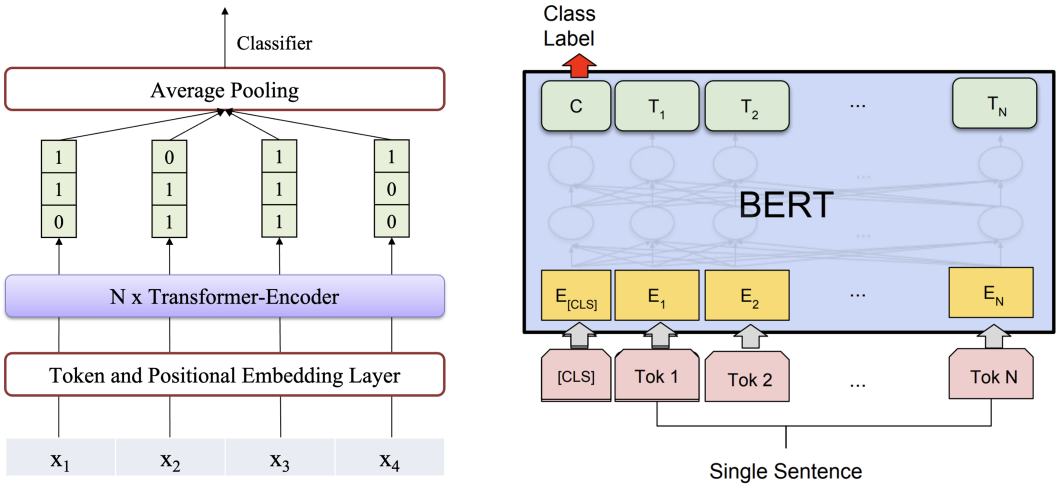
AI VIET NAM – COURSE 2024

Transformer Applications

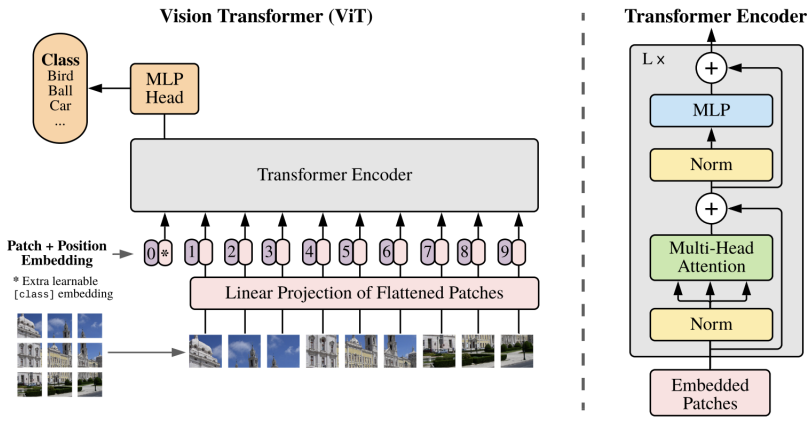
Thai Nguyen-Quoc, Vinh Dinh-Quang

Ngày 21 tháng 12 năm 2024

Phần 1. Transformer



Hình 1: Mô hình Transformer-Encoder và BERT cho bài toán phân loại văn bản.

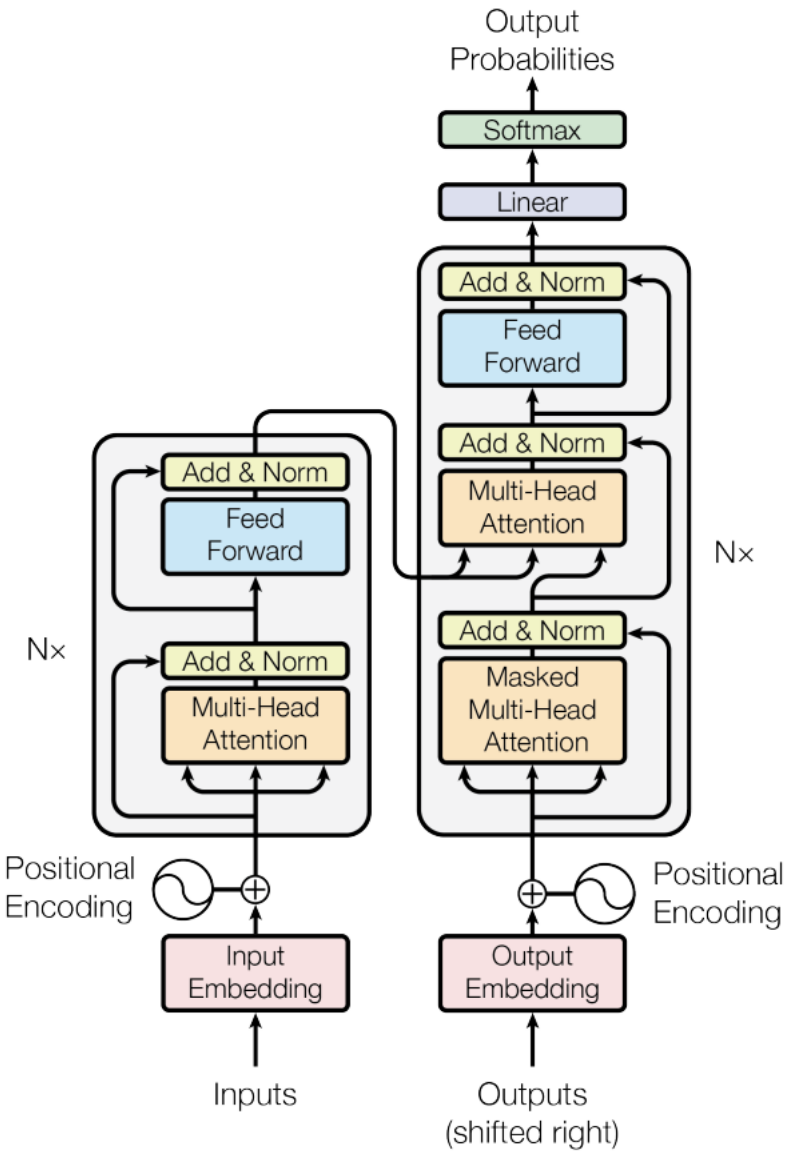


Hình 2: Mô hình Vision Transformer cho bài toán phân loại hình ảnh.

Mô hình Transformer ra đời với kỹ thuật cốt lõi và cơ chế Attention, đã đạt được những kết quả ấn tượng cho các bài toán về dữ liệu văn bản rồi từ đó mở rộng cho các kiểu dữ liệu như hình ảnh và

âm thanh,... Trong phần này, chúng ta sẽ tìm hiểu các thành phần trong mô hình Transformer và ứng dụng cho bài toán phân loại văn bản.

1.1. Kiến trúc Transformer



Hình 3: Mô hình kiến trúc Transformer.

