# Explicación de Hiperparámetros de Vision Transformer (ViT)

## Configuración

```
config = {
    "patch_size": 4,  # Input image size: 32x32 -> 8x8 patches
    "hidden_size": 48,
    "num_hidden_layers": 4,
    "num_attention_heads": 4,
    "intermediate_size": 4 * 48,  # 4 * hidden_size
    "hidden_dropout_prob": 0.0,
    "attention_probs_dropout_prob": 0.0,
    "initializer_range": 0.02,
    "image_size": 32,
    "num_classes": 10,  # num_classes of CIFAR10
    "num_channels": 3,
    "qkv_bias": True,
    "use_faster_attention": True,
}
```

## Descripción

1. patch_size: 4 Descripción: Define el tamaño de los parches extraídos de la imagen de entrada. Con un tamaño de parche de 4, una imagen de 32x32 se dividirá en 8x8 parches. Impacto: Parches más pequeños capturan más detalles, pero aumentan la complejidad computacional.
2. hidden_size: 48 Descripción: Tamaño de las proyecciones lineales de cada parche. Controla la dimensionalidad de los vectores de características. Impacto: Aumentar este valor incrementa la capacidad del modelo para aprender patrones complejos, pero también el costo computacional.
3. num_hidden_layers: 4 Descripción: Número de capas Transformer en el modelo. Impacto: Más capas aumentan la profundidad del modelo y su capacidad para aprender, pero pueden hacer que el modelo sea más difícil de entrenar y más propenso a sobreajustarse.
4. num_attention_heads: 4 Descripción: Número de cabezas en el mecanismo de atención multi-cabeza. Impacto: Más cabezas permiten al modelo enfocarse en diferentes partes de la entrada, pero aumentan el costo computacional.
5. intermediate_size: 192 Descripción: Tamaño intermedio de la red feed-forward en cada capa Transformer, definido como 4 * hidden_size. Impacto: Un tamaño mayor permite aprender relaciones más complejas, pero aumenta el costo computacional.
6. hidden_dropout_prob: 0.0 Descripción: Probabilidad de aplicar dropout a las activaciones. Impacto: Dropout ayuda a evitar el sobreajuste, pero un valor de 0 significa que no se aplica esta regularización.

7. attention_probs_dropout_prob: 0.0 Descripción: Probabilidad de aplicar dropout a las probabilidades de atención. Impacto: Regulariza el mecanismo de atención para prevenir sobreajuste.
8. initializer_range: 0.02 Descripción: Controla el rango de inicialización aleatoria de los pesos. Impacto: Valores altos pueden causar inestabilidad en el entrenamiento, mientras que valores bajos pueden hacer que el modelo aprenda lentamente.
9. image_size: 32 Descripción: Tamaño de las imágenes de entrada. Impacto: Este valor debe coincidir con el tamaño del conjunto de datos de entrada.
10. num_classes: 10 Descripción: Número de clases en la tarea de clasificación (10 clases en CIFAR-10). Impacto: Debe ajustarse según el número de clases en tu tarea.
11. num_channels: 3 Descripción: Número de canales en las imágenes de entrada. Para imágenes en color, este valor es 3 (RGB). Impacto: Para imágenes en escala de grises, este valor sería 1.
12. qkv_bias: True Descripción: Define si se usa un sesgo en las capas que calculan las matrices Q, K, y V del mecanismo de atención. Impacto: Añadir sesgo puede ayudar a aprender relaciones más complejas, pero introduce más parámetros.
13. use_faster_attention: True Descripción: Habilita una implementación optimizada del mecanismo de atención. Impacto: Mejora la velocidad de entrenamiento e inferencia sin afectar la precisión del modelo.