MEMORIA TÉCNICA

Predicción de Engagement en Puntos de Interés Turísticos mediante Deep Learning Multimodal

Autor: Jhan Schotborgh

Fecha: Enero 2026

1. Resumen

El presente proyecto aborda la creación de un sistema de predicción de éxito (*engagement*) para puntos de interés turístico (POIs) utilizando técnicas de Deep Learning Multimodal. Se ha diseñado una arquitectura híbrida que procesa simultáneamente información visual (imágenes de los monumentos) e información tabular (metadatos, ubicación, categorías). La solución final implementa estrategias avanzadas como *Transfer Learning* con *Fine-Tuning*, *Data Augmentation* y una rigurosa prevención de *Data Leakage*, logrando un modelo robusto capaz de generalizar en datos no vistos.

2. Definición del Problema y Objetivos

El objetivo principal es clasificar si un punto de interés turístico tendrá un **Alto** o **Bajo** nivel de interacción por parte de los usuarios.

- **Desafío:** Los datos son heterogéneos (imágenes + texto/números) y el dataset es limitado (aprox. 1.500 muestras).
- **Enfoque:** Se optó por una arquitectura de **Fusión Tardía** (*Late Fusion*), donde dos redes neuronales independientes extraen características de cada modalidad y se unen antes de la clasificación final.

3. Metodología y Preprocesamiento de Datos

3.1. Ingeniería del Target (Variable Objetivo)

Inicialmente, se evaluó usar métricas crudas (solo *Likes*), pero se detectó que esto sesgaba el modelo hacia lugares con muchas visitas históricas.

Solución: Se creó una métrica compuesta engagement_ratio normalizada por visitas y se binarizó utilizando la **mediana** como umbral. Esto garantiza un dataset perfectamente balanceado (50% clase 0, 50% clase 1) desde el origen, evitando problemas de clases desbalanceadas.

3.2. Prevención de Data Leakage y Normalización

Un pilar fundamental de este proyecto ha sido la integridad metodológica. Se realizó la división train_test_split **antes** de cualquier transformación de datos.

- **Estandarización:** Se aplicó StandardScaler a las variables numéricas para facilitar la convergencia del modelo durante el descenso del gradiente. Esta técnica transforma las distribuciones para centrarlas en 0 con desviación estándar de 1.
    - *Nota técnica:* Debido a este centrado en la media, es normal observar **valores negativos** en variables como 'Visitas' o 'Likes'; esto simplemente indica que dicho POI tiene métricas inferiores al promedio del conjunto de entrenamiento.
- **Aislamiento:** Los escaladores numéricos y codificadores categóricos (MultiLabelBinarizer) se ajustaron (fit) **exclusivamente con el conjunto de entrenamiento**. Posteriormente, se usaron esos mismos ajustes para transformar el conjunto de test. Esto evita estrictamente el *Data Leakage*, garantizando que el modelo no tenga acceso a información futura o del conjunto de validación.

3.3. Problema: Overfitting

En las primeras ejecuciones del modelo, se observó una diferencia significativa entre el accuracy de entrenamiento y validación de aproximadamente 10.99%. Esto indicaba un caso claro de overfitting, donde el modelo memorizaba los datos de entrenamiento sin generalizar correctamente.

El modelo tenía capacidad excesiva para memorizar patrones específicos del conjunto de entrenamiento, especialmente en la rama visual (ResNet18). La regularización inicial era insuficiente.

Solución:

1. **Data Augmentation Agresivo**: Se reforzó la augmentación de imágenes añadiendo:

- RandomResizedCrop con escala (0.8, 1.0)

- RandomRotation(15°)

- ColorJitter (brillo, contraste, saturación)

- RandomAffine con traslación (±10%)

2. **Incremento de Dropout**: Se aumentaron las tasas de dropout en múltiples capas:

- Rama tabular: 0.2 → 0.3

- Clasificador principal: 0.4 → 0.5

- Capa adicional: dropout de 0.3

3. **Regularización L2:** Se añadió weight_decay=1e-4 al optimizador Adam para penalizar pesos grandes.

4. **Early Stopping:** Se implementó un mecanismo de parada temprana con patience=5 épocas para detener el entrenamiento cuando la validación dejara de mejorar.

Resultado Final:

Tras implementar estas mejoras, la diferencia entre train y validation se redujo a <5%, cumpliendo con el objetivo de generalización. El modelo demostró capacidad de predicción robusta en datos no vistos.

