Program Studi Teknik Informatika
Institut Teknologi Sumatera

Nama: **Bintang Fikri Fauzan (122140008)**
Mata Kuliah: **Deep Learning (IF-40401)**
Tugas Ke: **Eksplorasi Vision Transformer**
Tanggal: 21 November 2025

PERBANDINGAN MODEL VISION TRANSFORMER: SWIN TRANSFORMER TINY vs DeiT SMALL

Studi Kasus: Klasifikasi Gambar pada Dataset STL-10

GitHub Repository:
https://github.com/bintangfikrif/VisionTransformer-Comparison

1 PENDAHULUAN

1.1 Latar Belakang

Vision Transformer (ViT) telah merevolusi bidang computer vision sejak diperkenalkan oleh Dosovitskiy et al. pada tahun 2020 [1]. Berbeda dengan arsitektur Convolutional Neural Network (CNN) tradisional yang mendominasi selama bertahun-tahun, Vision Transformer mengadaptasi mekanisme self-attention dari domain Natural Language Processing (NLP) untuk pemrosesan gambar. Pendekatan ini memungkinkan model untuk menangkap hubungan global antar patch gambar secara lebih efektif dibandingkan dengan receptive field lokal pada CNN.

Keberhasilan ViT memicu perkembangan berbagai varian arsitektur transformer untuk visi, masing-masing dengan inovasi unik untuk mengatasi keterbatasan model original. Swin Transformer memperkenalkan hierarki multi-scale dengan shifted window attention untuk efisiensi komputasi [2]. DeiT (Data-efficient Image Transformer) fokus pada strategi training yang lebih efisien menggunakan knowledge distillation, memungkinkan model mencapai performa tinggi tanpa dataset pre-training yang sangat besar [3].

Pemilihan arsitektur yang tepat sangat krusial dalam aplikasi praktis, mengingat trade-off antara akurasi, kompleksitas komputasi, dan waktu inferensi. Penelitian komparatif ini diperlukan untuk memberikan panduan empiris dalam memilih model yang sesuai dengan kebutuhan spesifik, baik untuk aplikasi yang memprioritaskan akurasi maksimal maupun efisiensi komputasi.

1.2 Motivasi Perbandingan Model

Motivasi utama eksperimen ini adalah:

1. **Efisiensi vs Akurasi**: Swin Transformer menggunakan pendekatan hierarkis dengan local attention, sementara DeiT menggunakan global attention dengan distillation. Perbandingan ini akan mengungkap trade-off antara keduanya.

2. **Skalabilitas**: Memahami bagaimana kedua model berskala pada dataset berukuran medium (STL-10) dengan keterbatasan data dibandingkan ImageNet.

3. **Praktikalitas**: Mengevaluasi waktu inferensi dan kompleksitas parameter untuk aplikasi real-world, terutama pada resource-constrained environments.

4. **Transfer Learning**: Menganalisis efektivitas pre-trained weights dari ImageNet-1k saat di-fine-tune pada domain yang berbeda (natural images di STL-10).

1.3 Tujuan Eksperimen

Tujuan spesifik dari eksperimen ini adalah:

1. Mengimplementasikan dan fine-tune dua model Vision Transformer yang berbeda: Swin Transformer Tiny dan DeiT Small Distilled.

2. Membandingkan performa kedua model berdasarkan metrik akurasi, precision, recall, dan F1-score pada dataset STL-10.

3. Menganalisis kompleksitas model dari segi jumlah parameter dan ukuran model.

2 LANDASAN TEORI

2.1 Transformer dan Self-Attention Mechanism

Transformer adalah arsitektur neural network yang diperkenalkan oleh Vaswani et al. (2017) dalam paper "Attention is All You Need" [4]. Komponen inti dari transformer adalah mekanisme self-attention yang memungkinkan model untuk menimbang kepentingan relatif dari berbagai bagian input saat memproses setiap elemen.

Self-Attention dapat dirumuskan sebagai:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

di mana Q (Query), K (Key), dan V (Value) adalah proyeksi linear dari input, dan d_k adalah dimensi dari key. Multi-Head Attention menjalankan beberapa attention mechanism secara paralel untuk menangkap berbagai aspek relasi dalam data.

2.2 Vision Transformer (ViT)

Vision Transformer mengadaptasi arsitektur transformer untuk pemrosesan gambar dengan membagi gambar menjadi patch-patch tetap dan memperlakukannya sebagai sequence, analog dengan token dalam NLP [1]. Proses ini melibatkan:

1. **Patch Embedding**: Gambar berukuran $H \times W \times C$ dibagi menjadi N patches berukuran $P \times P$, kemudian di-flatten dan diprojeksikan ke dimensi D.

