Phi-3-Vision on Apple Silicon:

MLX Porting Guide

Josef Albers

August 1, 2024

Abstract

This tutorial series details the process of porting Microsoft's Phi-3-Vision to Apple's MLX framework for efficient execution on Apple Silicon. The four-part guide covers: 1) basic model implementation, 2) integrating Su-scaled Rotary Position Embeddings for handling long sequences, 3) implementing batching for improved efficiency, and 4) adding caching to accelerate text generation. Aimed at AI enthusiasts and developers, this work demonstrates how to run advanced AI models efficiently on consumer-grade hardware, making cutting-edge AI more accessible on Apple devices.

Contents

1 Porting Phi-3-Vision to MLX: A Python Hobbyist's Journey

1.1 Introduction:

Welcome to an exciting series on optimizing cutting-edge AI models for Apple Silicon! Over the next few weeks, we'll dive deep into the process of porting Phi-3-Vision, a powerful and compact vision-language model, from Hugging Face to MLX.

This series is designed for AI enthusiasts, developers, and researchers interested in running advanced models efficiently on Mac devices. For those eager to get started, you can find the MLX ports of both Phi-3-Mini-128K and Phi-3-Vision in my GitHub repository: https://github.com/JosefAlbers/Phi-3-Vision-MLX

1.2 Why Phi-3-Vision?

When Microsoft Research released Phi-3, I was immediately intrigued. Despite its relatively small size of 3.8 billion parameters, it was performing on par with or even surpassing models with 7 billion parameters. This efficiency was impressive and hinted at the potential for running sophisticated AI models on consumer-grade hardware.

The subsequent release of Phi-3-Vision further piqued my interest. As an owner of a Mac Studio and a Python hobbyist, I saw an exciting opportunity to bring this capable vision-language model to Apple Silicon. While llama.cpp was a popular option for running large language models on Mac, its C++ codebase was way beyond my skill level, so I looked for a more accessible option. This led me to MLX, Apple's machine learning framework that not only offered a Python-friendly environment but also promised better performance than llama.cpp on Apple Silicon.

What made this journey even more exciting was that it marked my first foray into contributing to open source projects. As I worked on porting Phi-3-Vision, I found myself making my first pull requests to repositories like "mlx-examples" and "mlx-vlm". This experience was an invaluable learning experience that helped me gain a better understanding of the MLX framework and how to apply it to real-world projects. This experience also connected me with the broader AI development community.

1.3 Useful Resources:

Before we dive into the series, I want to highlight some excellent resources that have been invaluable in my journey:

1. **MLX Examples** (https://github.com/ml-explore/mlx-examples): This official repository from the MLX team at Apple is a treasure trove of examples and tutorials that showcase the capabilities of the MLX framework. With a wide range of standalone examples, from basic MNIST to advanced language models and image generation, this repository is an excellent starting point for anyone looking to learn MLX. The quality and depth of the examples are truly impressive, and they demonstrate the team's commitment to making MLX accessible to developers of all levels. I also want to give a special shoutout to awni, the repo owner, who was incredibly kind and patient with me when I made my first-ever pull request to this repository. Despite my lack of experience with Git and GitHub, awni guided me through the process and helped me navigate the precommit hooks and other nuances of the repository. Their patience and willingness to help a newcomer like me was truly appreciated, and I'm grateful for the opportunity to have contributed to this repository. If you're new to MLX or Git, I highly recommend checking out this repository and reaching out to awni - they're a great resource and a pleasure to work with!

2. **MLX-VLM** (https://github.com/Blaizzy/mlx-vlm): A package specifically for running Vision Language Models on Mac using MLX. This repository was particularly helpful in understanding how to handle multimodal inputs, and I found the well-organized and well-written code to be incredibly valuable in learning not only Vision Language Models (VLMs) but also the MLX framework in general. The codebase is a great example of how to structure and implement complex AI models using MLX, making it an excellent resource for anyone looking to learn from experienced developers and improve their own MLX skills. For those interested in other models, Prince Canuma has an excellent tutorial on running Google's Gemma 2 locally on Mac using MLX: https://www.youtube.com/watch?v=CKznaU1HpVQ

3. **Hugging Face** (https://huggingface.co/): A popular platform for natural language processing (NLP) and computer vision tasks, providing a vast range of pre-trained models, datasets, and tools. Hugging Face's Transformers library is particularly useful for working with transformer-based models like Phi-3-Vision. Their documentation and community support are also top-notch, making it an excellent resource for anyone looking to learn more about NLP and computer vision.

These resources provide a great foundation for anyone looking to explore MLX and run advanced AI models on Apple Silicon.

1.4 What to Expect in This Series:

1.4.1 MLX vs. Hugging Face: A Code Comparison

We'll start by comparing the original Hugging Face implementation with our MLX port, highlighting key differences in syntax and how MLX leverages Apple Silicon's unified memory architecture.

1.4.2 Implementing Su-RoPE for 128K Context

We'll explore the Surrogate Rotary Position Embedding (Su-RoPE) implementation that enables Phi-3-Vision to handle impressive 128K token contexts.

1.4.3 Optimizing Text Generation in MLX: From Batching to Advanced Techniques

Learn how to implement efficient batch text generation, a crucial feature for many real-world applications. We'll also cover custom KV-Cache implementation, streaming capabilities, and other text generation optimizations.

1.4.4 LoRA Fine-tuning and Evaluation on MLX

Discover how to perform Low-Rank Adaptation (LoRA) training, enabling efficient fine-tuning of Phi-3-Vision on custom datasets. We'll also develop comprehensive evaluation techniques to ensure our LoRA-adapted model meets or exceeds the original's performance.

1.4.5 Building a Versatile AI Agent

In our finale, we'll create a multi-modal AI agent showcasing Phi-3-Vision's capabilities in handling both text and visual inputs.

1.5 Why This Series Matters:

Phi-3-Vision represents a significant advancement in compact, high-performing vision-language models. By porting it to MLX, we're making it more accessible and efficient for a wide range of applications on Apple Silicon devices. This project demonstrates the potential of running advanced AI models on consumer-grade hardware, specifically Apple Silicon Macs.

1.5.1 Throughout this series, we'll highlight:

- Performance gains on Apple Silicon
- Challenges in porting and how to overcome them
- The process of contributing to open source AI projects
- Practical applications of the optimized model

1.5.2 Who This Series Is For:

- AI enthusiasts and hobbyists looking to dive deeper into model optimization
- Researchers exploring efficient AI on consumer hardware
- Mac users eager to leverage their devices for AI tasks
- Anyone curious about the intersection of AI and Apple Silicon
- Beginners interested in contributing to open source AI projects

1.6 Stay Tuned!

Our journey into optimizing Phi-3-Vision for MLX promises to be full of insights, challenges, and exciting breakthroughs. Whether you're a fellow hobbyist looking to run advanced AI models on your Mac or simply curious about the future of AI on consumer devices, this series has something for you.

Join me on this adventure in AI optimization, and let's unlock the full potential of Phi-3-Vision on Apple Silicon together!