Kiến trúc Transformer gồm các thành phần:

- Input Embedding: Biểu diễn các token đầu vào (Thường được tách bởi Subword-based Tokenization) thành các dense vector.
- Positional Encoding: Biểu diễn vị trí (thứ tự) của các token trong câu. Thường được tính dựa vào hàm sinusoid hoặc được học trong quá trình huấn luyện mô hình.
- Các khối encoder: Để mã hoá các tokens đầu vào thành các contextual embedding. Bao gồm: Multi-Head Attention, Add - Normalization, Feed Forward

- Các khối decoder: nhận input là các token lịch sử và trạng thái mã hoá từ encoder, giải mã dự đoán token tiếp theo. Gồm: Masked Multi-Head Attention (dựa vào token lịch sử cửa decoder), Multi-Head Attention(dựa vào encoder và trạng thái hiện tại decoder), Add - Normalization, Feed Forward
- Language Model Head: Projection và Softmax dự đoán token tiếp theo vs xác suất lớn nhất.

1. Input Embedding, Positional Encoding

```python
class TokenAndPositionEmbedding(nn.Module):
    def __init__(self, vocab_size, embed_dim, max_length, device='cpu'):
        super().__init__()
        self.device = device
        self.word_emb = nn.Embedding(
            num_embeddings=vocab_size,
            embedding_dim=embed_dim
        )
        self.pos_emb = nn.Embedding(
            num_embeddings=max_length,
            embedding_dim=embed_dim
        )
            đầu vào x là một tensor
    def forward(self, x):
        N, seq_len = x.size()
        positions = torch.arange(0, seq_len).expand(N, seq_len).to(self.device)
        output1 = self.word_emb(x)
        output2 = self.pos_emb(positions)
        output = output1 + output2
        return output
```

(?)1. word embedding khác với pos embedding như thế nào

(?)2. batch size: là số từ trong từ điển có

(?)3. sequence length: số chữ số mà một một từ được biểu diễn thành (số token trong mỗi mẫu)

2. Encoder

```python
class TransformerEncoderBlock(nn.Module):
    def __init__(self, embed_dim, num_heads, ff_dim, dropout=0.1):
        super().__init__()
        self.attn = nn.MultiheadAttention(
            embed_dim=embed_dim,
            num_heads=num_heads,
            batch_first=True
        )
        self.ffn = nn.Sequential(
            nn.Linear(in_features=embed_dim, out_features=ff_dim, bias=True),
            nn.ReLU(),
            nn.Linear(in_features=ff_dim, out_features=embed_dim, bias=True)
        )
        self.layernorm_1 = nn.LayerNorm(normalized_shape=embed_dim, eps=1e-6)
        self.layernorm_2 = nn.LayerNorm(normalized_shape=embed_dim, eps=1e-6)
        self.dropout_1 = nn.Dropout(p=dropout)
        self.dropout_2 = nn.Dropout(p=dropout)

    def forward(self, query, key, value):
        attn_output, _ = self.attn(query, key, value)
        attn_output = self.dropout_1(attn_output)
        out_1 = self.layernorm_1(query + attn_output)
        ffn_output = self.ffn(out_1)
        ffn_output = self.dropout_2(ffn_output)
        out_2 = self.layernorm_2(out_1 + ffn_output)
        return out_2

class TransformerEncoder(nn.Module):
    def __init__(self,
```

(!) Học lại lý thuyết MultiheadAttention

```
                src_vocab_size, embed_dim, max_length, num_layers, num_heads, ff_dim,
                dropout=0.1, device='cpu'
        ):
        super().__init__()
        self.embedding = TokenAndPositionEmbedding(
            src_vocab_size, embed_dim, max_length, device
        )
        self.layers = nn.ModuleList(
            [
                TransformerEncoderBlock(
                    embed_dim, num_heads, ff_dim, dropout
                ) for i in range(num_layers)
            ]
        )

    def forward(self, x):
        output = self.embedding(x)
        for layer in self.layers:
            output = layer(output, output, output)
        return output
```

3. Decoder

```
class TransformerDecoderBlock(nn.Module):
    def __init__(self, embed_dim, num_heads, ff_dim, dropout=0.1):
        super().__init__()
        self.attn = nn.MultiheadAttention(
            embed_dim=embed_dim,
            num_heads=num_heads,
            batch_first=True
        )
        self.cross_attn = nn.MultiheadAttention(
            embed_dim=embed_dim,
            num_heads=num_heads,
            batch_first=True
        )
        self.ffn = nn.Sequential(
            nn.Linear(in_features=embed_dim, out_features=ff_dim, bias=True),
            nn.ReLU(),
            nn.Linear(in_features=ff_dim, out_features=embed_dim, bias=True)
        )
        self.layernorm_1 = nn.LayerNorm(normalized_shape=embed_dim, eps=1e-6)
        self.layernorm_2 = nn.LayerNorm(normalized_shape=embed_dim, eps=1e-6)
        self.layernorm_3 = nn.LayerNorm(normalized_shape=embed_dim, eps=1e-6)
        self.dropout_1 = nn.Dropout(p=dropout)
        self.dropout_2 = nn.Dropout(p=dropout)
        self.dropout_3 = nn.Dropout(p=dropout)

    def forward(self, x, enc_output, src_mask, tgt_mask):
        attn_output, _ = self.attn(x, x, x, attn_mask=tgt_mask)
        attn_output = self.dropout_1(attn_output)
        out_1 = self.layernorm_1(x + attn_output)

        attn_output, _ = self.cross_attn(
            out_1, enc_output, enc_output, attn_mask=src_mask
        )
        attn_output = self.dropout_2(attn_output)
        out_2 = self.layernorm_2(out_1 + attn_output)

        ffn_output = self.ffn(out_2)
        ffn_output = self.dropout_3(ffn_output)
```

```
        out_3 = self.layernorm_3(out_2 + ffn_output)
        return out_3

class TransformerDecoder(nn.Module):
    def __init__(self,
            tgt_vocab_size, embed_dim, max_length, num_layers, num_heads, ff_dim,
            dropout=0.1, device='cpu'
        ):
        super().__init__()
        self.embedding = TokenAndPositionEmbedding(
            tgt_vocab_size, embed_dim, max_length, device
        )
        self.layers = nn.ModuleList(
            [
                TransformerDecoderBlock(
                    embed_dim, num_heads, ff_dim, dropout
                ) for i in range(num_layers)
            ]
        )

    def forward(self, x, enc_output, src_mask, tgt_mask):
        output = self.embedding(x)
        for layer in self.layers:
            output = layer(output, enc_output, src_mask, tgt_mask)
        return output
```

4. Transformer

```
class Transformer(nn.Module):
    def __init__(self,
            src_vocab_size, tgt_vocab_size,
            embed_dim, max_length, num_layers, num_heads, ff_dim,
            dropout=0.1, device='cpu'
        ):
        super().__init__()
        self.device = device
        self.encoder = TransformerEncoder(
            src_vocab_size, embed_dim, max_length, num_layers, num_heads, ff_dim
        )
        self.decoder = TransformerDecoder(
            tgt_vocab_size, embed_dim, max_length, num_layers, num_heads, ff_dim
        )
        self.fc = nn.Linear(embed_dim, tgt_vocab_size)

    def generate_mask(self, src, tgt):
        src_seq_len = src.shape[1]
        tgt_seq_len = tgt.shape[1]

        src_mask = torch.zeros(
            (src_seq_len, src_seq_len),
            device=self.device).type(torch.bool)

        tgt_mask = (torch.triu(torch.ones(
            (tgt_seq_len, tgt_seq_len),
            device=self.device)
        ) == 1).transpose(0, 1)
        tgt_mask = tgt_mask.float().masked_fill(
            tgt_mask == 0, float('-inf')).masked_fill(tgt_mask == 1, float(0.0))
        return src_mask, tgt_mask

    def forward(self, src, tgt):
```

```
        src_mask, tgt_mask = self.generate_mask(src, tgt)
        enc_output = self.encoder(src)
        dec_output = self.decoder(tgt, enc_output, src_mask, tgt_mask)
        output = self.fc(dec_output)
        return output
```