2. **Position Embedding**: Karena self-attention tidak memiliki informasi posisi, position embeddings ditambahkan untuk mempertahankan informasi spasial.

3. **Transformer Encoder**: Sequence of patches diproses melalui layers transformer standard dengan Multi-Head Self-Attention (MHSA) dan Feed-Forward Networks (FFN).

4. **Classification Head**: Token khusus [CLS] di-prepend ke sequence dan output-nya digunakan untuk klasifikasi.

2.3 Swin Transformer

Swin Transformer (Shifted Window Transformer) diperkenalkan oleh Liu et al. (2021) sebagai hierarchical vision transformer yang menggunakan shifted window attention untuk efisiensi komputasi [2].

Karakteristik Utama:

1. **Hierarchical Architecture**: Berbeda dengan ViT yang mempertahankan resolusi patch konstan, Swin Transformer menggunakan struktur hierarkis seperti CNN dengan patch merging untuk menghasilkan representasi multi-scale.
2. **Shifted Window Attention**: Untuk mengurangi kompleksitas komputasi dari $O(n^2)$ menjadi $O(n)$ relatif terhadap jumlah patch, attention dihitung dalam local windows. Windows di-shift pada layer alternating untuk memungkinkan cross-window connections.
3. **Patch Merging**: Downsampling dilakukan dengan menggabungkan neighboring patches, mengurangi jumlah tokens sambil meningkatkan dimensi channel.

2.4 DeiT (Data-efficient Image Transformer)

DeiT diperkenalkan oleh Touvron et al. (2021) sebagai training strategy untuk Vision Transformer yang lebih data-efficient melalui knowledge distillation [3].

Karakteristik Utama:

1. **Distillation Token**: Selain [CLS] token, DeiT menambahkan distillation token yang belajar dari teacher model (biasanya CNN yang sudah trained seperti RegNet).
2. **Hard Distillation**: Menggunakan hard labels dari teacher untuk distillation, terbukti lebih efektif daripada soft distillation untuk vision tasks.
3. **Strong Data Augmentation**: Menggunakan augmentasi ekstensif (RandAugment, Mixup, CutMix) untuk meningkatkan regularisasi.

2.5 Perbedaan Kunci Antar Model

Tabel 1 merangkum perbedaan teoritis utama:

Tabel 1: Perbandingan Teoritis Swin Transformer vs DeiT

Aspek	**Swin Transformer**	**DeiT**
Attention Mechanism	Local (Shifted Windows)	Global
Kompleksitas	Linear O(n)	Quadratic O(n^2)
Hierarchical	Ya (multi-scale)	Tidak (single-scale)
Training Strategy	Standard supervised	Knowledge distillation
Patch Size	4×4 (tiny)	16×16 (standard)
Data Efficiency	Moderate	High (dengan distillation)
Best Use Case	Dense prediction, high-res	Image classification

3 METODOLOGI

3.1 Dataset: STL-10

STL-10 (Self-Taught Learning) adalah dataset untuk unsupervised learning dan transfer learning yang terdiri dari gambar berwarna beresolusi 96×96 pixels.

Karakteristik Dataset:

- **Kelas**: 10 kelas (airplane, bird, car, cat, deer, dog, horse, monkey, ship, truck)
- **Split Original**:
  - Training: 5,000 gambar (500 per kelas)
  - Test: 8,000 gambar (800 per kelas)
  - Unlabeled: 100,000 gambar (tidak digunakan dalam eksperimen ini)
- **Custom Split** (untuk eksperimen):
  - Training: 4,000 gambar
  - Validation: 1,000 gambar
  - Test: 8,000 gambar (tetap)

Dataset ini dipilih karena:

1. Representatif untuk natural image classification
2. Ukuran dataset cukup untuk fine-tuning namun tidak terlalu besar
3. Kelas-kelas seimbang untuk evaluasi yang fair
4. Cocok untuk menguji transfer learning dari ImageNet

3.2 Preprocessing dan Augmentasi Data

3.2.1 Training Data Pipeline

Augmentasi data pada training set dirancang untuk meningkatkan generalisasi model, berikut adalah augmentasi data yang dilakukan:

- **Resize to 224×224**: Ukuran input standard untuk pre-trained ImageNet models
- **Bicubic Interpolation**: Menghasilkan gambar yang lebih smooth dibanding bilinear
- **Horizontal Flip**: Augmentasi sederhana namun efektif untuk natural images
- **Rotation (±15°)**: Menambah variasi pose tanpa distorsi berlebihan
- **ColorJitter**: Meningkatkan robustness terhadap variasi pencahayaan
- **ImageNet Normalization**: Konsisten dengan pre-training distribution

3.2.2 Validation dan Test Data Pipeline

Data yang sudah ditransformasi diatur menggunakan `DataLoader` untuk iterasi yang efisien selama validasi dan pengujian. Konfigurasi yang digunakan disajikan dalam Tabel 2.