2 Part 1: Basic Implementation of Phi-3-Vision in MLX

2.1 Introduction

Welcome to Part 1 of the tutorial series on porting Phi-3-Vision from PyTorch to Apple's MLX framework. Our goal is to create a minimal functional implementation of Phi-3-Vision in MLX through:

1. Analyzing the original PyTorch code
2. Translating core components to MLX
3. Building a basic MLX implementation
4. Loading and running the ported model

By the end of this tutorial, we will have a basic working model capable of generating text, setting the stage for further optimizations in subsequent parts of the series.

The full implementation of this tutorial is available at https://github.com/JosefAlbers/Phi-3-Vision-MLX/tree/main/assets/tutorial_1.py

2.2 Analyzing the Source Code

Our first task is to locate the source code for the original Phi-3-Vision:

1. Visit the Hugging Face model hub: https://huggingface.co/models
2. Search for "phi-3-vision"
3. Click on the model to access its repository
4. Look for a file named `modeling_phi3_v.py`

Now let's examine the code:

1. Scroll to the bottom of the file to find the top-level model class (`Phi3VForCausalLM` in our case)
2. Look for the `forward` method in this class
3. Trace the flow of data through the model by following method calls

Through this process, we can identify five key components of the model:

1. `Phi3VModel`: Main model
2. `Phi3DecoderLayer`: Decoder layers
3. `Phi3Attention`: Attention mechanism
4. `Phi3MLP`: Feed-forward network
5. `Phi3ImageEmbedding`: Image embedding

With these components identified, we're ready to begin the translation process to MLX.

2.3 MLX-Specific Considerations

A few differences between PyTorch and MLX to note before we begin the porting:

1. **Array Creation**: MLX doesn't require specifying device location.
2. **Lazy Computation**: Arrays in MLX are only materialized when `eval()` is called.
3. **Model Definition**: MLX uses `__call__` instead of `forward` for the model's forward pass.

2.4 Understanding the Model Structure

Let's now examine each key component of Phi-3-Vision, translating them to MLX as we go:

2.4.1 Top-Level Model: Phi3VForCausalLM

This class serves as the main entry point of the model. It encapsulates the core `Phi3VModel` and adds a language modeling head.

```
class Phi3VForCausalLM(nn.Module):
    # ...
    def __call__(self, input_ids, pixel_values=None, image_sizes=None):
        x = self.model(input_ids, pixel_values, image_sizes)
        return self.lm_head(x)
```

This top-level class serves two main functions:

1. **Encapsulating the core model**: It wraps the `Phi3VModel`, which produces contextualized representations of the input.
2. **Applying the language model head**: It uses a linear transformation to convert the contextualized representations into logits over the entire vocabulary, representing the model's predictions for the next token in the sequence.

2.4.2 Core Model: Phi3VModel

The Phi3VModel implements the main transformer architecture.

```
class Phi3VModel(nn.Module):
    # ...
    def __call__(self, input_ids, pixel_values, image_sizes):
        x = self.embed_tokens(input_ids)
        x = self.vision_embed_tokens(x, pixel_values, image_sizes)
        for l in self.layers:
            x = l(x)
        return self.norm(x)
```

This class processes inputs through four stages:

1. **Text Embedding**: Input tokens are converted to dense vector representations.
2. **Vision Embedding**: If present, image inputs are processed and integrated with the text embeddings.
3. **Transformer Layers**: The combined embeddings are then passed through a series of decoder layers.
4. **Normalization**: The output is normalized before being returned.

2.4.3 Image Embedding: Phi3ImageEmbedding

This component processes image inputs and integrates them with text embeddings.

```
class Phi3ImageEmbedding(nn.Module):
    # ...
    def __call__(self, txt_embeds, img_embeds, img_sizes, positions):
        # Process images with CLIP
        img_features = self.img_processor.vision_model(img_embeds)

        # Reshape and concatenate features
        global_features = self._process_global_features(img_features)
        local_features = self._process_local_features(img_features, img_sizes)
```

```
        # Apply additional projections
        x = mx.concatenate([local_features, global_features], axis=1)
        for layer in self.img_projection:
            x = layer(x)

        # Integrate with text embeddings
        txt_embeds = self._integrate_features(txt_embeds, x, positions)
        return txt_embeds
```

This class combines a CLIP (Contrastive Language-Image Pre-training) model with custom processing steps:

1. **CLIP Processing**: The model uses a pre-trained CLIP vision model to extract initial features from the input images.
2. **Additional Processing**: After CLIP processing, the model applies additional processing steps:
   - It reshapes and concatenates the features for both global and local (sub-image) representations.
   - It applies additional linear projections and non-linear activations (GELU) to further process these features.
3. **Integration with Text Embeddings**: Finally, the processed image features are integrated with the text embeddings at specific positions in the input sequence.

2.4.4 Decoder Layer: Phi3DecoderLayer

Each decoder layer is a fundamental building block of the transformer architecture.

```
class Phi3DecoderLayer(nn.Module):
    # ...
    def __call__(self, x):
        r = self.self_attn(self.input_layernorm(x))
        h = x + r
        r = self.mlp(self.post_attention_layernorm(h))
        return h + r
```

The decoder layer performs a series of operations to its input:

1. **Self-Attention**: This mechanism allows the model to weigh the importance of different parts of the input when processing each element, enabling it to capture long-range dependencies in the sequence.
2. **Feedforward Neural Network (MLP)**: This subnet processes each position independently, introducing non-linearity and increasing the model's capacity to learn complex functions.
3. **Residual Connections**: After both the self-attention and MLP operations, the input is added to the output. This technique helps in mitigating the vanishing gradient problem and allows for easier training of deep networks.
4. **Layer Normalization**: Applied before the self-attention and MLP operations, this normalizes the inputs to each sub-layer, stabilizing the learning process and allowing for deeper networks.

The combination of these components enables each layer to refine and enrich the representations passed through the model.

2.4.5 Attention Mechanism: Phi3Attention

The attention mechanism allows the model to weigh the importance of different parts of the input when processing each element.

```
class Phi3Attention(nn.Module):
    # ...
    def __call__(self, x):
        B, L, _ = x.shape
        qkv = self.qkv_proj(x)
        q, k, v = mx.split(qkv, self.chop, axis=-1)
        q = q.reshape(B, L, self.n_heads, -1).transpose(0, 2, 1, 3)
        k = k.reshape(B, L, self.n_kv_heads, -1).transpose(0, 2, 1, 3)
        v = v.reshape(B, L, self.n_kv_heads, -1).transpose(0, 2, 1, 3)
        mask = mx.triu(mx.full((x.shape[1], x.shape[1]), -mx.inf), k=1)
        w = (q * self.scale) @ k.transpose(0, 2, 3, 1)
        w += mask
        w = mx.softmax(w, axis=-1)
        o = w @ v
        o = o.transpose(0, 2, 1, 3).reshape(B, L, -1)
        return self.o_proj(o).astype(qkv.dtype)
```

Key aspects of this implementation:

1. **Projection and Splitting**: The input is first projected into query (q), key (k), and value (v) representations using a single linear projection (`qkv_proj`), which is then split.
2. **Multi-head Reshaping**: The q, k, and v tensors are reshaped to separate the heads and prepare for the attention computation.
3. **Attention Mask**: A causal mask is created to ensure that each position can only attend to previous positions.
4. **Scaled Dot-Product Attention**: The core attention computation is performed. Alternatively, you can use a faster, optimized version of this operation available in mlx.core.fast:

```
# This:
w = (q * self.scale) @ k.transpose(0, 1, 3, 2)
w += mask
w = mx.softmax(w, axis=-1)
o = w @ v

# Is equivalent to:
o = mx.fast.scaled_dot_product_attention(q,k,v,scale=self.scale,mask=mask)
```

5. **Output Projection**: The attention output is reshaped and projected back to the original dimensionality.