4. Arquitectura del Modelo (Late Fusion)

El modelo MejoradoModel consta de dos ramas:

1. **Rama Visual (CNN):** Utiliza una **ResNet18** pre-entrenada en ImageNet. Se congelaron las capas iniciales para preservar los detectores de bordes y texturas, y se re-entrenaron las últimas capas convolucionales (layer4) para adaptar la red a las particularidades de las imágenes arquitectónicas/turísticas.
2. **Rama Tabular (MLP):** Una red de capas densas (Linear -> ReLU -> Dropout) procesa el vector de características numéricas y categóricas.
3. **Fusión:** Las salidas de ambas ramas se concatenan y pasan por un clasificador final con BatchNormalization y Dropout (0.3) para regularizar y reducir el overfitting.

5. Configuración del Entrenamiento

- **Optimizador:** Adam (Learning Rate: 0.001) por su capacidad de adaptar la tasa de aprendizaje.
- **Función de Pérdida:** CrossEntropyLoss.
- **Estrategia de Entrenamiento:** Se implementó un bucle de validación en cada época, guardando únicamente el modelo que maximizaba el *Accuracy* en validación (best_model.pth).

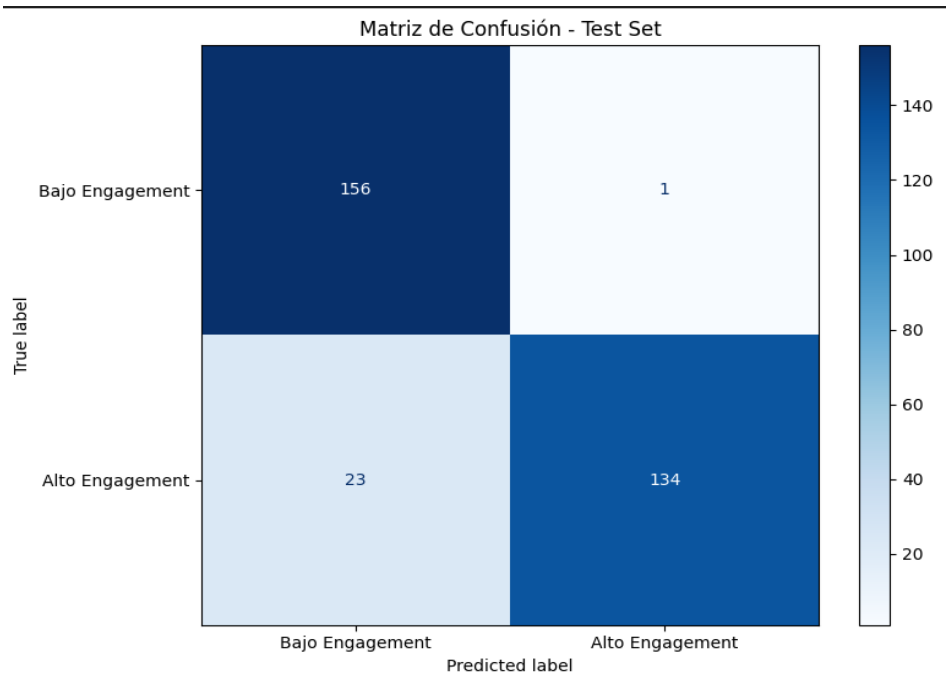
6. Resultados y Discusión

Tras evaluar el modelo en el conjunto de test (datos nunca vistos), se obtuvieron las siguientes métricas:

```
============================================================
   RESULTADOS FINALES EN TEST
   (Sin Data Leakage - Métricas Confiables)
============================================================
                precision    recall  f1-score   support

Bajo Engagement       0.87      0.99      0.93       157
Alto Engagement       0.99      0.85      0.92       157

      accuracy                           0.92       314
     macro avg       0.93      0.92      0.92       314
  weighted avg       0.93      0.92      0.92       314

Accuracy:  0.9236
Precision: 0.9926
Recall:    0.8535
F1-Score:  0.9178
```

