```
import requests
from PIL import Image
from transformers import ViltForQuestionAnswering, ViltProcessor

from swarm_models.base_multimodal_model import BaseMultiModalModel


class Vilt(BaseMultiModalModel):
    """
    Vision-and-Language Transformer (ViLT) model fine-tuned on VQAv2.
    It was introduced in the paper ViLT: Vision-and-Language Transformer Without
    Convolution or Region Supervision by Kim et al. and first released in this repository.

    Disclaimer: The team releasing ViLT did not write a model card for this model
    so this model card has been written by the Hugging Face team.

    https://huggingface.co/dandelin/vilt-b32-finetuned-vqa


    Example:
        >>> model = Vilt()
        >>> output = model("What is this image",
"http://images.cocodataset.org/val2017/000000039769.jpg")

    """
```

```
def __init__(

    self,

    model_name: str = "dandelin/vilt-b32-finetuned-vqa",

    *args,

    **kwargs,

):

    super().__init__(model_name, *args, **kwargs)

    self.processor = ViltProcessor.from_pretrained(

        model_name, *args, **kwargs

    )

    self.model = ViltForQuestionAnswering.from_pretrained(

        model_name, *args, **kwargs

    )


def run(self, task: str = None, img: str = None, *args, **kwargs):

    """

    Run the model




    Args:


    """

    # Download the image

    image = Image.open(requests.get(img, stream=True).raw)


    encoding = self.processor(image, task, return_tensors="pt")
```

```
# Forward pass
outputs = self.model(**encoding)
logits = outputs.logits
idx = logits.argmax(-1).item()
print("Predicted Answer:", self.model.config.id2label[idx])
```