The attention mechanism supports both standard multi-head attention and grouped-query attention by allowing different numbers of heads for queries (`n_heads`) versus keys/values (`n_kv_heads`). In the current configuration, however, these are set to the same value (32), resulting in standard multi-head attention.

2.4.6 MLP Layer: Phi3MLP

The MLP layer applies non-linear transformations to the attention outputs.

```python
class Phi3MLP(nn.Module):
    # ...
    def __call__(self, x):
        x = self.gate_up_proj(x)
        gate, x = mx.split(x, 2, axis=-1)
        return self.down_proj(nn.silu(gate) * x)
```

This implements a gated feedforward network:

1. **Gated Architecture**:
    - The input is first projected into two separate spaces: one for the 'gate' and one for the 'values'.
    - This is achieved through a single linear projection followed by a split operation.
2. **Activation Function**:
    - The gate portion uses the SiLU (Sigmoid Linear Unit) activation, also known as the swish function.
    - SiLU is defined as f(x) = x * sigmoid(x), which has been shown to perform well in deep networks.
3. **Gating Mechanism**:
    - The activated gate is element-wise multiplied with the value portion.
    - This allows the network to dynamically control information flow, potentially helping with gradient flow and enabling more complex functions to be learned.
4. **Final Projection**:
    - The gated output is then projected back to the model's hidden size through a final linear layer.

This design combines the benefits of gating mechanisms (often seen in LSTMs and GRUs) with the simplicity and effectiveness of feedforward networks, potentially allowing for more expressive computations within each transformer layer.

2.5 Loading and Using the Model

Now that we've ported our model to MLX, let's load and use it for text generation.

First, we'll download the model configuration and weights from huggingface:

```python
model_path = snapshot_download('microsoft/Phi-3-vision-128k-instruct')
```

Next, we'll load the model configuration:

```python
with open(f"{model_path}/config.json", "r") as f:
    config = json.load(f)
model_config = SimpleNamespace(**config)
```

Now, let's load and "sanitize" the model weights:

```python
model_weight = [(k, v.transpose(0, 2, 3, 1) if "patch_embedding.weight" in k
 ↪  else v)
                for wf in glob.glob(f"{model_path}/*.safetensors")
                for k, v in mx.load(wf).items()]
```

The line `v.transpose(0, 2, 3, 1) if "patch_embedding.weight" in k else v` adapts the patch embedding weights to MLX's format by converting them from PyTorch's NCHW (batch, channel, height, width) to MLX's NHWC (batch, height, width, channel) format. This transposition, often called "weight sanitization", is necessary when porting the model from PyTorch to MLX.

With our configuration and weights ready, we can initialize and load our model:

```
model = Phi3VForCausalLM(model_config)
model.load_weights(model_weight)
mx.eval(model.parameters())
model.eval()
```

Now that our model is loaded, let's use it to generate some text. First, we'll load the pretrained processor:

```
processor =
↪  AutoProcessor.from_pretrained('microsoft/Phi-3-vision-128k-instruct',
↪  trust_remote_code=True)
```

Then, we'll process our input text and generate the first token:

```
inputs = processor('Hello world!', return_tensors='np')
input_ids = mx.array(inputs['input_ids'])
logits = model(input_ids)
token = mx.argmax(logits[:, -1, :], axis=-1)
list_tokens = token.tolist()
```

This code processes the input text "Hello world!" and generates the first token. We use the `AutoProcessor` to tokenize the input, then pass it through the model to get logits. The token with the highest probability is selected as the next token.

To generate more tokens, we can use a simple loop:

```
for i in range(5):
    input_ids = mx.concatenate([input_ids, token[None]], axis=-1)
    logits = model(input_ids)
    token = mx.argmax(logits[:, -1, :], axis=-1)
    list_tokens += token.tolist()
```

This loop generates five additional tokens by repeatedly feeding our model's output back as input.

```
print(processor.tokenizer.decode(list_tokens))
# Output: How are you doing?<|end|>
```

And there we have it!

We've successfully ported Phi-3-Vision to MLX, loaded the model, and generated text. While this implementation is basic, it demonstrates that our port is functional and capable of generating coherent text.

2.6 Limitations

Our minimal implementation works for short sequences, but you'll notice it starts producing gibberish with longer contexts. This is because we haven't yet implemented position encoding, which we'll address in the next part with RoPE (Rotary Position Embedding).

2.7 Conclusion:

We've successfully created a barebones implementation of Phi-3-Vision in MLX. While it's not yet fully functional, it provides a solid foundation for the optimizations we'll explore in upcoming tutorials.

In Part 2, we'll implement Su-scaled Rotary Position Embeddings (RoPE) to enhance our model's ability to handle long sequences.

3 Part 2: Implementing Su-scaled Rotary Position Embeddings (RoPE) for Phi-3-Vision

3.1 Introduction

Welcome to Part 2 of our Phi-3-Vision porting series. In Part 1, we've created a basic implementation of the model in MLX. However, we also noted that it struggles with longer sequences. Today, we'll address this limitation by implementing Su-scaled Rotary Position Embeddings (RoPE), which will significantly enhance our model's ability to handle long contexts of up to 128K tokens.

The full implementation of this tutorial is available at https://github.com/JosefAlbers/Phi-3-Vision-MLX/tree/main/assets/tutorial_2.py

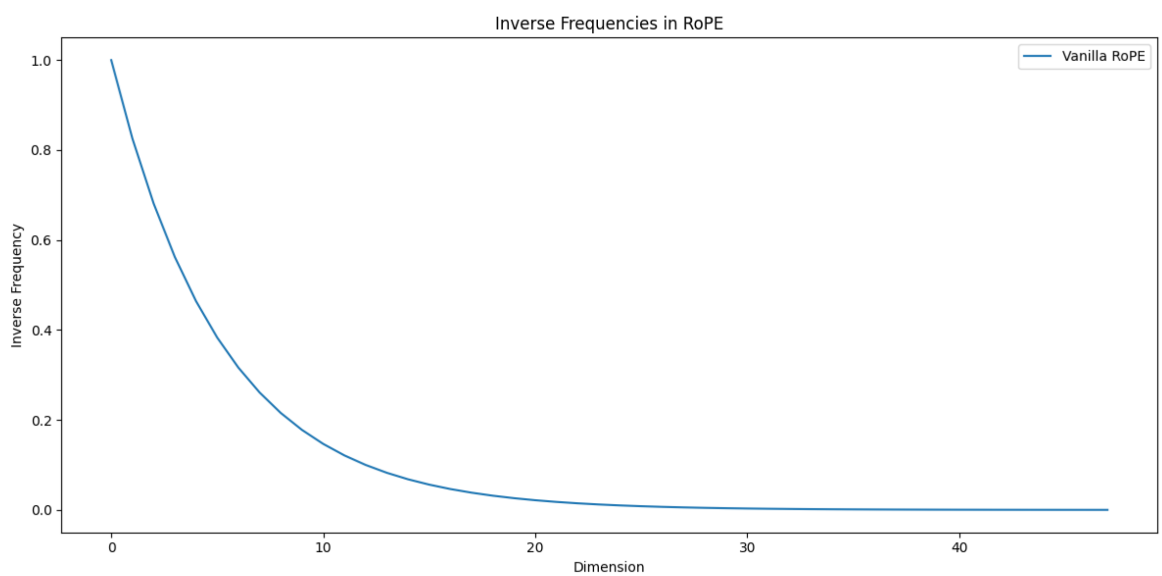
3.2 Understanding Rotary Position Embeddings (RoPE)

Before we delve into Su-scaled RoPE, let's first understand the basics of Rotary Position Embeddings.