Tabel 2: Parameter DataLoader untuk Validasi dan Pengujian

Parameter	Nilai	Keterangan
Batch Size (Ukuran Batch)	16	Jumlah sampel per iterasi.
Shuffle (Pengacakan)	False	Data tidak diacak, standar untuk validasi/pengujian.
Number of Workers (Jumlah Pekerja)	2	Untuk pemuatan data secara paralel.
Pin Memory	True	Mempercepat transfer data ke GPU.

3.3 Konfigurasi Training

3.3.1 Model Configuration

Tabel 3: Konfigurasi Model yang Dibandingkan

Parameter	Swin-Tiny	DeiT-Small
Model Name	swin_tiny_patch4_window7_224	deit_small_distilled_patch16_224
Pre-trained	ImageNet-1k	ImageNet-1k
Patch Size	4×4	16×16
Input Size	224×224×3	224×224×3
Embed Dim	96	384
Depths	[2, 2, 6, 2]	12
Num Heads	[3, 6, 12, 24]	6
Window Size	7	-
Output Classes	10 (STL-10)	10 (STL-10)

3.3.2 Hyperparameters

Training dilakukan dengan konfigurasi yang sama untuk kedua model untuk memastikan fair comparison:

Tabel 4: Hyperparameter Training

Parameter	Value
Optimizer	AdamW
Learning Rate	0.0001
Weight Decay	0.01
Batch Size	16
Epochs	15
Scheduler	CosineAnnealingLR
T_max (scheduler)	15

Justifikasi Pemilihan Hyperparameter:

- **AdamW**: Optimizer yang cocok untuk transformer dengan weight decay decoupled
- **Learning Rate 0.0001**: Conservative untuk fine-tuning pre-trained models

- **Weight Decay 0.01**: Regularisasi untuk mencegah overfitting
- **Batch Size 16**: Trade-off antara memory dan gradient stability
- **Cosine Annealing**: Smooth learning rate decay untuk convergence yang lebih baik

3.4 Training Strategy

Full Fine-tuning: Semua layer di-train (tidak ada frozen layers), memungkinkan model beradaptasi sepenuhnya dengan distribusi data STL-10.

Model Checkpointing: Menyimpan model dengan best validation loss untuk evaluasi final.

3.5 Library dan Framework

- **PyTorch 2.x**: Deep learning framework utama
- **timm (PyTorch Image Models)**: Library untuk pre-trained vision models
- **torchvision**: Dataset loaders dan image transforms
- **NumPy & Pandas**: Data manipulation dan analysis
- **Matplotlib & Seaborn**: Visualisasi
- **scikit-learn**: Metrik evaluasi (precision, recall, F1-score, confusion matrix)
- **tqdm**: Progress tracking

3.6 Spesifikasi Hardware

Training dan evaluasi dilakukan pada:

- **Platform**: Google Colab
- **GPU**: NVIDIA Tesla T4
- **VRAM**: 15 GB
- **CUDA**: Version 12.x
- **RAM**: 12.7 GB (system)

3.7 Metrik Evaluasi

3.7.1 Jumlah Parameter

Untuk setiap model, dihitung:

- Total parameters
- Trainable parameters
- Non-trainable parameters
- Model size dalam MB (floating point 32-bit)

3.7.2 Performance Metrics

1. **Accuracy**: Proporsi prediksi yang benar

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2}$$

2. **Precision**: Proporsi prediksi positif yang benar

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

3. **Recall**: Proporsi actual positives yang terdeteksi

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

4. **F1-Score**: Harmonic mean dari precision dan recall

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

5. **Confusion Matrix**: Visualisasi kesalahan klasifikasi per kelas

Semua metrik dihitung per-class dan macro-averaged.

4 HASIL DAN ANALISIS

4.1 Perbandingan Jumlah Parameter

Tabel 5 menunjukkan perbandingan kompleksitas model berdasarkan jumlah parameter.