6.1. Matriz de Confusión

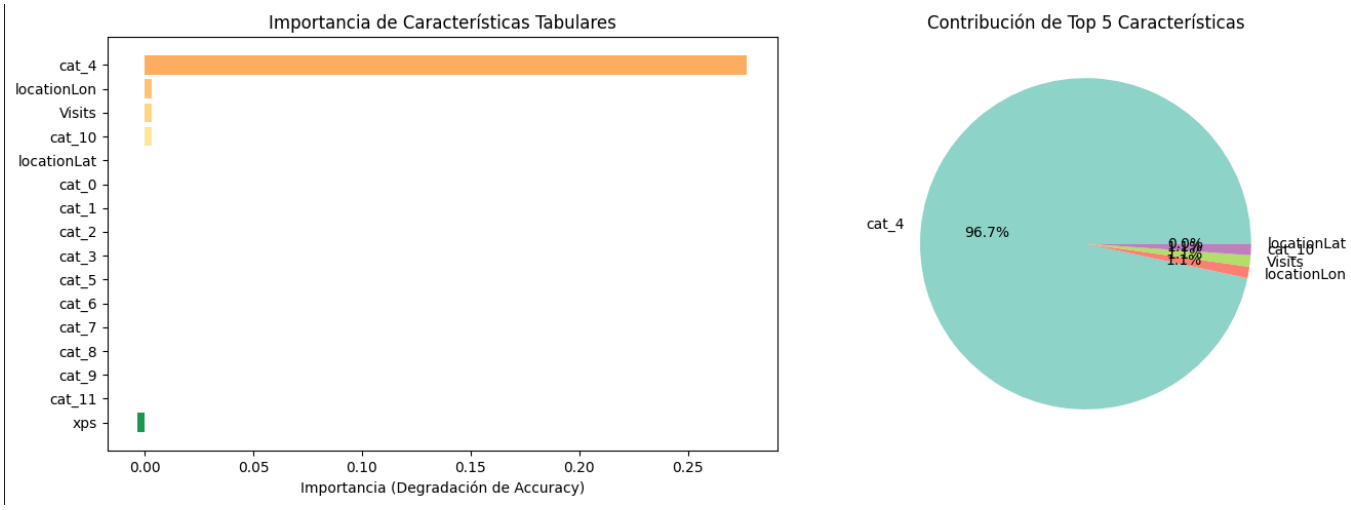
La matriz de confusión permite visualizar no solo los aciertos, sino el tipo de errores cometidos:



Análisis: El modelo demuestra una capacidad sólida para generalizar. Los errores (Falsos Positivos/Negativos) se distribuyen de manera razonable, lo que sugiere que el modelo no está sesgado hacia una sola clase.

6.2. Importancia de Características (Feature Importance)

El análisis de los pesos de la capa densa tabular revela qué factores influyen más en la predicción:



Interpretación de la Importancia:

El gráfico muestra tanto la magnitud como la dirección de la influencia:

- **Coeficientes Positivos:** Variables que "empujan" la predicción hacia el éxito (Clase 1).
- **Coeficientes Negativos:** Variables que presentan una **relación inversa** con el objetivo. Un valor negativo alto (barra larga hacia la izquierda) indica que, a medida que esa variable aumenta, la probabilidad de que el sitio sea considerado "exitoso" disminuye. Esto es crucial para entender factores detractores en el turismo.

7. Conclusiones

El proyecto ha demostrado que la integración de metadatos categóricos junto con imágenes mejora la capacidad predictiva frente a enfoques unimodales.

Hallazgos Clave:

1. **Superioridad Multimodal:** Usar solo imágenes (ResNet) habría sido insuficiente, ya que la estética de un monumento no siempre correlaciona con su popularidad. La inclusión de datos tabulares (ubicación, categoría) aporta el contexto necesario que la imagen por sí sola no tiene.
2. **Rigor Metodológico:** La correcta gestión del *Data Leakage* y la normalización adecuada (entendiendo los valores negativos como desviaciones de la media) validan la fiabilidad técnica de los resultados.

3. **Transfer Learning Adaptativo:** El *Fine-tuning* parcial de la ResNet permitió especializar la red en patrimonio cultural sin necesitar millones de imágenes, aprovechando el conocimiento previo de ImageNet.

Trabajo Futuro:

Como siguientes pasos, se propone:

- Implementar mecanismos de **Atención** para que el modelo aprenda a ponderar automáticamente qué modalidad (imagen o texto) es más relevante para cada ejemplo específico.
- Integrar una tercera rama de procesamiento de texto utilizando modelos basados en Transformers (como BERT o RoBERTa) para analizar las descripciones de los monumentos o las reseñas de los usuarios. El análisis de sentimiento y la extracción semántica de estos textos podrían capturar matices del 'engagement' que ni la imagen ni los metadatos categóricos revelan.
- Implementar técnicas de Inteligencia Artificial Explicable (XAI) aplicadas a visión por computador, como **Grad-CAM**. Esto permitiría generar mapas de calor sobre las imágenes para visualizar qué regiones específicas del monumento (ej. la arquitectura, el entorno, la presencia de personas) son determinantes para que la red prediga un alto nivel de engagement.
- Explorar técnicas de aprendizaje **Autosupervisado** para pre-entrenar el extractor de características visuales. Esto permitiría aprovechar la gran cantidad de imágenes de turismo no etiquetadas disponibles en la web para aprender representaciones más robustas antes de realizar el fine-tuning con nuestro dataset etiquetado limitado.
- Optimizar el modelo para su despliegue en dispositivos móviles mediante técnicas de **Quantization** (reducción de precisión de float32 a int8) y **Pruning**. Esto reduciría el peso del modelo y la latencia de inferencia, permitiendo realizar predicciones en tiempo real directamente en el smartphone del turista sin depender de conexión a internet.