RoPE is a technique that injects positional information into the model's token representations without adding extra tokens or increasing the model's parameter count. The key idea is to apply a rotation to each token's embedding based on its position in the sequence.

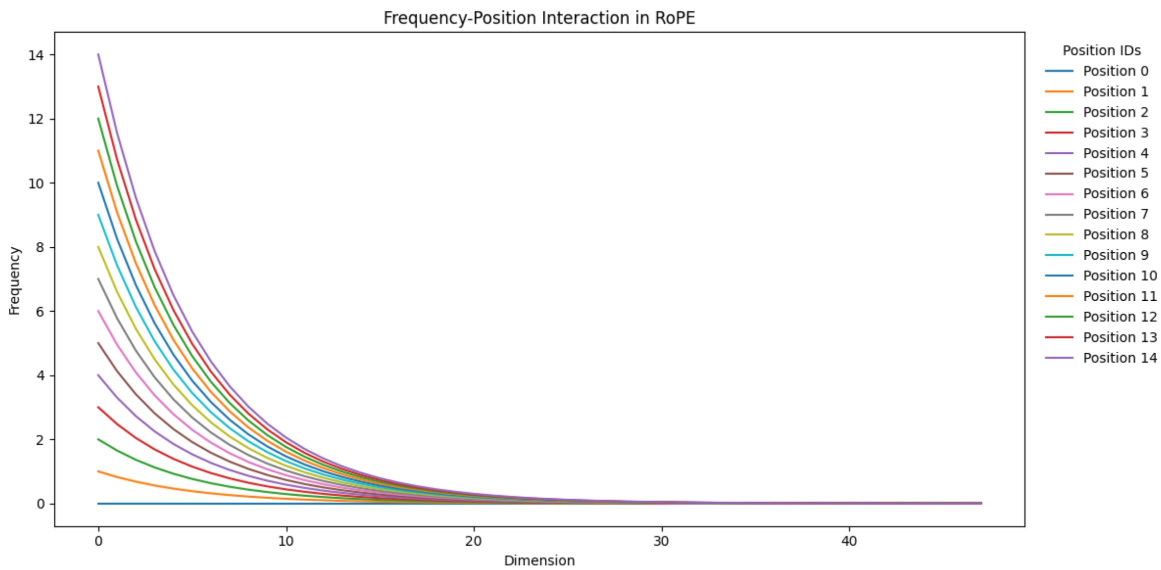
1. **Frequency Calculation**: For each dimension d in the embedding space, RoPE calculates a frequency:

```
inv_freq = 1 / (theta ** (d / dim))
```



2. **Position-Frequency Interaction**: These frequencies are then multiplied by the token positions to create unique sinusoidal patterns for each position.

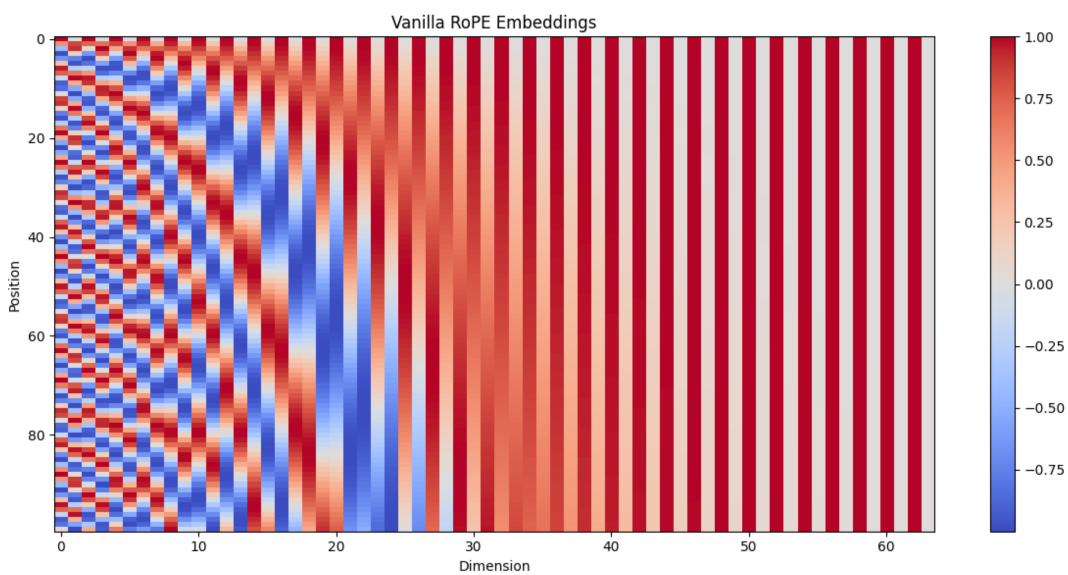
```
freqs = inv_freq @ position_ids.T
```



3. **Rotation Application**: The resulting patterns are used to rotate the token embeddings in 2D planes.

   For a token at position `pos`, RoPE applies the following rotation:

```
x_rotated = [x * cos(pos * freq) - y * sin(pos * freq),
             y * cos(pos * freq) + x * sin(pos * freq)]
```



Now that we understand RoPE, let's explore how Su-scaled RoPE builds upon and enhances this concept.

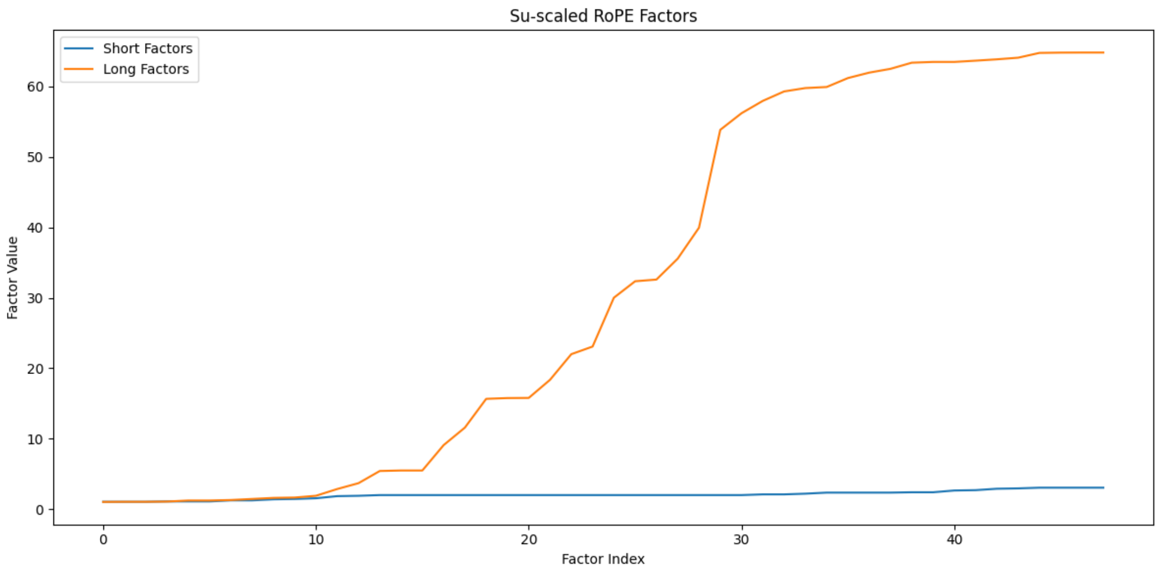
3.3 Understanding Su-RoPE

Su-RoPE extends RoPE by introducing scaling factors for different sequence length ranges.