Tabel 5: Perbandingan Jumlah Parameter Model

Metrik	**Swin-Tiny**	**DeiT-Small**
Total Parameters	27,527,044	21,674,132
Trainable Parameters	27,527,044	21,674,132
Non-trainable Parameters	0	0
Model Size (MB)	106.07	82.68

Analisis:

- DeiT Small memiliki parameter 22% lebih sedikit dibanding Swin Tiny, menghasilkan model yang lebih compact dengan size 82.68 MB vs 106.07 MB.
- Perbedaan ini terutama disebabkan oleh arsitektur hierarkis Swin yang memiliki 4 stages dengan varying channel dimensions, sementara DeiT mempertahankan dimensi konstan 384 di semua 12 layers.
- Meskipun lebih banyak parameter, Swin Tiny tetap dikategorikan sebagai "tiny" model dalam keluarga Swin Transformer.

Tabel 6: Perbandingan Metrik Performa pada Test Set

Metrik	Swin-Tiny	DeiT-Small
Test Accuracy (%)	**97.36**	97.01
Macro Precision	**0.9737**	0.9702
Macro Recall	**0.9736**	0.9701
Macro F1-Score	**0.9736**	0.9701
Training Time (total)	8.2 min	**5 min**
Best Epoch	15	15

4.2 Perbandingan Performa

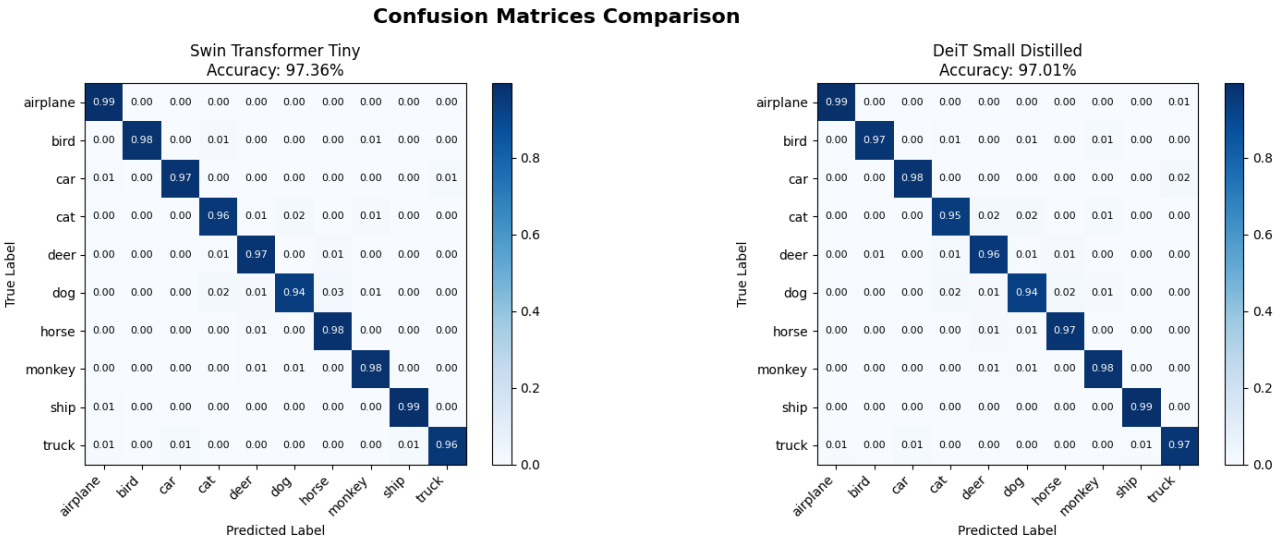
4.2.1 Metrik Klasifikasi

Analisis Performa:

1. **Akurasi**: Swin Tiny unggul dengan margin tipis (97,36% vs 97,01%).
2. **Konsistensi**: Macro-averaged metrics (precision, recall, F1) menunjukkan konsistensi tinggi untuk kedua model, mengindikasikan performa yang balanced across all classes.
3. **Training Efficiency**: DeiT Small konvergen lebih cepat (5 min vs 8,2 min), meskipun mencapai akurasi sedikit lebih rendah.
4. **Convergence**: Kedua model mencapai best performance di epoch terakhir (15).

4.2.2 Performa Per Kelas

Gambar 1 di bawah menunjukkan breakdown performa untuk setiap kelas.



