNet dataset.
The CLIP-Diffusion-LM model operates differently during training and inference stages. Here's a breakdown for each:

Training and Inference:
Input: An image and its corresponding caption.
CLIP Model (pre-trained):
Extracts image features using the ViT-Base/32 architecture.
Extracts caption features using a modified transformer encoder.
Both image and caption features are not fine-tuned during training.
Diffusion Process:
The model goes through multiple diffusion stages.
At each stage:



Learning Process

The current caption embedding (or noise in the first stage) and CLIP features (image & text) are fed into the model.
A Multi-Layer Perceptron (MLP) projects CLIP features to the same dimension as the caption embedding.
The fused embedding (caption + image & text features) is fed into the DistilBert model.
DistilBert predicts the previous stage's caption embedding (denoising).
This denoising process continues through all stages.
Loss Calculation:
The final predicted caption embedding is compared to the original caption embedding using a loss function.
The model backpropagates to optimize the DistilBert weights (except embedding layer) for better denoising.
During training, the model learns to improve denoising by comparing the generated captions with the ground truth captions.
In inference, there's no ground truth available. The model uses the pre-trained CLIP features and DistilBert to iteratively refine the caption based on the image content.

Website

Conclusion

We present the application of diffusion in the image caption task and prove its validity on Flickr8k and Flickr30k datasets. Particularly, we identify the importance of rounding term in loss function to help model converge, and introduce the adaptive ratio adjustment to balance its importance with other terms. By training on the combined Flickr8k and Flickr30k dataset, our model achieves good BLEU-4 score. Finally, after evaluating this, we will develop a website implementing this methodology using flask.