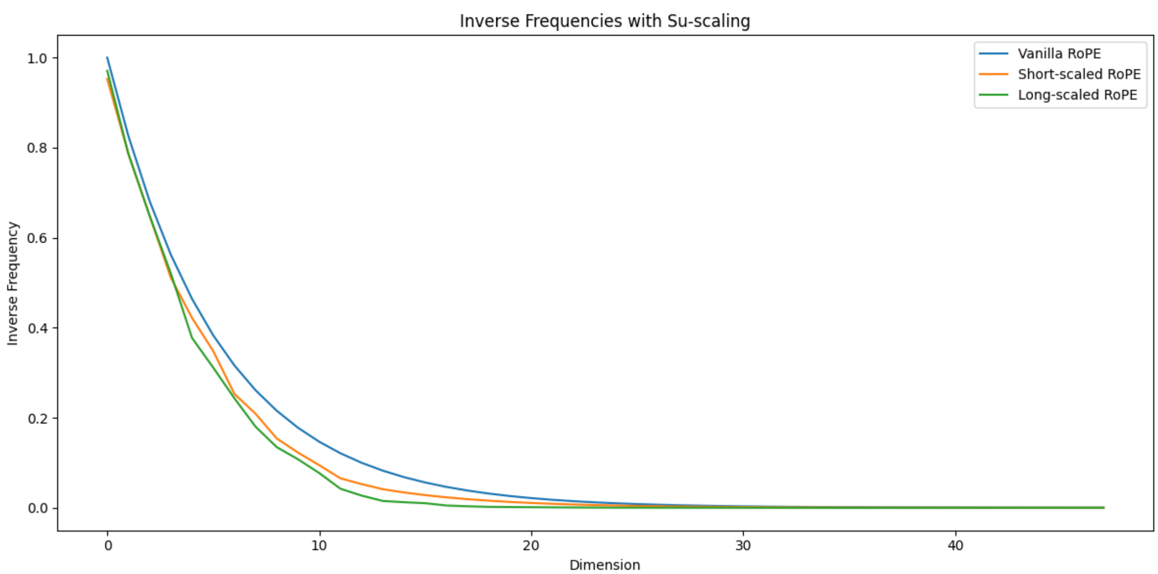
```
freq = 1 / (SU_FACTOR * theta ** (d / dim))
```

This allows the model to better generalize to sequences longer than those seen during training.

1. **Short and Long Factors**: Two sets of scaling factors are used, one for shorter sequences and one for longer sequences.



2. **Adaptive Scaling**: The choice between short and long factors is made based on the sequence length.



3. **Scaling Factor**: An additional scaling factor is applied to adjust for the extended maximum position embeddings.

3.4 Implementing Su-scaled RoPE

Now that we understand the theory behind Su-scaled RoPE, let's implement it in code. We'll create a `SuRoPE` class that encapsulates all the functionality we've discussed:

```
import mlx.core as mx
import mlx.nn as nn
import math

class SuRoPE:
    def __init__(self, config):
        self.dim = config.hidden_size // config.num_attention_heads
        self.original_max_position_embeddings =
         ↪ config.original_max_position_embeddings
        self.rope_theta = config.rope_theta
        self.scaling_factor = math.sqrt(1 +
         ↪ math.log(config.max_position_embeddings /
         ↪ config.original_max_position_embeddings) /
         ↪ math.log(config.original_max_position_embeddings))
        self.long_factor = config.rope_scaling["long_factor"]
        self.short_factor = config.rope_scaling["short_factor"]

    def __call__(self, q, k, position_ids=None):
        position_ids = mx.arange(q.shape[2], dtype=mx.float32)[None] if
 ↪  position_ids is None else position_ids
        cos, sin = self._get_cos_sin(position_ids)
        q = (q * cos) + (self._rotate_half(q) * sin)
        k = (k * cos) + (self._rotate_half(k) * sin)
        return q, k

    def _get_cos_sin(self, position_ids):
        su_factor = self.long_factor if mx.max(position_ids) >
 ↪  self.original_max_position_embeddings else self.short_factor
        position_ids_expanded = position_ids[:, None, :]
        inv_freq = 1.0 / (mx.array(su_factor, dtype=mx.float32) *
 ↪  self.rope_theta**(mx.arange(0, self.dim, 2, dtype=mx.float32) / self.dim))
        inv_freq_expanded = mx.repeat(inv_freq[None, :, None],
 ↪  position_ids.shape[0], axis=0)
        freqs = (inv_freq_expanded @ position_ids_expanded).transpose(0, 2, 1)
        emb = mx.concatenate([freqs, freqs], axis=-1)
        cos = mx.expand_dims(mx.cos(emb) * self.scaling_factor, axis=1)
        sin = mx.expand_dims(mx.sin(emb) * self.scaling_factor, axis=1)
        return cos, sin

    @staticmethod
    def _rotate_half(x):
        midpoint = x.shape[-1] // 2
        x1, x2 = x[..., :midpoint], x[..., midpoint:]
        return mx.concatenate([-x2, x1], axis=-1)
```

3.5 Integrating Su-scaled RoPE into Phi-3-Vision

Integrating our Su-scaled RoPE implementation into the Phi-3-Vision model is straightforward. We only need to add two lines to our `Phi3Attention` module:

```
class Phi3Attention(nn.Module):
    def __init__(self, config):
```

```
        # ...
        self.rope = SuRoPE(config)

    def __call__(self, x):
        # ...
        q, k = self.rope(q, k)
        # ...
```

These simple modifications allow our model to leverage Su-scaled RoPE, enabling it to handle sequences up to 128K tokens effectively.

3.6 Using the Updated Phi-3-Vision Model

Let's try an example that includes both text and an image:

```
from PIL import Image
import requests
prompt = f"<|user|>\n<|image_1|>\nWhat is shown in this
↪  image?<|end|>\n<|assistant|>\n"
images = [Image.open(requests.get("https://assets-
↪  c4akfrf5b4d3f4b7.z01.azurefd.net/assets/2024/04/BMDataViz_661fb89f3845e.png"
↪  , stream=True).raw)]
inputs = processor(prompt, images, return_tensors="np")
input_ids, pixel_values, image_sizes = [mx.array(inputs[i]) for i in
↪  ['input_ids', 'pixel_values', 'image_sizes']]
print(input_ids.shape)
# Output: (1, 1939)
```

Note that the input is translated into 1939 tokens.