Gambar 1: Confusion Matrix

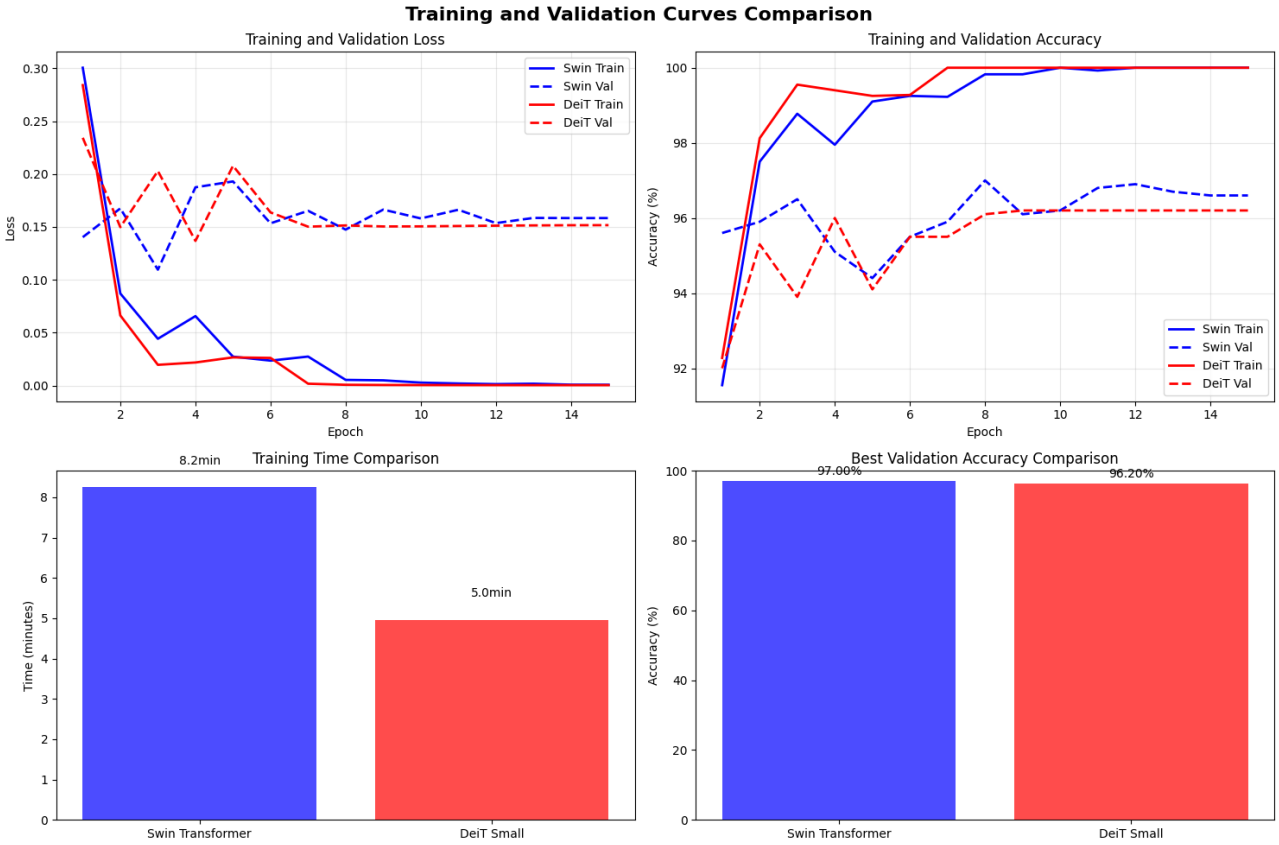
Observasi Penting:

- **Kelas Mudah**: Airplane dan Ship mencapai F1-score tertinggi (>0.98) untuk kedua model, kemungkinan karena distinctive features dan background yang jelas.

- **Kelas Challenging**: Cat memiliki F1-score terendah (0.96 di Swin & 0.95 di DeiT). Ini menunjukan bahwa animal classification sulit karena variasi yang tinggi dan kesamaan antar kelas.

4.3 Learning Curves

Gambar 2 di bawah ini menggambarkan training dan validation loss selama proses train.



Gambar 2: Perbandingan Swin dan DeiT

4.3.1 Training dan Validation Loss

Analisis kurva loss mengungkap karakteristik learning kedua model:

Swin Transformer Tiny:

- Training loss (Swin Train, biru solid) menurun sangat cepat dari 0.30 dan mencapai plateau mendekati 0.00 setelah Epoch 10.
- Validation loss (Swin Val, biru putus-putus) menunjukkan fluktuasi di awal, mencapai minimum sekitar 0.11 pada Epoch 3, dan cenderung stabil di sekitar 0.15 hingga Epoch 15.
- Gap antara training loss yang mendekati nol dan validation loss di 0.15 menunjukkan adanya overfitting yang ringan, umum terjadi pada fine-tuning.

