

---

# Clase 141 — Encoder-Decoder para traducción

Parte: 2 — Deep Learning · Fuente: Géron, cap. 16 § Encoder–Decoder Network for Neural Machine Translation. Duración estimada: 80 min.

## Clase 141 — Encoder-Decoder para traducción

Parte: 2 — Deep Learning · Fuente: Géron, cap. 16 § Encoder-Decoder Network for Neural Machine Translation. Duración estimada: 80 min.

### Objetivo

Implementar la arquitectura seq2seq (Sutskever, Vinyals & Le 2014) — un encoder que comprime la oración fuente en un vector de contexto + un decoder que genera la oración destino token por token. Conocer teacher forcing (durante training, el decoder ve los targets reales como input) vs inference autoregresiva. Esta arquitectura es la antesala de atención (clase 125) y de Transformers (clase 126).

### Resultados de aprendizaje

Al finalizar, el estudiante podrá:

- Construir un seq2seq con dos LSTM: encoder devuelve state, decoder usa ese state como inicialización.
- Aplicar teacher forcing: en training, decoder input = target shifted; en inference, autoregresivo.
- Tokens especiales: <start>, <end>, <pad>, <unk>.
- Reconocer la limitación del cuello de botella (todo el meaning de la oración fuente en un vector fijo) — motivación para atención.
- Evaluar traducción con BLEU (nltk.translate.bleu\_score).

### Temas

- Arquitectura clásica seq2seq con 2 LSTM.
- Teacher forcing vs scheduled sampling.
- Inference autoregresiva con generate loop.
- Beam search (mejor que greedy).
- BLEU como métrica.
- Limitación del bottleneck → motivó atención (Bahdanau 2014).

### Definiciones y características

- Encoder: LSTM que procesa la fuente, devuelve final\_state = (h<sub>T</sub>, c<sub>T</sub>).
- Decoder: LSTM inicializado con final\_state del encoder; genera target.
- Teacher forcing: durante training, decoder recibe el target real (shifted) en lugar de su propia predicción.
- <start> / <end> tokens: marcadores para iniciar/parar generación.
- BLEU: métrica de calidad de traducción basada en n-grams overlapping. Va de 0 a 100.
- Beam search: durante inference, mantiene los k mejores prefijos en lugar de greedy.

### Dataset / recursos

- Tatoeba English-Spanish (small): <https://www.manythings.org/anki/>
- Librerías: tensorflow, keras, nltk (para BLEU).

## Ejercicios

1. Preparar datos: tokenizar source y target, agregar <start> y <end> al target, padding.
2. Encoder: Embedding → LSTM(256, return\_state=True). Mantener state\_h, state\_c.
3. Decoder en training: Embedding(decoder\_input) → LSTM(256, initial\_state=encoder\_state) → Dense(target\_vocab, softmax).
4. Inference loop: feed <start>, predecir, alimentar prediction como next input, hasta <end> o max\_len.
5. BLEU: calcular sobre el test set.

## Homework verificable

Traducción inglés → español con seq2seq:

1. Dataset Tatoeba (~10k frases).
2. Encoder + decoder LSTM con 256 units.
3. Train 30 épocas con teacher forcing.
4. Inference autoregresiva + BLEU.

Criterio de aceptación: BLEU > 10 en test (modesto pero válido para LSTM básico); muestras de traducción reconocibles aunque imperfectas.

## Errores comunes

Síntoma / mensaje	Causa y cómo arreglar
Modelo genera <end> inmediatamente	El bias del decoder lleva siempre a <end>.
Greedy inference repite palabras	"the the the". Fix: beam search o repetiti
Inference más lento que esperado	Loop Python en cada token. Fix: usar tf.fu
BLEU = 0	Mismatch en tokens entre prediction y refe
Vocabulario target muy grande	Sample softmax o tied embeddings. Fix: lim

## Preguntas frecuentes

¿seq2seq aún se usa?

Conceptualmente sí (Transformer es seq2seq con atención). Como RNN-seq2seq pure: solo histórico o casos muy específicos. Para traducción sería → Transformer + Hugging Face (clase 127).

¿Bottleneck del estado fijo?

Una sola tupla (h, c) para resumir 50 tokens de input es muy pobre. Atención (clase 125) resuelve dando al decoder acceso a TODOS los estados del encoder.

¿Teacher forcing vs scheduled sampling?

Teacher forcing = siempre target real. Funciona, pero hay mismatch entre train y inference. Scheduled sampling alterna entre target y predicción propia → mejor.

¿Cómo manejo vocabulario grande?

BPE / SentencePiece (clase 127). Vocab típico 32k subwords cubre cualquier idioma con OOV manejable.

¿BLEU es buena métrica?

Aceptable para traducción "literal". Falla con paráfrasis. Para evaluación humana o LLM-as-judge, mejores opciones.

## Referencias

- Géron, cap. 16 — Encoder-Decoder Network for Neural Machine Translation.
- Sutskever, Vinyals & Le (2014), Sequence to Sequence Learning with Neural Networks, NeurIPS.
- Cho et al. (2014), Learning Phrase Representations using RNN Encoder-Decoder.
- Papineni et al. (2002), BLEU, ACL.

## Siguiente clase

Clase 142 — Mecanismos de atención

## Apéndice: notebook (primer bloque)

Primera celda ejecutable del notebook de la clase.

```
# Imports y configuración inicial
```

## Archivos complementarios

- notebook.ipynb