Now, let's generate a response:

```
logits = model(input_ids, pixel_values, image_sizes)
token = mx.argmax(logits[:, -1, :], axis=-1)
list_tokens = token.tolist()
for i in range(50):
    input_ids = mx.concatenate([input_ids, token[None]], axis=-1)
    logits = model(input_ids)
    token = mx.argmax(logits[:, -1, :], axis=-1)
    list_tokens += token.tolist()
print(processor.tokenizer.decode(list_tokens))
# Output: The image displays a chart with a series of connected dots forming a
↪  line that trends upwards, indicating a positive correlation between two
↪  variables. The chart is labeled with 'X' on the horizontal axis and 'Y' on
↪  the vertical axis,
```

This example showcases the model's ability to process a long input sequence (1939 tokens from the image plus the text prompt) and generate a coherent response, demonstrating the effectiveness of our Su-scaled RoPE implementation.

3.7 Limitations

While our Su-scaled RoPE implementation enhances the model's capacity for long sequences, two key limitations remain:

1. **Single Input Processing**: The current implementation processes only one input at a time, limiting throughput for multiple queries.
2. **Inefficient Generation**: Our token-by-token generation without caching leads to unnecessary repeated computations, slowing down the process.

These issues will be addressed in upcoming tutorials, where we'll explore efficient batching and caching mechanisms to improve the model's speed and inefficiency.

3.8 Conclusion

In this tutorial, we implemented Su-scaled Rotary Position Embeddings (RoPE), enabling our model to handle sequences up to 128K tokens.

In Part 3, we'll explore batching techniques to further optimize our Phi-3-Vision implementation in MLX.

4 Part 3: Implementing Batching for Phi-3-Vision in MLX

4.1 Introduction

In this tutorial, we will explore how to implement batching for the Phi-3-Vision model in MLX. Batching enables the model to process multiple inputs in parallel, significantly enhancing computational efficiency and accelerating text generation.

The full implementation of this tutorial is available at https://github.com/JosefAlbers/Phi-3-Vision-MLX/tree/main/assets/tutorial_3.py

4.2 Understanding Batching

Batching is a technique that allows the model to process multiple inputs simultaneously. This approach is particularly advantageous for smaller large language models (sLLMs) like Phi-3, as it can massively speed up the text generation process.

4.3 Implementing Batching Utilities

To implement batching, we need to create utility functions that can handle padding, updating inputs, and generating attention masks.

4.3.1 Padding Function

The `pad_to_batch` function takes in a dictionary of inputs and returns a padded version of the inputs, along with the corresponding position IDs and attention masks.

```
def pad_to_batch(inputs):
    input_ids = [i.tolist() for i in inputs['input_ids']]
    max_length = max(len(sublist) for sublist in input_ids)
    return {
        'input_ids': mx.array([[0]*(max_length-len(sublist)) + sublist for
        ↪  sublist in input_ids]),
        'position_ids': mx.array([[1]*(max_length-len(sublist)) +
        ↪  list(range(len(sublist))) for sublist in input_ids]),
        'attention_mask': mx.array([[0]*(max_length-len(sublist)) +
        ↪  [1]*len(sublist) for sublist in input_ids]),
    }
```

This function pads the inputs to the same length, adjusts the position IDs, and creates attention masks. Note that we're padding on the left side to preserve the causal structure of the input sequence, as required by autoregressive models.

4.3.2 Input Update Function

The `update_inputs` function updates the inputs with newly generated tokens, maintaining the correct structure for position IDs and attention masks.

```
def update_inputs(inputs, token):
    input_ids, position_ids, attention_mask = inputs['input_ids'],
↪   inputs['position_ids'], inputs['attention_mask']
    return {
```

```
        'input_ids': mx.concatenate([input_ids, token[:,None]], axis=-1),
        'position_ids': mx.concatenate([position_ids, position_ids[:, -1:] +
         ↪  1], axis=1),
        'attention_mask': mx.concatenate([attention_mask,
         ↪  mx.ones((input_ids.shape[0], 1), dtype=attention_mask.dtype)],
         ↪  axis=1),
    }
```

This function updates our inputs with newly generated tokens, maintaining the correct structure for position IDs and attention masks.

4.4 Modifying the Model for Batched Inputs

To enable batching, we need to update our model to use the `position_ids` and `attention_mask`.

4.4.1 Updating the Model Interface

We modify the top-level `Phi3VForCausalLM` class to accept the batched inputs and pass them to its model.

```
class Phi3VForCausalLM(nn.Module):
    # ...
    def __call__(self, input_ids, pixel_values=None, image_sizes=None,
     ↪  position_ids=None, attention_mask=None):
        x = self.model(input_ids, pixel_values, image_sizes, position_ids,
 ↪  attention_mask)
        return self.lm_head(x)
```

4.4.2 Updating the Phi3VModel

Next, modify the `Phi3VModel` to pass `position_ids` and `attention_mask` to each layer:

```
class Phi3VModel(nn.Module):
    # ...
    def __call__(self, input_ids, pixel_values, image_sizes, position_ids,
     ↪  attention_mask):
        x = self.embed_tokens(input_ids)
        x = self.vision_embed_tokens(x, pixel_values, image_sizes)
        for l in self.layers:
            x = l(x, position_ids, attention_mask)
        return self.norm(x)
```

4.4.3 Updating the Attention Mechanism

Finally, update the Phi3Attention module to utilize `position_ids` and `attention_mask`:

```
class Phi3Attention(nn.Module):
    # ...
    def __call__(self, x, position_ids, attention_mask):
        # ...
        q, k = self.rope(q, k, position_ids)
```

```
        mask = mx.triu(mx.full((v.shape[2], v.shape[2]), -mx.inf), k=1)
        if attention_mask is not None:
            mask += mx.where(attention_mask[:, :, None]*attention_mask[:, None, :]==1, 0, -mx.inf)
            mask = mx.expand_dims(mask, 1)
        # ...
```

4.5 Using Batched Inputs

Here's an example of batched text generation:

```
# Prepare batched inputs
inputs = processor(['Hello World!', 'Guten Tag!'], return_tensors='np')
inputs = pad_to_batch(inputs)

# Generate tokens
logits = model(**inputs)
token = mx.argmax(logits[:, -1, :], axis=-1)
list_tokens = [token]
for i in range(5):
    inputs = update_inputs(inputs, token)
    logits = model(**inputs)
    token = mx.argmax(logits[:, -1, :], axis=-1)
    list_tokens.append(token)
list_tokens = mx.stack(list_tokens, axis=1).tolist()
print(processor.tokenizer.batch_decode(list_tokens))
# Output: ['How are you doing today?', 'Was möchten Sie w']
```

4.6 Conclusion

By implementing custom batching for Phi-3-Vision, we've enabled our model to efficiently handle multiple inputs while ensuring correct behavior for autoregressive generation. This approach provides fine-grained control over input processing, position IDs, and attention masks, which is crucial for optimal model performance.

In the next part, we'll explore implementing efficient caching mechanisms to further accelerate text generation, especially for longer sequences.