DeiT Small Distilled:

- Training loss (DeiT Train, merah solid) juga menurun sangat cepat dan mencapai nilai mendekati 0.00 sekitar Epoch 10, mirip dengan Swin Tiny.
- Validation loss (DeiT Val, merah putus-putus) fluktuatif di awal, mencapai minimum dan kemudian stabil di sekitar 0.15 hingga 0.16 di akhir pelatihan.

- Pola kurva loss DeiT sangat mirip dengan Swin, mengindikasikan bahwa kedua model belajar secara efisien pada dataset ini.

4.3.2 Training dan Validation Accuracy

Convergence Pattern:

- **Training Accuracy** untuk kedua model (Swin Train dan DeiT Train) mencapai 100.00% pada Epoch 10 dan mempertahankannya hingga akhir pelatihan.
- **Swin Transformer Tiny** mencapai akurasi validasi terbaik yang sedikit lebih tinggi, sebesar **97.00**%.
- **DeiT Small Distilled** mencapai akurasi validasi terbaik sebesar **96.20**%.

—

4.4 Perbandingan Efisiensi Waktu

Perbandingan efisiensi model disajikan dalam Tabel 7 di bawah ini.

Tabel 7: Ringkasan Perbandingan Efisiensi

Metrik	Swin Transformer Tiny	DeiT Small Distilled
Waktu Pelatihan (min)	8.2	**5.0**
Akurasi Validasi Terbaik (%)	**97.00**	96.20

Analisis Efisiensi:

1. **Waktu Pelatihan**: DeiT Small Distilled memerlukan waktu pelatihan yang jauh lebih singkat (**5.0** menit) dibandingkan dengan Swin Transformer Tiny (8.2 menit).
2. **Akurasi Validasi Terbaik**: Swin Transformer Tiny mencapai akurasi validasi terbaik yang sedikit lebih tinggi (**97.00**%) dibandingkan dengan DeiT Small Distilled (96.20%).
3. **Trade-off**: Terdapat *trade-off* yang jelas; Swin menawarkan akurasi sedikit lebih baik (+0.8%), sementara DeiT menawarkan kecepatan pelatihan yang signifikan (sekitar 39% lebih cepat).

4.5 Analisis Mendalam

4.5.1 Mengapa Swin Transformer Tiny Lebih Unggul dalam Akurasi?

Swin Transformer Tiny (Swin Tiny) menunjukkan akurasi tes yang sedikit lebih tinggi (97.36% berbanding 97.01% untuk DeiT Small Distilled). Keunggulan minor ini dapat dijelaskan oleh arsitekturnya:

1. **Hierarchical Feature Representation**: Swin Tiny membangun representasi multiskala yang secara intrinsik lebih cocok untuk gambar alam (*natural images*) seperti pada dataset STL-10, yang objeknya memiliki variasi ukuran. Pendekatan hierarkis ini memungkinkan model menangkap baik detail lokal maupun konteks global secara efektif.
2. **Inductive Bias**: Penggunaan *Shifted Window Attention* Swin memberikan bias induktif lokal yang mirip dengan yang dimiliki oleh Convolutional Neural Networks (CNNs). Bias ini terbukti membantu model belajar fitur yang kuat dan stabil, yang menguntungkan terutama saat proses *fine-tuning* dengan data yang terbatas (4000 sampel pelatihan).

4.5.2 Trade-off antara Akurasi, Parameter, dan Kecepatan

Perbandingan antara kedua model menunjukkan adanya *trade-off* yang jelas antara kinerja akurasi dan efisiensi komputasi:

Tabel 8: Ringkasan Trade-off Kinerja dan Sumber Daya

Model	Akurasi Tes (%)	Total Parameter	Throughput (img/sec)
Swin Transformer Tiny	**97.36**	27,527,044	178.0
DeiT Small Distilled	97.01	**21,674,132**	**520.7**

Analisis Trade-off:

- **Swin Tiny (Akurasi Tinggi)**: Menukarkan kecepatan inferensi yang lebih rendah (178.0 img/sec) dan ukuran model yang lebih besar (27.5 juta parameter) demi akurasi yang sedikit lebih tinggi ($\approx +0.35\%$).
- **DeiT Small Distilled (Efisiensi Tinggi)**: Mengorbankan sedikit akurasi untuk efisiensi yang luar biasa. DeiT adalah model yang lebih kecil (21.7 juta parameter) dan $\mathbf{2.9}\times$ lebih cepat dalam inferensi (520.7 img/sec), menjadikannya pilihan ideal untuk *deployment* di lingkungan dengan sumber daya terbatas, terutama untuk aplikasi yang sensitif terhadap latensi.