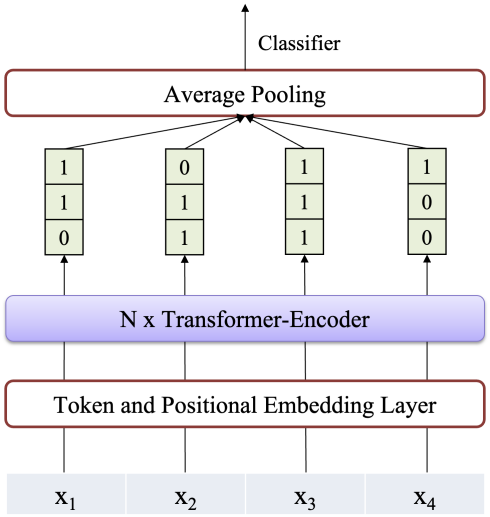
5. Thử nghiệm

```
batch_size = 128
src_vocab_size = 1000
tgt_vocab_size = 2000
embed_dim = 200
max_length = 100
num_layers = 2
num_heads = 4
ff_dim = 256

model = Transformer(
    src_vocab_size, tgt_vocab_size,
    embed_dim, max_length, num_layers, num_heads, ff_dim
)

src = torch.randint(
    high=2,
    size=(batch_size, max_length),
    dtype=torch.int64
)

tgt = torch.randint(
    high=2,
    size=(batch_size, max_length),
    dtype=torch.int64
)

prediction = model(src, tgt)
prediction.shape # batch_size x max_length x tgt_vocab_size
```

(?) cho đầu vào là các kích thước chứ không phải là các câu à

1. Mục đích của model này
2. Chạy thử nghiệm transformer một câu xem như nào
3. Sử dụng như thế nào

1.2. Text Classification



Hình 4: Phân loại văn bản sử dụng Transformer-Encoder

Ở phần này, chúng ta sử dụng mô hình Transformer-Encoder cho bài toán phân loại văn bản. Kiến trúc mô hình gồm các phần:

- Input Embedding và Positional Encoding
- Các lớp Encoder
- Lớp Average Pooling: lấy trung bình biểu diễn các từ thành biểu diễn cho câu
- Classifier: Linear layer

1. Load Dataset

```
!pip install datasets

from datasets import load_dataset

ds = load_dataset('thainq107/ntc-scv')
```

2. Preprocessing

Áp dụng hàm tiền xử lý sau trên cột 'sentence' hoặc có thể bỏ qua bước tiền xử lý khi áp dụng trên cột 'preprocessed_sentence'.

```
import re
import string

def preprocess_text(text):
    # remove URLs https://www.
    url_pattern = re.compile(r'https?://\s+\wwww\.\s+')
    text = url_pattern.sub(r" ", text)

    # remove HTML Tags: <>
    html_pattern = re.compile(r'<[^<>]+>')
    text = html_pattern.sub(" ", text)

```

```
    # remove puncs and digits
    replace_chars = list(string.punctuation + string.digits)
    for char in replace_chars:
        text = text.replace(char, " ")

    # remove emoji
    emoji_pattern = re.compile("["
        u"\U0001F600-\U0001F64F"  # emoticons
        u"\U0001F300-\U0001F5FF"  # symbols & pictographs
        u"\U0001F680-\U0001F6FF"  # transport & map symbols
        u"\U0001F1E0-\U0001F1FF"  # flags (iOS)
        u"\U0001F1F2-\U0001F1F4"  # Macau flag
        u"\U0001F1E6-\U0001F1FF"  # flags
        u"\U0001F600-\U0001F64F"
        u"\U00002702-\U000027B0"
        u"\U000024C2-\U0001F251"
        u"\U0001f926-\U0001f937"
        u"\U0001F1F2"
        u"\U0001F1F4"
        u"\U0001F620"
        u"\u200d"
        u"\u2640-\u2642"
        "]+", flags=re.UNICODE)
    text = emoji_pattern.sub(r" ", text)

    # normalize whitespace
    text = " ".join(text.split())

    # lowercasing
    text = text.lower()
    return text
```

3. Representation

```
# !pip install -q torchtext==0.17.2 (install before import torch)

def yield_tokens(sentences, tokenizer):
    for sentence in sentences:
        yield tokenizer(sentence)


# word-based tokenizer
from torchtext.data import get_tokenizer
tokenizer = get_tokenizer("basic_english")

# build vocabulary
from torchtext.vocab import build_vocab_from_iterator

vocab_size = 10000
vocabulary = build_vocab_from_iterator(
    yield_tokens(ds['train']['preprocessed_sentence'], tokenizer),
    max_tokens=vocab_size,
    specials=["<pad>", "<unk>"]
)
vocabulary.set_default_index(vocabulary["<unk>"])

# convert torchtext dataset
from torchtext.data.functional import to_map_style_dataset

def prepare_dataset(df):
    # create iterator for dataset: (sentence, label)
```

```
    for row in df:
        sentence = row['preprocessed_sentence']
        encoded_sentence = vocabulary(tokenizer(sentence))
        label = row['label']
        yield encoded_sentence, label

train_dataset = prepare_dataset(ds['train'])
train_dataset = to_map_style_dataset(train_dataset)

valid_dataset = prepare_dataset(ds['valid'])
valid_dataset = to_map_style_dataset(valid_dataset)

test_dataset = prepare_dataset(ds['test'])
test_dataset = to_map_style_dataset(test_dataset)
```

4. Dataloader

```
import torch

seq_length = 100

def collate_batch(batch):
    # create inputs, offsets, labels for batch
    sentences, labels = list(zip(*batch))
    encoded_sentences = [
        sentence+([0]* (seq_length-len(sentence))) if len(sentence) < seq_length else sentence[:seq_length]
        for sentence in sentences
    ]

    encoded_sentences = torch.tensor(encoded_sentences, dtype=torch.int64)
    labels = torch.tensor(labels)

    return encoded_sentences, labels

from torch.utils.data import DataLoader

batch_size = 128

train_dataloader = DataLoader(
    train_dataset,
    batch_size=batch_size,
    shuffle=True,
    collate_fn=collate_batch
)
valid_dataloader = DataLoader(
    valid_dataset,
    batch_size=batch_size,
    shuffle=False,
    collate_fn=collate_batch
)

test_dataloader = DataLoader(
    test_dataset,
    batch_size=batch_size,
    shuffle=False,
    collate_fn=collate_batch
)
```

