**به نام خدا**

**۱. معماری Encoder-Decoder**

### **۱.۱ اصول معماری Encoder-Decoder:**

**Encoder:**

* **ورودی:** متن مبدأ (مثلاً انگلیسی)
* **خروجی:** Representation فشرده از متن
* **کار:** فهم و encode کردن متن ورودی

**Decoder:**

* **ورودی:** خروجی Encoder + متن تولیدشده تا الآن
* **خروجی:** متن مقصد (مثلاً فارسی)
* **کار:** تولید متن به صورت قدم‌به‌قدم

### **۱.۲ نقشه داده‌ها:**

[Encoder]

Input: "Hello world" → Embedding → Positional Encoding → Self-Attention → ...

[Decoder]

Input: "<start>" + Encoder Output → Self-Attention → Cross-Attention → ...

Output: "سلام" → "سلام دنیا" → "سلام دنیا <end>"

### **۱.۳ انواع مدل‌های Encoder-Decoder:**

* **T5 (Text-to-Text):** همه چیز به صورت متن-به-متن
* **BART:** Denoising autoencoder
* **MarianMT:** مخصوص ترجمه
* **PEGASUS:** مخصوص خلاصه‌سازی

## ****۲. مدل‌های T5 و BART****

### **۲.۱ مدل T5 (Text-to-Text Transfer Transformer):**

**ایده اصلی:** همه tasks به صورت text-to-text

"translate English to German: The house is wonderful" → "Das Haus ist wunderbar"

"summarize: long text..." → "short summary"

"cola sentence: The dog plays" → "acceptable"

**کاربرد های T5-small (مثال پیاده شده با T5-Small)**

**T5-small برای این کارها عالی است:**

* ✅ **خلاصه‌سازی** متن‌های طولانی
* ✅ **بازنویسی** و پارافریز
* ✅ **تحلیل احساسات**
* ✅ **پاسخ به سوال**
* ✅ **تولید متن** خلاق
* ✅ **سیستم‌های چت‌بات** ساده

**اما برای این کارها مناسب نیست:**

* ❌ **ترجمه** بین زبان‌ها
* ❌ **کارهای بسیار تخصصی**
* ❌ **متون بسیار طولانی**

## ****نتیجه‌گیری:****

* **T5-small برای ترجمه مناسب نیست** - فقط متن را تکرار می‌کند
* برای **ترجمه** از مدل‌های مخصوص مانند Helsinki-NLP/opus-mt-en-fr استفاده کنید
* **T5-base** یا **T5-large** برای ترجمه بهتر هستند اما حجم بیشتری دارند
* **Pipeline** ساده‌ترین روش برای ترجمه است

**پیشنهاد:** از راه حل ۳ (Pipeline) استفاده کنید چون ساده‌تر و قابل اطمینان‌تر است! 🚀