

Día 12 · Tema 4 — Ejercicios: difusión y modelos multimodales

Cómo se usa esta hoja. Hay dos tipos de ejercicio, y sólo dos.

El ejercicio del día. Uno por hoja. Es el obligatorio y es el que entra en el examen, con el cálculo incluido. Basta con hacer ese para seguir el curso con solvencia.

Los ejercicios de ampliación. El resto. Están para enseñar de qué es capaz el método, no para exigir: **no se piden calculados en el examen.** De ellos sólo puede preguntarse la idea, a nivel conceptual.

EL EJERCICIO DEL DÍA *obligatorio · entra en el examen*

1. Proceso forward: schedule de ruido

Consideramos un proceso de difusión con $T = 3$ pasos y schedule de varianzas $\beta_1 = 0,1$, $\beta_2 = 0,2$, $\beta_3 = 0,3$. Partimos de un dato escalar $x_0 = 1,0$.

Recordamos que $\alpha_t = 1 - \beta_t$ y $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$, y que la distribución marginal es $q(x_t | x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t} x_0, (1 - \bar{\alpha}_t) I)$.

- Calcula $\alpha_1, \alpha_2, \alpha_3$.
- Calcula $\bar{\alpha}_1, \bar{\alpha}_2, \bar{\alpha}_3$.
- Para cada $t \in \{1, 2, 3\}$, calcula la media $\sqrt{\bar{\alpha}_t} x_0$ y la desviación típica $\sqrt{1 - \bar{\alpha}_t}$ de $q(x_t | x_0)$.
- Interpreta cómo evoluciona la relación señal/ruido a medida que t crece.

AMPLIACIÓN *opcional · no se pide calculado en el examen*

2. Muestreo directo desde x_0

Utilizando los valores de $\bar{\alpha}_t$ calculados en el ejercicio anterior, supongamos que hemos extraído una realización concreta $\varepsilon = 0,5$ de $\mathcal{N}(0, 1)$. Usando la fórmula de muestreo directo

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \varepsilon,$$

- calcula x_1, x_2 y x_3 numéricamente (cuatro decimales).
- ¿A qué valor convergería x_t si siguiéramos añadiendo pasos ($T \rightarrow \infty$ con $\bar{\alpha}_T \rightarrow 0$)?

AMPLIACIÓN *opcional · no se pide calculado en el examen*

3. Paso reverso: de x_2 a x_1

En el proceso reverso, la red neuronal ε_θ predice el ruido que se añadió, y la media de la distribución reversa se calcula como

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \varepsilon_\theta(x_t, t) \right).$$

Supongamos que estamos en $t = 2$ con $x_2 = 0,6$ y que la red predice $\varepsilon_\theta(x_2, 2) = 0,4$. Usando $\alpha_2 = 0,8$, $\beta_2 = 0,2$ y $\bar{\alpha}_2 = 0,72$:

- Calcula $\frac{\beta_2}{\sqrt{1 - \bar{\alpha}_2}}$.
- Calcula $\mu_\theta(x_2, 2)$. Este es el valor estimado de x_1 .
- Si la predicción de ruido fuese perfecta ($\varepsilon_\theta = \varepsilon$ real), ¿qué significaría geométicamente μ_θ ?

AMPLIACIÓN *opcional · no se pide calculado en el examen*

4. Classifier-free guidance

Dado un modelo de difusión condicionado por texto, la predicción combinada con *classifier-free guidance* es

$$\hat{\varepsilon} = \varepsilon_\theta(x_t, t, \varnothing) + w(\varepsilon_\theta(x_t, t, c) - \varepsilon_\theta(x_t, t, \varnothing)),$$

donde $\varnothing$ indica la predicción incondicional, c es el prompt de texto y w es la escala de guía.

Supón que, para un cierto x_t y paso t , las predicciones bidimensionales son:

$$\varepsilon_\theta(x_t, t, \varnothing) = (0,3, 0,5), \quad \varepsilon_\theta(x_t, t, c) = (0,1, 0,8),$$

donde c es el prompt `cat`.

- Calcula la dirección condicional: $\Delta = \varepsilon_\theta(x_t, t, c) - \varepsilon_\theta(x_t, t, \varnothing)$.

- (b) Calcula $\hat{\epsilon}$ para $w = 1$ (estándar), $w = 3$, $w = 7,5$ (valor por defecto en Stable Diffusion) y $w = 15$.
- (c) Calcula la norma $\|\hat{\epsilon}\|$ para cada valor de w . ¿Qué observas?
- (d) Discute el compromiso calidad/diversidad: ¿qué ocurre con valores de w muy altos?

AMPLIACIÓN *opcional · no se pide calculado en el examen*

5. Relación señal-ruido (SNR)

La *signal-to-noise ratio* (SNR) en el paso t se define como

$$\text{SNR}(t) = \frac{\bar{\alpha}_t}{1 - \bar{\alpha}_t}.$$

- (a) Usando los $\bar{\alpha}_t$ del ejercicio 1, calcula $\text{SNR}(1)$, $\text{SNR}(2)$ y $\text{SNR}(3)$.
- (b) ¿Para qué valor de $\bar{\alpha}_t$ se cumple $\text{SNR} = 1$ (señal y ruido iguales)? ¿Está cerca de alguno de nuestros pasos t ?
- (c) En los modelos de difusión reales (e.g., DDPM con $T = 1000$), el SNR decae de $\approx 10^3$ a $\approx 10^{-3}$. ¿Qué implica esto sobre los pasos iniciales y finales del proceso forward?

AMPLIACIÓN *opcional · no se pide calculado en el examen*

6. Coste de una campaña publicitaria multimodal

Un equipo de marketing quiere generar contenido para el lanzamiento de un producto utilizando modelos generativos. El plan de producción y los costes unitarios son:

Modalidad	Cantidad	Coste unitario	Detalle
Imágenes de producto	500	\$0,04/imagen	DALL-E 3 / Midjourney
Descripciones de texto	500	ver abajo	Claude Sonnet
Locuciones de voz	100	\$0,04/segundo	ElevenLabs (15 s cada una)
Clips de vídeo	10	\$1,50/clip	Sora (5 s cada uno)

Para las descripciones de texto, cada llamada usa 500 tokens de entrada (\$3/Mtok) y genera 200 tokens de salida (\$15/Mtok).

- (a) Calcula el coste total de cada modalidad por separado.
- (b) Calcula el coste total de toda la campaña.
- (c) ¿Qué modalidad domina el presupuesto? ¿Es sorprendente?
- (d) Si el presupuesto es \$500, ¿cuántas campañas completas se podrían financiar?

AMPLIACIÓN *opcional · no se pide calculado en el examen*

7. Coste de tokenización de imágenes

Un modelo visión-lenguaje (e.g., GPT-4o, Claude) procesa imágenes convirtiéndolas en ~ 1000 tokens por imagen. El precio de entrada es \$3/Mtok.

- (a) ¿Cuál es el coste por imagen analizada?
- (b) Una empresa de e-commerce quiere analizar su catálogo de 50 000 imágenes de producto para extraer atributos (color, material, estilo). ¿Cuál es el coste total vía API?
- (c) Compara con la alternativa humana: 5 anotadores a \$2 000/mes, cada uno etiqueta 200 imágenes/día (22 días laborables/mes). ¿Cuántos meses tardan en etiquetar las 50 000 imágenes y cuánto cuesta en total?
- (d) Calcula el ratio coste-humano/coste-API. ¿Qué factores cualitativos podrían justificar aún así el etiquetado humano?