5. Trainer

```
# train epoch
```

```python
import time

def train_epoch(model, optimizer, criterion, train_dataloader, device, epoch=0, log_interval=50):
    model.train()
    total_acc, total_count = 0, 0
    losses = []
    start_time = time.time()

    for idx, (inputs, labels) in enumerate(train_dataloader):
        inputs = inputs.to(device)
        labels = labels.to(device)

        optimizer.zero_grad()

        predictions = model(inputs)

        # compute loss
        loss = criterion(predictions, labels)
        losses.append(loss.item())

        # backward
        loss.backward()
        optimizer.step()
        total_acc += (predictions.argmax(1) == labels).sum().item()
        total_count += labels.size(0)
        if idx % log_interval == 0 and idx > 0:
            elapsed = time.time() - start_time
            print(
                "| epoch {:3d} | {:5d}/{:5d} batches "
                "| accuracy {:8.3f}".format(
                    epoch, idx, len(train_dataloader), total_acc / total_count
                )
            )
            total_acc, total_count = 0, 0
            start_time = time.time()

    epoch_acc = total_acc / total_count
    epoch_loss = sum(losses) / len(losses)
    return epoch_acc, epoch_loss

# evaluate
def evaluate_epoch(model, criterion, valid_dataloader, device):
    model.eval()
    total_acc, total_count = 0, 0
    losses = []

    with torch.no_grad():
        for idx, (inputs, labels) in enumerate(valid_dataloader):
            inputs = inputs.to(device)
            labels = labels.to(device)

            predictions = model(inputs)

            loss = criterion(predictions, labels)
            losses.append(loss.item())

            total_acc += (predictions.argmax(1) == labels).sum().item()
            total_count += labels.size(0)

```

```python
    epoch_acc = total_acc / total_count
    epoch_loss = sum(losses) / len(losses)
    return epoch_acc, epoch_loss


# train
def train(model, model_name, save_model, optimizer, criterion, train_dataloader, valid_dataloader, num_epochs, device):
    train_accs, train_losses = [], []
    eval_accs, eval_losses = [], []
    best_loss_eval = 100
    times = []
    for epoch in range(1, num_epochs+1):
        epoch_start_time = time.time()
        # Training
        train_acc, train_loss = train_epoch(model, optimizer, criterion, train_dataloader, device, epoch)
        train_accs.append(train_acc)
        train_losses.append(train_loss)

        # Evaluation
        eval_acc, eval_loss = evaluate_epoch(model, criterion, valid_dataloader, device)
        eval_accs.append(eval_acc)
        eval_losses.append(eval_loss)

        # Save best model
        if eval_loss < best_loss_eval:
            torch.save(model.state_dict(), save_model + f'/{model_name}.pt')

        times.append(time.time() - epoch_start_time)
        # Print loss, acc end epoch
        print("-" * 59)
        print(
            "| End of epoch {:3d} | Time: {:5.2f}s | Train Accuracy {:8.3f} | Train Loss {:8.3f} "
            "| Valid Accuracy {:8.3f} | Valid Loss {:8.3f} ".format(
                epoch, time.time() - epoch_start_time, train_acc, train_loss, eval_acc, eval_loss
            )
        )
        print("-" * 59)

    # Load best model
    model.load_state_dict(torch.load(save_model + f'/{model_name}.pt'))
    model.eval()
    metrics = {
        'train_accuracy': train_accs,
        'train_loss': train_losses,
        'valid_accuracy': eval_accs,
        'valid_loss': eval_losses,
        'time': times
    }
    return model, metrics

# report
import matplotlib.pyplot as plt

def plot_result(num_epochs, train_accs, eval_accs, train_losses, eval_losses):
    epochs = list(range(num_epochs))
```

```
    fig, axs = plt.subplots(nrows = 1, ncols =2 , figsize = (12,6))
    axs[0].plot(epochs, train_accs, label = "Training")
    axs[0].plot(epochs, eval_accs, label = "Evaluation")
    axs[1].plot(epochs, train_losses, label = "Training")
    axs[1].plot(epochs, eval_losses, label = "Evaluation")
    axs[0].set_xlabel("Epochs")
    axs[1].set_xlabel("Epochs")
    axs[0].set_ylabel("Accuracy")
    axs[1].set_ylabel("Loss")
    plt.legend()
```

6. Modeling

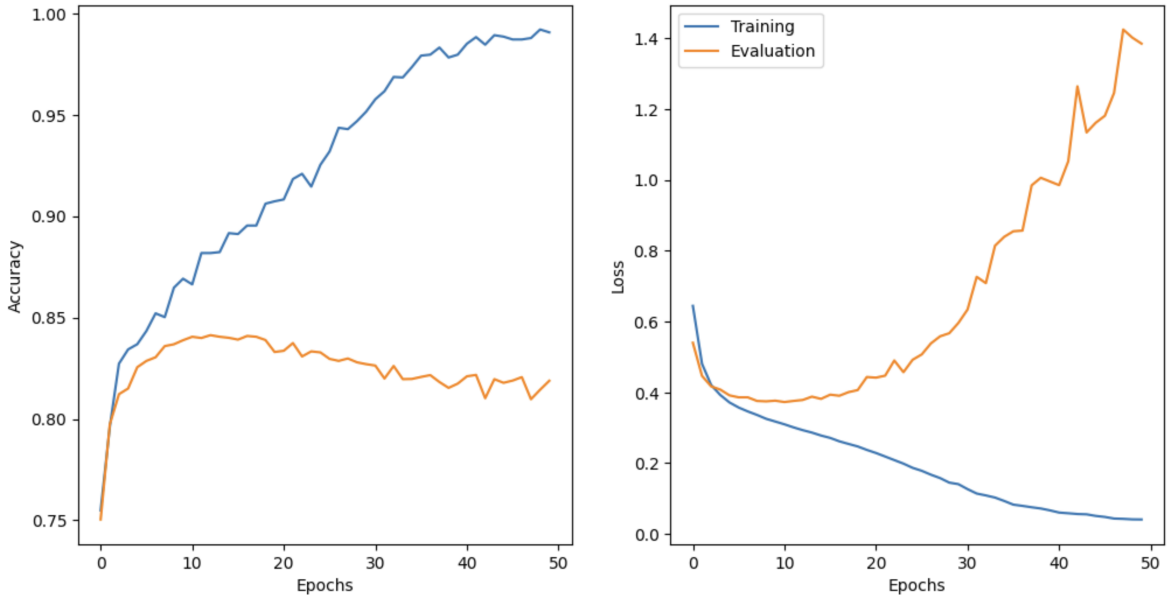
```
class TransformerEncoderCls(nn.Module):
    def __init__(self,
                 vocab_size, max_length, num_layers, embed_dim, num_heads, ff_dim,
                 dropout=0.1, device='cpu'
        ):
        super().__init__()
        self.encoder = TransformerEncoder(
            vocab_size, embed_dim, max_length, num_layers, num_heads, ff_dim, dropout,
    device
        )
        self.pooling = nn.AvgPool1d(kernel_size=max_length)
        self.fc1 = nn.Linear(in_features=embed_dim, out_features=20)
        self.fc2 = nn.Linear(in_features=20, out_features=2)
        self.dropout = nn.Dropout(p=dropout)
        self.relu = nn.ReLU()
    def forward(self, x):
        output = self.encoder(x)
        output = self.pooling(output.permute(0,2,1)).squeeze()
        output = self.dropout(output)
        output = self.fc1(output)
        output = self.dropout(output)
        output = self.fc2(output)
        return output
```