5 Part 4: Implementing Caching for Phi-3-Vision in MLX

5.1 Introduction

In this tutorial, we'll implement caching for our Phi-3-Vision model in MLX.

The full implementation of this tutorial is available at https://github.com/JosefAlbers/Phi-3-Vision-MLX/tree/main/assets/tutorial_4.py

5.2 The Need for Caching

Our previous implementation of the Phi-3-Vision model processes the entire input sequence from scratch for each new token. This approach becomes inefficient as the sequence grows:

```
Without Caching:

Iteration 1: [Prompt] -> Model -> Token 1
Iteration 2: [Prompt, Token 1] -> Model -> Token 2
Iteration 3: [Prompt, Token 1, Token 2] -> Model -> Token 3
...
```

This repetitive processing leads to unnecessary computations.

5.3 How Caching Helps

Caching solves this problem by storing and reusing intermediate computations from previous iterations:

```
With Caching:

Iteration 1: [Prompt] -> Model -> Token 1, Cache
Iteration 2: Cache + [Token 1] -> Model -> Token 2, Cache
Iteration 3: Cache + [Token 2] -> Model -> Token 3, Cache
...
```

Instead of processing the entire sequence each time, the model processes only the new token and uses the cached information for the rest.

5.4 Implementing Caching

To implement caching, we need to modify the attention mechanism and the model layers to handle the cache.

5.4.1 Modifying the Attention Mechanism

We modify the attention mechanism to handle both cached and non-cached scenarios. We add a cache parameter to the `__call__` method, which is used to store and retrieve the cached values:

```python
class Phi3Attention(nn.Module):
    # ...
    def __call__(self, x, position_ids, attention_mask, cache):
        # ...
        if cache is None:
            position_ids = mx.arange(q.shape[2], dtype=mx.float32)[None] if
 ↪  position_ids is None else position_ids
```

```
            q, k = self.rope(q, k, position_ids)
            mask = mx.triu(mx.full((v.shape[2], v.shape[2]), -mx.inf), k=1)
            if attention_mask is not None:
                mask += mx.where(attention_mask[:, :, None]*attention_mask[:,
↪   None, :]==1, 0, -mx.inf)
                mask = mx.expand_dims(mask, 1)
        else:
            past_k, past_v, past_p, past_m = cache
            position_ids = past_p[:,-1:]+1
            mask = mx.pad(past_m[:,:,-1:,:], ((0,0),(0,0),(0,0),(0,1)))
            q, k = self.rope(q, k, position_ids)
            k = mx.concatenate([past_k, k], axis=2)
            v = mx.concatenate([past_v, v], axis=2)

        cache = (k, v, position_ids, mask)
        # ...
        return self.o_proj(o).astype(qkv.dtype), cache
```

This modification allows the attention mechanism to either compute from scratch or use and update the cache, depending on whether a cache is provided.

5.4.2 Updating the Model Layers

Next, we update the model layers to handle the cache by adding a cache parameter to the `__call__` method and passing it through each layer.

```
class Phi3DecoderLayer(nn.Module):
    # ...
    def __call__(self, x, position_ids, attention_mask, cache):
        r, cache = self.self_attn(self.input_layernorm(x), position_ids,
↪   attention_mask, cache)
        h = x + r
        r = self.mlp(self.post_attention_layernorm(h))
        return h + r, cache

class Phi3VModel(nn.Module):
    # ...
    def __call__(self, input_ids, pixel_values, image_sizes, position_ids,
    ↪   attention_mask, cache):
        x = self.embed_tokens(input_ids)
        x = self.vision_embed_tokens(x, pixel_values, image_sizes)
        cache = [None]*len(self.layers) if cache is None else cache
        for i, l in enumerate(self.layers):
            x, cache[i] = l(x, position_ids, attention_mask, cache[i])
        return self.norm(x), cache

class Phi3VForCausalLM(nn.Module):
    # ...
    def __call__(self, input_ids, pixel_values=None, image_sizes=None,
    ↪   position_ids=None, attention_mask=None, cache=None):
        x, cache = self.model(input_ids, pixel_values, image_sizes,
↪   position_ids, attention_mask, cache)
```

```
        return self.lm_head(x), cache
```

5.5 Using Caching

Here's an example use of caching in text generation:

```
# Initial input processing
inputs = processor('Hello world!', return_tensors='np')
input_ids = mx.array(inputs['input_ids'])

# Initial forward pass
logits, cache = model(input_ids)
token = mx.argmax(logits[:, -1, :], axis=-1)
list_tokens = token.tolist()

# Generate additional tokens using cache
for i in range(5):
    logits, cache = model(token[:,None], cache=cache)
    token = mx.argmax(logits[:, -1, :], axis=-1)
    list_tokens += token.tolist()

print(processor.tokenizer.decode(list_tokens))
```

In this example, we first process the initial input and obtain the cache. Then, for each subsequent token generation, we use and update this cache, significantly reducing computation time for longer sequences.

5.6 Conclusion

By implementing caching in our Phi-3-Vision model, we've significantly improved its efficiency for token generation, especially for longer sequences. This optimization is important for practical applications of large language models, enabling faster and more efficient text generation.

In the next part of our series, we'll explore techniques for fine-tuning our model on custom datasets, allowing us to adapt Phi-3-Vision for specific tasks or domains. Stay tuned!

6 Addendum: Extending MLX Porting Techniques to PaliGemma

6.1 Introduction

In our previous tutorials, we explored porting Phi-3-Vision to MLX. Now, let's extend these techniques to PaliGemma, a powerful open Vision-Language Model (VLM) developed by Google. PaliGemma combines the SigLIP-So400m vision encoder with the Gemma-2B language model, creating a versatile and knowledgeable base model for various tasks.

The full implementation of this tutorial is available at https://github.com/JosefAlbers/Phi-3-Vision-MLX/blob/main/assets/paligemma_dissected.py

6.2 Key Differences Relevant for Porting PaliGemma

1. Dual-Model Architecture:
   - PaliGemma uses separate vision (SigLIP-So400m) and language (Gemma-2B) models
   - We'll need to implement both components and their integration in MLX
2. Multimodal Projection:
   - A linear adapter connects the vision and language models
   - This needs to be implemented to properly combine visual and textual features
3. Full Block Attention:
   - PaliGemma processes both image tokens and text tokens in a single attention mechanism
   - Our MLX implementation must handle this combined attention approach
4. Input Processing:
   - Visual tokens are prepended to text input
   - We need to implement this specific input preparation in MLX
5. Multiple Resolutions:
   - Support for different input sizes (224x224, 448x448, 896x896)
   - Our implementation should be flexible to handle these variants
6. Task Prefixes:
   - PaliGemma uses task-specific prefixes for conditioning
   - We'll need to incorporate this feature in our input processing and model logic

These key differences will guide our porting process, ensuring we accurately translate PaliGemma's architecture to MLX while optimizing for Apple Silicon.