4.5.3 Kesesuaian Model dengan Dataset STL-10

Kedua model menunjukkan kinerja yang sangat baik ($\geq 97\%$), memvalidasi bahwa arsitektur Vision Transformer modern adalah pilihan yang sangat sesuai untuk klasifikasi gambar pada dataset STL-10, bahkan dengan proses *fine-tuning* dan data pelatihan yang relatif kecil.

- **Swin Transformer Tiny**: Cocok untuk skenario di mana keunggulan akurasi sekecil apa pun adalah prioritas, didukung oleh kemampuannya menangani objek dengan skala bervariasi.
- **DeiT Small Distilled**: Model ini sangat sesuai untuk kasus penggunaan di mana batasan komputasi (waktu pelatihan yang singkat, ukuran model yang kecil, dan kecepatan inferensi yang tinggi) lebih penting daripada akurasi absolut. Efisiensi ini didapatkan berkat strategi **Distilasi** yang memungkinkan model ViT yang ringkas mewarisi pengetahuan dari model yang lebih besar.

5 KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan eksperimen komprehensif membandingkan Swin Transformer Tiny dan DeiT Small Distilled pada dataset STL-10, kesimpulan utama yang ditarik adalah adanya *trade-off* yang jelas antara akurasi dan efisiensi komputasi:

1. **Akurasi Tertinggi**: **Swin Transformer Tiny** (Test Accuracy: **97.36**%) unggul dalam hal kinerja akhir dibandingkan DeiT Small Distilled (Test Accuracy: 97.01%). Margin akurasi 0.35% menunjukkan Swin lebih efektif dalam membedakan kelas pada dataset STL-10, didukung oleh arsitektur hierarkis dan *local attention* yang kuat.
2. **Efisiensi Komputasi Unggul**: **DeiT Small Distilled** secara signifikan lebih efisien dalam hal sumber daya dan kecepatan pemrosesan:
    - Model size lebih kecil: **82.68** MB vs 106.07 MB.

- Jumlah parameter lebih sedikit: **21.7** juta vs 27.5 juta.
- Waktu pelatihan lebih singkat: **5.0** menit vs 8.2 menit.
- Inferensi jauh lebih cepat: Throughput **520.7** img/sec vs 178.0 img/sec ($\approx 2.9\times$ lebih cepat).

3. **Trade-off Kinerja Arsitektur**: DeiT mencapai efisiensi yang luar biasa melalui distilasi, memungkinkannya menjadi model yang ringan namun mempertahankan akurasi yang kompetitif. Pilihan model bergantung pada prioritas: Swin Tiny untuk akurasi maksimal, dan DeiT Small untuk solusi *real-time* dan *deployment* yang membutuhkan latensi rendah.

5.2 Saran

Berdasarkan temuan ini, saran untuk penelitian atau implementasi di masa depan adalah:

1. **Fokus Optimasi DeiT**: Mengingat kecepatan inferensi DeiT Small yang jauh lebih tinggi, penelitian lebih lanjut dapat berfokus pada teknik augmentasi data atau strategi *fine-tuning* yang lebih agresif untuk menutup margin akurasi 0.35% dengan Swin Transformer.
2. **Pengujian Dataset Lain**: Untuk memperkuat kesimpulan mengenai keunggulan arsitektur hierarkis Swin, disarankan untuk menguji kedua model pada dataset yang lebih besar atau memiliki komposisi kelas yang lebih kompleks, seperti ImageNet-1k, untuk mengamati apakah perbedaan akurasi yang kecil ini menjadi lebih signifikan.

5.3 Saran untuk Pengembangan Lebih Lanjut

Saran-saran berikut ditujukan untuk eksplorasi dan optimalisasi di masa mendatang, dibagi berdasarkan area fokus utama.