7. Training

```
import torch.optim as optim

vocab_size = 10000
max_length = 100
embed_dim = 200
num_layers = 2
num_heads = 4
ff_dim = 128
dropout=0.1

model = TransformerEncoderCls(
    vocab_size, max_length, num_layers, embed_dim, num_heads, ff_dim, dropout
)

device = torch.device('cuda' if torch.cuda.is_available() else 'cpu')

model = TransformerEncoderCls(
    vocab_size, max_length, num_layers, embed_dim, num_heads, ff_dim, dropout, device
)
model.to(device)

criterion = torch.nn.CrossEntropyLoss()
optimizer = optim.Adam(model.parameters(), lr=0.00005)
```

```
num_epochs = 50
save_model = './model'
os.makedirs(save_model, exist_ok = True)
model_name = 'model'

model, metrics = train(
    model, model_name, save_model, optimizer, criterion, train_dataloader,
    valid_dataloader, num_epochs, device
)
```

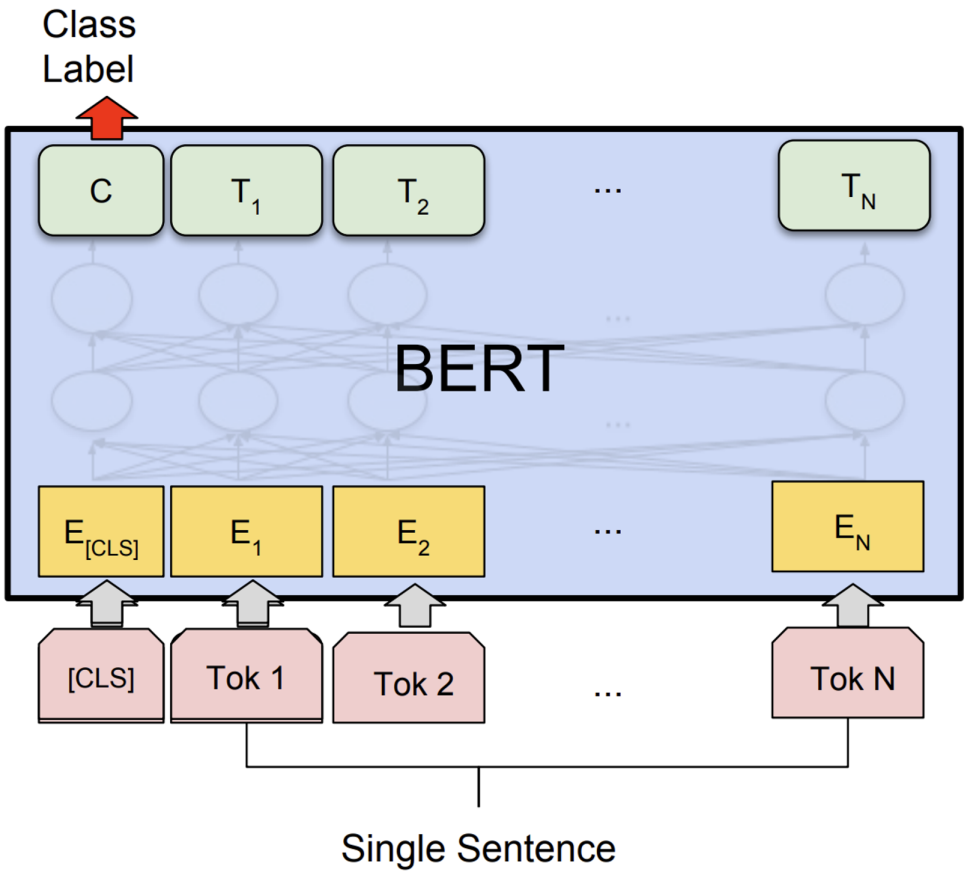
Kết quả training và accuracy trên tập test là 82%



Hình 5: Quá trình huấn luyện bộ dữ liệu NTC-SCV với mô hình Transformer-Encoder.

Phần 2. Text Classification using BERT

Một trong những mô hình pretrained đầu tiên cho dữ liệu văn bản dựa vào kiến trúc mô hình Transformer được ứng dụng cho các downstream task khác nhau đó là BERT. Trong phần này chúng ta sẽ fine tuning BERT cho bài toán phân loại trên bộ dữ liệu NTC-SCV dựa vào thư viện transformers của huggingface.



Hình 6: Fine-Tuning BERT cho bài toán phân loại văn bản.

1. Load Dataset

```
# install libs
!pip install -q -U transformers datasets accelerate evaluate

from datasets import load_dataset

ds = load_dataset('thainq107/ntc-scv')
```

2. Preprocessing

```
# tokenization
from transformers import AutoTokenizer

model_name = "distilbert-base-uncased" # bert-base-uncased

```

```
tokenizer = AutoTokenizer.from_pretrained(
    model_name,
    use_fast=True
)
max_seq_length = 100
max_seq_length = min(max_seq_length, tokenizer.model_max_length)

def preprocess_function(examples):
    # Tokenize the texts

    result = tokenizer(
        examples["preprocessed_sentence"],
        padding="max_length",
        max_length=max_seq_length,
        truncation=True
    )
    result["label"] = examples['label']

    return result

# Running the preprocessing pipeline on all the datasets
processed_dataset = ds.map(
    preprocess_function,
    batched=True,
    desc="Running tokenizer on dataset",
)
```

3. Modeling

```
from transformers import AutoConfig, AutoModelForSequenceClassification

num_labels = 2

config = AutoConfig.from_pretrained(
    model_name,
    num_labels=num_labels,
    finetuning_task="text-classification"
)
model = AutoModelForSequenceClassification.from_pretrained(
    model_name,
    config=config
)
```

4. Metric

```
import numpy as np
import evaluate

metric = evaluate.load("accuracy")
def compute_metrics(eval_pred):
    predictions, labels = eval_pred
    predictions = np.argmax(predictions, axis=1)
    result = metric.compute(predictions=predictions, references=labels)
    return result
```

5. Trainer

```
from transformers import TrainingArguments, Trainer

```

```
training_args = TrainingArguments(
    output_dir="save_model",
    learning_rate=2e-5,
    per_device_train_batch_size=128,
    per_device_eval_batch_size=128,
    num_train_epochs=10,
    eval_strategy="epoch",
    save_strategy="epoch",
    load_best_model_at_end=True
)

trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=processed_dataset["train"],
    eval_dataset=processed_dataset["valid"],
    compute_metrics=compute_metrics,
    tokenizer=tokenizer,
)

trainer.train()
```

6. Training

Kết quả training và accuracy trên tập test là 85%

Epoch	Training Loss	Validation Loss	Accuracy
1	No log	0.457404	0.790200
2	No log	0.432181	0.804500
3	No log	0.419964	0.812100
4	No log	0.395226	0.827100
5	0.431800	0.398116	0.831200
6	0.431800	0.388693	0.834300
7	0.431800	0.403755	0.831600
8	0.431800	0.408512	0.833300
9	0.316800	0.409772	0.835000
10	0.316800	0.410361	0.836200