6.3 Porting PaliGemma to MLX

6.3.1 Core Components

Let's start by implementing the main components of PaliGemma:

```
class PGemmaModel(nn.Module):
    def __init__(self, config):
        super().__init__()
        self.vision_tower = VisionModel(config.vision_config)
        self.language_model = LanguageModel(config.text_config)
        self.multi_modal_projector = Projector(config)

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

```
        super().__init__()
        self.embeddings = VisionEmbeddings(config)
        self.layers = [EncoderLayer(config) for _ in
         ↪ range(config.num_hidden_layers)]
        self.post_layernorm = nn.LayerNorm(config.hidden_size)

    def __call__(self, x):
        x = self.embeddings(x)
        for l in self.layers:
            x = l(x)
        return self.post_layernorm(x[0])

class LanguageModel(nn.Module):
    def __init__(self, config):
        super().__init__()
        self.scale = config.hidden_size**0.5
        self.embed_tokens = nn.Embedding(config.vocab_size, config.hidden_size)
        self.layers = [TransformerBlock(config=config) for _ in
         ↪ range(config.num_hidden_layers)]
        self.norm = RMSNorm(config)

    def __call__(self, input_ids, inputs_embeds=None, attention_mask_4d=None,
     ↪ cache=None):
        cache = [None] * len(self.layers) if cache is None else cache
        h = self.embed_tokens(input_ids) if inputs_embeds is None else
 ↪ inputs_embeds
        h = h * self.scale
        for e, layer in enumerate(self.layers):
            h, cache[e] = layer(h, attention_mask_4d, cache[e])
        return self.embed_tokens.as_linear(self.norm(h)), cache
```

6.3.2 Attention Mechanism

PaliGemma uses a different attention mechanism. Here's how we can implement it:

```
class Attention(nn.Module):
    def __init__(self, config):
        super().__init__()
        dims = config.hidden_size
        self.n_heads = n_heads = config.num_attention_heads
        head_dim = dims // n_heads
        self.scale = head_dim**-0.5
        self.q_proj = nn.Linear(dims, n_heads * head_dim, bias=config.attn_bias)
        self.k_proj = nn.Linear(dims, n_heads * head_dim, bias=config.attn_bias)
        self.v_proj = nn.Linear(dims, n_heads * head_dim, bias=config.attn_bias)
        self.o_proj = nn.Linear(n_heads * head_dim, dims, bias=config.attn_bias)
        if getattr(config, 'rope_base', False):
            self.rope = nn.RoPE(head_dim, base = config.rope_base)
        else:
            self.rope = lambda x, *args, **kwargs: x

    def __call__(self, x, mask=None, cache = None):
```

```
        B, L, _ = x.shape
        queries = self.q_proj(x).reshape(B, L, self.n_heads, -1).transpose(0,
↪   2, 1, 3)
        keys = self.k_proj(x).reshape(B, L, self.n_heads, -1).transpose(0, 2,
↪   1, 3)
        values = self.v_proj(x).reshape(B, L, self.n_heads, -1).transpose(0, 2,
↪   1, 3)

        queries = self.rope(queries)
        keys = self.rope(keys)

        if cache is not None:
            key_cache, value_cache = cache
            keys = mx.concatenate([key_cache, keys], axis=2)
            values = mx.concatenate([value_cache, values], axis=2)

        output = mx.fast.scaled_dot_product_attention(queries, keys, values,
↪   scale=self.scale, mask=mask)
        output = output.transpose(0, 2, 1, 3).reshape(B, L, -1)
        return self.o_proj(output), (keys, values)
```

6.3.3 Image Processing

PaliGemma handles image processing differently. Here's how we can implement it:

```
class VisionEmbeddings(nn.Module):
    def __init__(self, config):
        super().__init__()
        self.patch_embedding = nn.Conv2d(in_channels=config.num_channels,
         ↪  out_channels=config.hidden_size, kernel_size=config.patch_size,
         ↪  stride=config.patch_size)
        self.num_patches = (config.image_size // config.patch_size) ** 2
        self.position_embedding = nn.Embedding(self.num_patches,
         ↪  config.hidden_size)

    def __call__(self, x: mx.array) -> mx.array:
        return mx.flatten(self.patch_embedding(x), start_axis=1, end_axis=2) +
         ↪  self.position_embedding(mx.arange(self.num_patches)[None, :])
```

6.3.4 Assembling Inputs

PaliGemma requires a specific way of assembling inputs:

```
def assemble(input_ids, inputs_embeds, image_features, attention_mask, config):
    inputs_embeds, image_features, attention_mask = [mx.array(i) for i in
↪   (inputs_embeds, image_features, attention_mask)]
    final_embedding = mx.zeros_like(inputs_embeds)
    text_mask = (input_ids != config.image_token_index) & (input_ids !=
↪   config.pad_token_id)
    text_mask_expanded = mx.repeat(mx.expand_dims(text_mask, -1),
↪   final_embedding.shape[-1], axis=-1)
```

```
    final_embedding = mx.where(text_mask_expanded, inputs_embeds,
↪   final_embedding)
    image_mask = input_ids == config.image_token_index
    image_mask_expanded = mx.repeat(mx.expand_dims(image_mask, -1),
↪   final_embedding.shape[-1], axis=-1)
    final_embedding = mx.where(image_mask_expanded, mx.pad(image_features,
↪   ((0,0), (0,input_ids.shape[1] - image_features.shape[1]), (0,0))),
↪   final_embedding)
    attention_mask_expanded = mx.expand_dims(attention_mask, (1, 2))
    final_attention_mask_4d = attention_mask_expanded *
↪   attention_mask_expanded.transpose(0, 1, 3, 2)
    mx.repeat(final_attention_mask_4d, config.text_config.num_key_value_heads,
↪   axis=1)
    return mx.array(final_embedding), mx.array(final_attention_mask_4d)
```

6.4 Using the Ported PaliGemma Model

Here's how we can use our ported PaliGemma model:

```
# Load model components
processor, language_model, vision_model, projector, config = load_parts()

# Process input
image_url = "https://huggingface.co/datasets/huggingface/documentation-
↪   images/resolve/main/transformers/tasks/car.jpg"
image = Image.open(requests.get(image_url, stream=True).raw)
processed = processor('Caption: ', image, return_tensors="np")
input_ids, pixel_values, attention_mask = [mx.array(processed[key]) for key in
↪   ["input_ids", "pixel_values", "attention_mask"]]

# Prepare inputs
inputs_embeds = language_model.embed_tokens(input_ids)
hidden_state = vision_model(pixel_values.transpose(0, 2, 3, 1))
image_features = projector(hidden_state[None]) / (config.hidden_size**0.5)
inputs_embeds, attention_mask_4d = assemble(input_ids, inputs_embeds,
↪   image_features, attention_mask, config)

# Generate output
logits, cache = language_model(input_ids, inputs_embeds, attention_mask_4d,
↪   None)
token = mx.argmax(logits[:, -1, :], axis=-1)
list_tokens = token.tolist()
for _ in range(100):
    logits, cache = language_model(token[None], None, None, cache)
    token = mx.argmax(logits[:, -1, :], axis=-1)
    list_tokens += token.tolist()
    if list_tokens[-1] == processor.tokenizer.eos_token_id:
        break

print(processor.tokenizer.decode(list_tokens))
```

6.5 Conclusion

This addendum demonstrates how the techniques we learned for porting Phi-3-Vision to MLX can be extended to other models like PaliGemma. While the specific implementations differ, the core principles of translating model architecture, attention mechanisms, and input processing remain the same. This flexibility allows us to adapt a wide range of models to run efficiently on Apple Silicon using MLX.