بهینه‌سازی، فشرده‌سازی و تحلیل امنیت مدل‌های یادگیری عمیق

مدل هدف: MobileNetV2

دیتاست‌ها:

۱. **CIFAR-10** (تصاویر ۳۲×۳۲ پیکسل - رزولوشن پایین)

۲. **STL-10** (تصاویر ۹۶×۹۶ پیکسل - رزولوشن بالا)

هدف پروژه

هدف این پروژه، پیاده‌سازی و مقایسه‌ی جامع تکنیک‌های آموزش (Training)، فشرده‌سازی (Compression) و استقرار (Deployment) مدل‌های عصبی است. دانشجویان موظف‌اند یک پایپ‌لاین کامل را روی دو دیتاست با کیفیت‌های متفاوت اجرا کرده و نتایج را تحلیل کنند.

بخش اول: استراتژی‌های آموزش

در این فاز باید مدل MobileNetV2 را با رویکردهای زیر آموزش دهید و دقت نهایی را ثبت کنید:

۱. تنظیم جزئی (Partial Fine-Tuning) – اجباری

- **روش:** بارگذاری وزن‌های ImageNet، فریز کردن تمام لایه‌های استخراج ویژگی و آموزش تنها لایه طبقه‌بند (Classifier).
- **هدف:** ایجاد خط مبنا (Baseline) برای سنجش سرعت همگرایی.

۲. تنظیم دقیق کامل (Full Fine-Tuning) – اجباری

- **روش:** آنفریز کردن کل شبکه پس از آموزش اولیه و آموزش مجدد تمام لایه‌ها با نرخ یادگیری بسیار پایین (1e-5).
- **هدف:** دستیابی به بالاترین دقت ممکن (Standard Baseline).

۳. تقطیر دانش (Knowledge Distillation) – امتیازی (Bonus)

- **توضیح:** انتقال دانش از یک مدل بزرگ (Teacher) به مدل کوچک (Student).
- **روش:** استفاده از **روش هینتون (Hinton's Response-Based Knowledge Distillation)**.
  - معلم: ResNet-50.
  - دانش‌آموز: MobileNetV2 (با مقداردهی تصادفی).
  - تابع هزینه (Loss): ترکیبی از CrossEntropy (برای لیبل‌های واقعی) و KL-Divergence (برای نزدیک کردن توزیع احتمال خروجی دانش‌آموز به معلم با پارامتر دما T).

بخش دوم: جراحی و بهینه‌سازی ساختاری (Structural Optimization)

روی مدل **Full Fine-Tuned** (و مدل Distilled در صورت انجام)، موارد زیر را اجرا کنید:

۱. تفسیرپذیری (Grad-CAM) – اجباری

- استخراج نقشه‌های حرارتی (Heatmaps) روی تصاویر تست.
- **تحلیل:** بررسی اثر رزولوشن تصویر (CIFAR vs STL) بر روی قابلیت تفسیرپذیری مدل.

۲. هرس کردن (Pruning) – اجباری

- **Iterative Unstructured:** حذف ٪۲۰ وزن‌ها و بازآموزی (در ۳ مرحله).
- **Structured:** حذف ٪۴۰ کانال‌های خروجی در لایه آخر.

۳. فرضیه بلیت بخت‌آزمایی (Lottery Ticket Hypothesis - LTH) – امتیازی (Bonus)

- **توضیح:** این فرضیه بیان می‌کند که در شبکه‌های عصبی متراکم، زیرشبکه‌هایی (Sub-networks) وجود دارند که اگر با **همان وزن‌های اولیه (Initialization Weights)** آموزش ببینند، به دقتی معادل مدل اصلی می‌رسند.
- **تسک:**

۱. ماسک هرس (Pruning Mask) مدل آموزش‌دیده را استخراج کنید.

۲. وزن‌های مدل را دقیقاً به مقادیر لحظه شروع (قبل از آموزش) بازگردانید (Rewind).

۳. ماسک را اعمال کرده و مدل تنک (Sparse) را مجدداً آموزش دهید. آیا مدل به دقت قبلی می‌رسد؟

۴. تجزیه ماتریس (Low-Rank Factorization / SVD) – امتیازی (Bonus)

- **توضیح:** ماتریس وزن لایه‌های کاملا متصل (fully connected) ($W \in R^{m \times n}$) دارای افزونگی اطلاعاتی هستند. با استفاده از **تجزیه مقادیر منفرد (SVD)**، می‌توان این ماتریس را به ضرب دو ماتریس کوچکتر ($U \in R^{m \times k}$ و $V \in R^{k \times n}$) تقریب زد که $k \ll m, n$.
- **تسک:** لایه آخر (Classifier) مدل را با SVD تجزیه کرده و مدل را مجدداً (Fine-Tune) کنید تا افت دقت جبران شود. تاثیر آن بر حجم مدل را گزارش دهید.

بخش سوم: تست استقرار و امنیت (Security & Deployment)

۱. کوانتیزه‌سازی (Quantization) – اجباری

- **روش:** تبدیل مدل‌ها به فرمت Int8 با استفاده از روش **Post-Training Quantization (FX Graph Mode)**.
- **تحلیل:** مقایسه میزان افت دقت در مدل "Full Fine-Tuned" نسبت به مدل "Distilled". (کدام مدل پایدارتر است؟)

۲. پایداری (Robustness) – اجباری

- ارزیابی دقت مدل روی تصاویر تخریب‌شده با نویز گاوسی (Gaussian Blur).

۳. حمله خصمانه (Adversarial Attack - FGSM) – امتیازی (Bonus)

- **توضیح:** روش **FGSM (Fast Gradient Sign Method)** با محاسبه گرادیان تابع هزینه نسبت به پیکسل‌های ورودی، نویزهای نامرئی اما جهت‌داری را به تصویر اضافه می‌کند که باعث خطای مدل می‌شود.
- **تسک:** پیاده‌سازی حمله با epsilon=0.05 و مقایسه مقاومت مدل‌های مختلف.

۴. خروجی نهایی – اجباری

- تبدیل مدل نهایی به فرمت ONNX.

۵. خروجی مورد انتظار

یک گزارش نهایی شامل جدول مقایسه‌ای زیر برای **هر دو دیتاست**:

روش	دقت پایه	دقت بعد از هرس	دقت کوانتیزه (Int8)	دقت در حمله FGSM
Partial Tuning	...	-	...	-
Full Fine-Tuning	...	...	...	...
Distillation (Bonus)	...	...	...	...