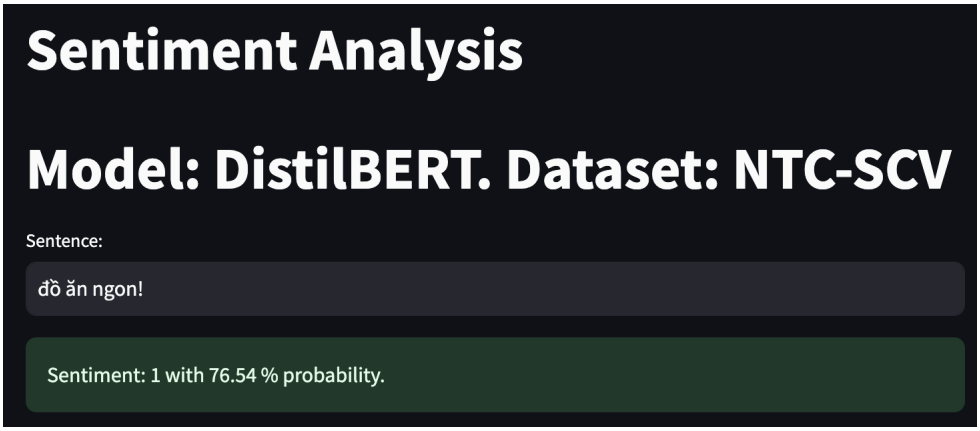
Hình 7: Quá trình huấn luyện bộ dữ liệu NTC-SCV dựa vào fine tuning BERT.

7. Triển khai mô hình

Dựa trên mô hình đã huấn luyện và thư viện streamlit để triển khai ứng dụng:

1. Link Streamlit
2. Github

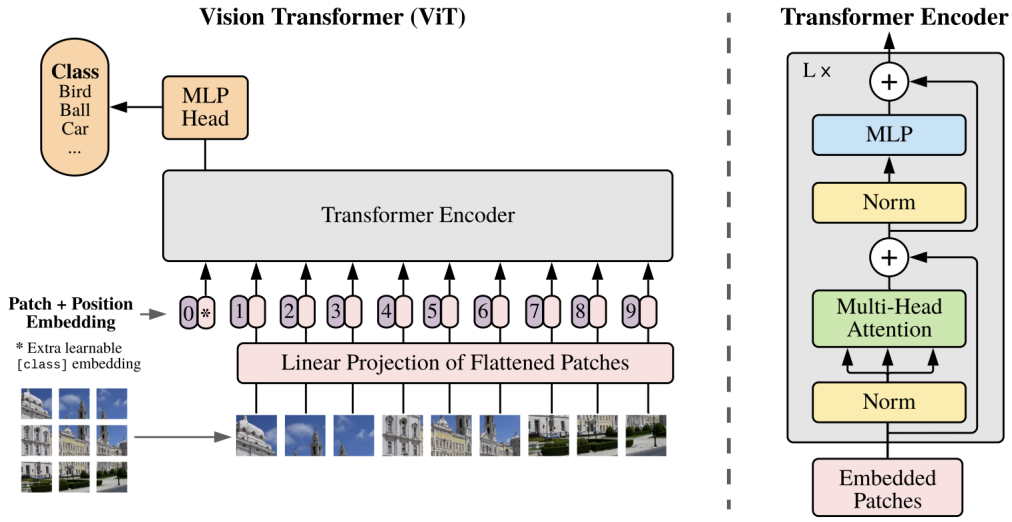
3. Giao diện:



Hình 8: Giao diện ứng dụng.

Phần 3. Vision Transformer

Trong phần này, chúng ta sẽ huấn luyện mô hình Vision Transformer cho bài toán phân loại hình ảnh.



Hình 9: Mô hình Vision Transformer cho bài toán phân loại hình ảnh.

1. Load Dataset

```
import torch
import torchvision.transforms as transforms
from torch.utils.data import DataLoader, random_split
import torch.optim as optim
from torchvision.datasets import ImageFolder
from torch import nn
import math
import os
```

```
# download
!gdown 1vSevps_hV5zhVf6aWuN8X7dd-qSAIgcc
!unzip ./flower_photos.zip

# load data
data_patch = "./flower_photos"
dataset = ImageFolder(root=data_patch)
num_samples = len(dataset)
classes = dataset.classes
num_classes = len(dataset.classes)

# split
TRAIN_RATIO, VALID_RATIO = 0.8, 0.1
n_train_examples = int(num_samples * TRAIN_RATIO)
n_valid_examples = int(num_samples * VALID_RATIO)
n_test_examples = num_samples - n_train_examples - n_valid_examples
train_dataset, valid_dataset, test_dataset = random_split(
    dataset,
    [n_train_examples, n_valid_examples, n_test_examples]
)
```

2. Preprocessing

```
# resize + convert to tensor
IMG_SIZE = 224

train_transforms = transforms.Compose([
    transforms.Resize((IMG_SIZE, IMG_SIZE)),
    transforms.RandomHorizontalFlip(),
    transforms.RandomRotation(0.2),
    transforms.ToTensor(),
    transforms.Normalize([0.5, 0.5, 0.5], [0.5, 0.5, 0.5])
])

test_transforms = transforms.Compose([
    transforms.Resize((IMG_SIZE, IMG_SIZE)),
    transforms.ToTensor(),
    transforms.Normalize([0.5, 0.5, 0.5], [0.5, 0.5, 0.5])
])

# apply
train_dataset.dataset.transform = train_transforms
valid_dataset.dataset.transform = test_transforms
test_dataset.dataset.transform = test_transforms
```

3. Dataloader

```
BATCH_SIZE = 512

train_loader = DataLoader(
    train_dataset,
    shuffle=True,
    batch_size=BATCH_SIZE
)

val_loader = DataLoader(
    valid_dataset,
    batch_size=BATCH_SIZE
)

test_loader = DataLoader(
    test_dataset,
    batch_size=BATCH_SIZE
)
```

4. Training from Scratch

4.1. Modeling

```
class TransformerEncoder(nn.Module):
    def __init__(self, embed_dim, num_heads, ff_dim, dropout=0.1):
        super().__init__()
        self.attn = nn.MultiheadAttention(
            embed_dim=embed_dim,
            num_heads=num_heads,
            batch_first=True
        )
        self.ffn = nn.Sequential(
            nn.Linear(in_features=embed_dim, out_features=ff_dim, bias=True),
            nn.ReLU(),
            nn.Linear(in_features=ff_dim, out_features=embed_dim, bias=True)
```

```
        )
        self.layernorm_1 = nn.LayerNorm(normalized_shape=embed_dim, eps=1e-6)
        self.layernorm_2 = nn.LayerNorm(normalized_shape=embed_dim, eps=1e-6)
        self.dropout_1 = nn.Dropout(p=dropout)
        self.dropout_2 = nn.Dropout(p=dropout)

    def forward(self, query, key, value):
        attn_output, _ = self.attn(query, key, value)
        attn_output = self.dropout_1(attn_output)
        out_1 = self.layernorm_1(query + attn_output)
        ffn_output = self.ffn(out_1)
        ffn_output = self.dropout_2(ffn_output)
        out_2 = self.layernorm_2(out_1 + ffn_output)
        return out_2
```