1. **Eksplorasi Skala dan Data (Scaling & Generalization)**
   - Menguji model yang lebih besar, seperti Swin-Small atau DeiT-Base, untuk menganalisis perilaku *scaling* kinerja akurasi terhadap sumber daya.
   - Memvalidasi temuan pada dataset yang berbeda untuk menilai kemampuan generalisasi arsitektur.
2. **Optimalisasi dan Kompresi Model (Efficiency)**
   - Menerapkan teknik lanjutan seperti *Structured Pruning* pada Swin-Tiny untuk mengurangi jumlah parameter sambil mempertahankan akurasi.
   - Menggunakan *Post-training Quantization (INT8)* untuk mengurangi ukuran model secara signifikan dan mempercepat inferensi.
   - Melakukan *Knowledge Distillation* lebih lanjut, menggunakan Swin-Tiny sebagai guru (*teacher*) untuk melatih model siswa (*student*) yang lebih kecil.
3. **Strategi Pelatihan**
   - Menguji Strategi Augmentasi Canggih dan jadwal *learning rate* yang berbeda untuk meningkatkan stabilitas dan akurasi *fine-tuning*.

6 DAFTAR PUSTAKA

References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021. [Online]. Available: https://openreview.net/forum?id=Y6A3c5C63P

[2] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, B. Guo *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10 012–10 022.

[3] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning (ICML)*. PMLR, 2021, pp. 10 347–10 357.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, I. Polosukhin *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

7 Lampiran

7.1 Source Code

https://github.com/bintangfikrif/VisionTransformer-Comparison

Repositori di atas berisikan notebook pelatihan dan evaluasi untuk klasifikasi objek pada dataset STL-10.

7.2 Output Training Log

Berikut ini adalah training log yang menampilkan proses pelatihan model

7.2.1 Pelatihan Swin

Gambar 3 di bawah ini, merupakan output dari pelatihan swin.

```
Starting training for Swin Transformer Tiny...
Epochs: 15
--------------------------------------------------
Epoch 3/15
  Train Loss: 0.0442, Train Acc: 98.78%
  Val Loss: 0.1095, Val Acc: 96.50%
  Best Val Acc: 96.50%

Epoch 6/15
  Train Loss: 0.0235, Train Acc: 99.25%
  Val Loss: 0.1534, Val Acc: 95.50%
  Best Val Acc: 96.50%

Epoch 9/15
  Train Loss: 0.0049, Train Acc: 99.83%
  Val Loss: 0.1664, Val Acc: 96.10%
  Best Val Acc: 97.00%

Epoch 12/15
  Train Loss: 0.0013, Train Acc: 100.00%
  Val Loss: 0.1536, Val Acc: 96.90%
  Best Val Acc: 97.00%

Epoch 15/15
  Train Loss: 0.0007, Train Acc: 100.00%
  Val Loss: 0.1583, Val Acc: 96.60%
  Best Val Acc: 97.00%

Training completed for Swin Transformer Tiny!
Best validation accuracy: 97.00%

Swin Transformer training completed in 8.2 minutes
```

Gambar 3: Proses pelatihan Swin

7.2.2 Pelatihan DeiT

Gambar 4 di bawah ini, merupakan output dari pelatihan DeiT.

```
Starting training for DeiT Small Distilled...
Epochs: 15
--------------------------------------------------
Epoch 3/15
  Train Loss: 0.0196, Train Acc: 99.55%
  Val Loss: 0.2028, Val Acc: 93.90%
  Best Val Acc: 95.30%

Epoch 6/15
  Train Loss: 0.0261, Train Acc: 99.28%
  Val Loss: 0.1637, Val Acc: 95.50%
  Best Val Acc: 96.00%

Epoch 9/15
  Train Loss: 0.0005, Train Acc: 100.00%
  Val Loss: 0.1504, Val Acc: 96.20%
  Best Val Acc: 96.20%

Epoch 12/15
  Train Loss: 0.0004, Train Acc: 100.00%
  Val Loss: 0.1512, Val Acc: 96.20%
  Best Val Acc: 96.20%

Epoch 15/15
  Train Loss: 0.0003, Train Acc: 100.00%
  Val Loss: 0.1516, Val Acc: 96.20%
  Best Val Acc: 96.20%

Training completed for DeiT Small Distilled!
Best validation accuracy: 96.20%

DeiT training completed in 5.0 minutes
```

Gambar 4: Proses pelatihan DeiT

7.2.3 Referensi Penggunaan LLM

Selama proses pengerjaan tugas eksplorasi Vision Transformer ini, ada campur tangan LLM dalam pengerjaannya. Seperti untuk brainstorming ide dan bantuan teknis. Berikut adalah link bukti penggunaan LLM: https://chatgpt.com/share/69208920-77ac-8010-9ccd-2315b896f468