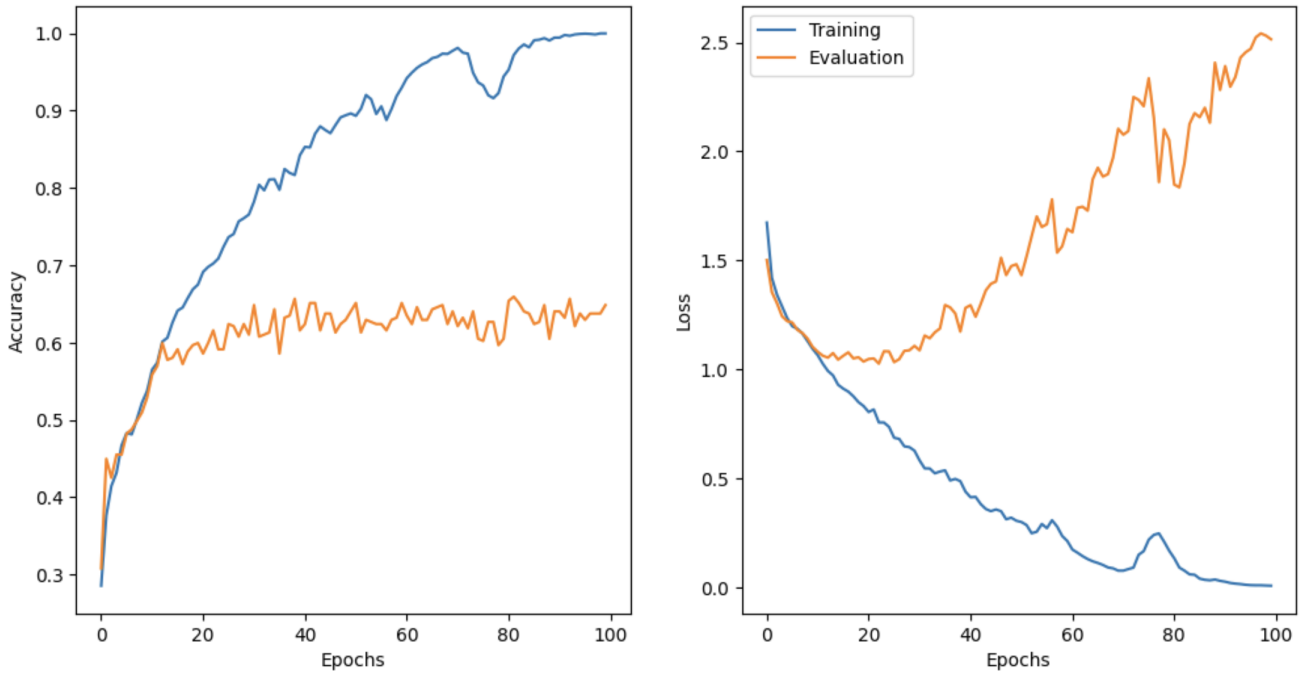
```
class PatchPositionEmbedding(nn.Module):
    def __init__(self, image_size=224, embed_dim=512, patch_size=16, device='cpu'):
        super().__init__()
        self.conv1 = nn.Conv2d(in_channels=3, out_channels=embed_dim, kernel_size=
    patch_size, stride=patch_size, bias=False)
        scale = embed_dim ** -0.5
        self.positional_embedding = nn.Parameter(scale * torch.randn((image_size //
    patch_size) ** 2, embed_dim))
        self.device = device

    def forward(self, x):
        x = self.conv1(x) # shape = [*, width, grid, grid]
        x = x.reshape(x.shape[0], x.shape[1], -1)  # shape = [*, width, grid ** 2]
        x = x.permute(0, 2, 1)  # shape = [*, grid ** 2, width]

        x = x + self.positional_embedding.to(self.device)
        return x
```

```
class VisionTransformerCls(nn.Module):
    def __init__(self,
                 image_size, embed_dim, num_heads, ff_dim,
                 dropout=0.1, device='cpu', num_classes = 10, patch_size=16
        ):
        super().__init__()
        self.embd_layer = PatchPositionEmbedding(
            image_size=image_size, embed_dim=embed_dim, patch_size=patch_size, device=
    device
        )
        self.transformer_layer = TransformerEncoder(
            embed_dim, num_heads, ff_dim, dropout
        )
        # self.pooling = nn.AvgPool1d(kernel_size=max_length)
        self.fc1 = nn.Linear(in_features=embed_dim, out_features=20)
        self.fc2 = nn.Linear(in_features=20, out_features=num_classes)
        self.dropout = nn.Dropout(p=dropout)
        self.relu = nn.ReLU()
    def forward(self, x):
        output = self.embd_layer(x)
        output = self.transformer_layer(output, output, output)
        output = output[:, 0, :]
        output = self.dropout(output)
        output = self.fc1(output)
        output = self.dropout(output)
        output = self.fc2(output)
        return output
```

4.2. Training

```
image_size=224
embed_dim = 512
num_heads = 4
ff_dim = 128
dropout=0.1

device = torch.device('cuda' if torch.cuda.is_available() else 'cpu')
model = VisionTransformerCls(
    image_size=224, embed_dim=512, num_heads=num_heads, ff_dim=ff_dim, dropout=dropout
    , num_classes=num_classes, device=device
)
model.to(device)

criterion = torch.nn.CrossEntropyLoss()
optimizer = optim.Adam(model.parameters(), lr=0.0005)

num_epochs = 100
save_model = './vit_flowers'
os.makedirs(save_model, exist_ok = True)
model_name = 'vit_flowers'

model, metrics = train(
    model, model_name, save_model, optimizer, criterion, train_loader, val_loader,
    num_epochs, device
)
```

Kết quả training và accuracy trên tập test là 60%



Hình 10: Quá trình huấn luyện bộ dữ liệu Flower với mô hình Vision Transformer.

5. Fine Tuning

5.1. Modeling

```
from transformers import ViTForImageClassification

id2label = {id:label for id, label in enumerate(classes)}
label2id = {label:id for id,label in id2label.items()}

model = ViTForImageClassification.from_pretrained('google/vit-base-patch16-224-in21k',
                                                  num_labels=num_classes,
                                                  id2label=id2label,
                                                  label2id=label2id)
device = torch.device('cuda' if torch.cuda.is_available() else 'cpu')
model.to(device)
```

5.2. Metric

```
import evaluate
import numpy as np

metric = evaluate.load("accuracy")

def compute_metrics(eval_pred):
    predictions, labels = eval_pred
    predictions = np.argmax(predictions, axis=1)
    return metric.compute(predictions=predictions, references=labels)
```

5.3. Trainer

```
import torch
from transformers import ViTImageProcessor
from transformers import TrainingArguments, Trainer

feature_extractor = ViTImageProcessor.from_pretrained("google/vit-base-patch16-224-
    in21k")

metric_name = "accuracy"

args = TrainingArguments(
    f"vit_flowers",
    save_strategy="epoch",
    evaluation_strategy="epoch",
    learning_rate=2e-5,
    per_device_train_batch_size=32,
    per_device_eval_batch_size=32,
    num_train_epochs=10,
    weight_decay=0.01,
    load_best_model_at_end=True,
    metric_for_best_model=metric_name,
    logging_dir='logs',
    remove_unused_columns=False,
)

def collate_fn(examples):
    # example => Tuple(image, label)
    pixel_values = torch.stack([example[0] for example in examples])
    labels = torch.tensor([example[1] for example in examples])
    return {"pixel_values": pixel_values, "labels": labels}

```

```
trainer = Trainer(
    model,
    args,
    train_dataset=train_dataset,
    eval_dataset=valid_dataset,
    data_collator=collate_fn,
    compute_metrics=compute_metrics,
    tokenizer=feature_extractor,
)
```

5.4. Training

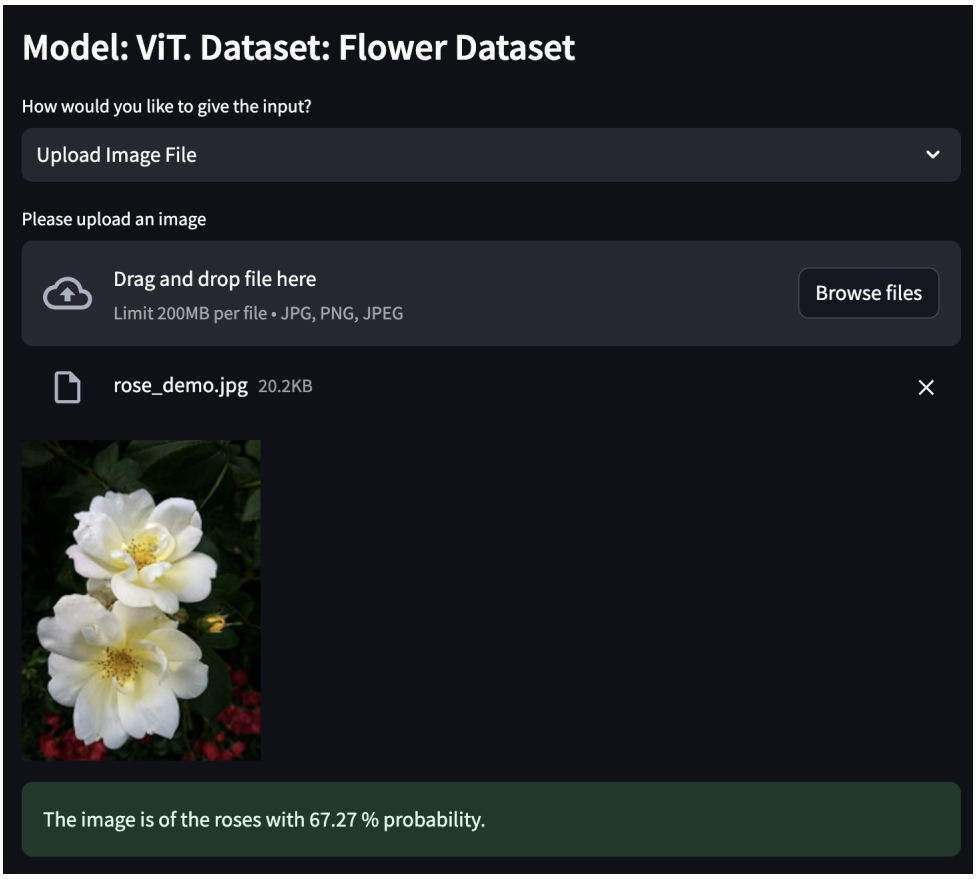
```
trainer.train()
outputs = trainer.predict(test_dataset)
outputs.metrics
```

Kết quả training và accuracy trên tập test là 97%

5.5. Triển khai mô hình

Dựa trên mô hình đã huấn luyện và thư viện streamlit để triển khai ứng dụng:

1. Link Streamlit
2. Github
3. Giao diện:



Hình 11: Giao diện ứng dụng.

Phần 4. Câu hỏi trắc nghiệm

Câu hỏi 1 Lớp Positional Encoding trong kiến trúc mô hình Transformer dùng để làm gì?
a) Biểu diễn vị trí của các tokens
b) Tính Attention
c) Dự đoán token tiếp theo
d) Lính Loss

Câu hỏi 2 Layer nào sau đây không có trong Transformer-Encoder?
a) Masked Multi-Head Attention
b) Multi-Head Attention
c) Layer Normaliation
d) Feed Forward

Câu hỏi 3 Layer nào sau đây không có trong Transformer-Decoder?
a) Masked Multi-Head Attention
b) Multi-Head Attention
c) CNN
d) Feed Forward

Câu hỏi 4 Masked Multi-Head Attention được sử dụng để làm gì?
a) Mask những token lịch sử
b) Mask vị trí các token lịch sử
c) Mask những token trong tương lai
d) Mask cả những token lịch sử và token trong tương lai

Câu hỏi 5 Phát biểu nào sau đây là đúng về BERT?
a) Bao gồm các khối Transformer-Encoder
b) Bao gồm các khối Transformer-Decoder
c) Cả 2 đáp án a và b đều đúng
d) Cả 2 đáp án a và b đều sai

Câu hỏi 6 Objective Function của BERT là?
a) Masked Language Model
b) Next Sentence Prediction
c) Cả a và b đều đúng
d) Cả a và b đều sai

Câu hỏi 7 Thành phần nào không có trong các giá trị Input của BERT
a) Token Embeddings
b) Seqment Embeddings

c) Position Embeddings
d) Language Model Head

Câu hỏi 8 Số lượng tham số của mô hình BERT-base và BERT-large là?
a) 340M và 110M
b) 110M và 110M
c) 340M và 340M
d) 110M và 340M

Câu hỏi 9 Các Patch trong Vision Transformer được xác định như thế nào?
a) Flatten ảnh đầu vào
b) Trung bình các giá trị điểm ảnh theo hàng
c) Trung bình các giá trị điểm ảnh theo cột
d) Chia ảnh đầu vào thành các Patch với kích thước các mảng cố định

Câu hỏi 10 Bộ dữ liệu nào sau đây được sử dụng trong phần thực nghiệm so sánh kết quả Vision Transformer với các phương pháp huấn luyện từ đầu và sử dụng pretrained?
a) CIFAR10
b) Flower
c) CIFAR100
d) MNIST

Phần 5. Phụ lục

1. **Hint**: Các file code gợi ý có thể được tải về tại đây:
   - Transformer-For-Text
   - Transformer-For-Image
2. **Solution**: Các file code cài đặt hoàn chỉnh và phần trả lời nội dung trắc nghiệm có thể được tải về tại đây (Lưu ý: Sáng thứ 3 khi hết deadline phần project, ad mới copy các nội dung bài giải nêu trên vào đường dẫn).
3. **Rubric**:

Phần	Kiến Thức	Đánh Giá
1	- Hiểu rõ kiến trúc mô hình Transformer - Hiểu rõ mô hình BERT - Áp dụng Transformer-Encoder và BERT cho bài toán phân loại văn bản	- Xây dựng mô hình phân loại văn bản sử dụng Transformer-Encoder và BERT
2.	- Hiểu rõ kiến trúc mô hình Vision Transformer	- Phân loại hình ảnh sử dụng mô hình Vision Transformer và các mô hình tiền huấn luyện cho Vision Transformer

- Hết -