Filtrado basado en contenido II Imágenes y Música

Denis Parra
PUC Chile

Introducción

- En la clase anterior revisamos la recomendación basada en contenido en comparación con el filtrado colaborativo y los modelos de feedback implícito.

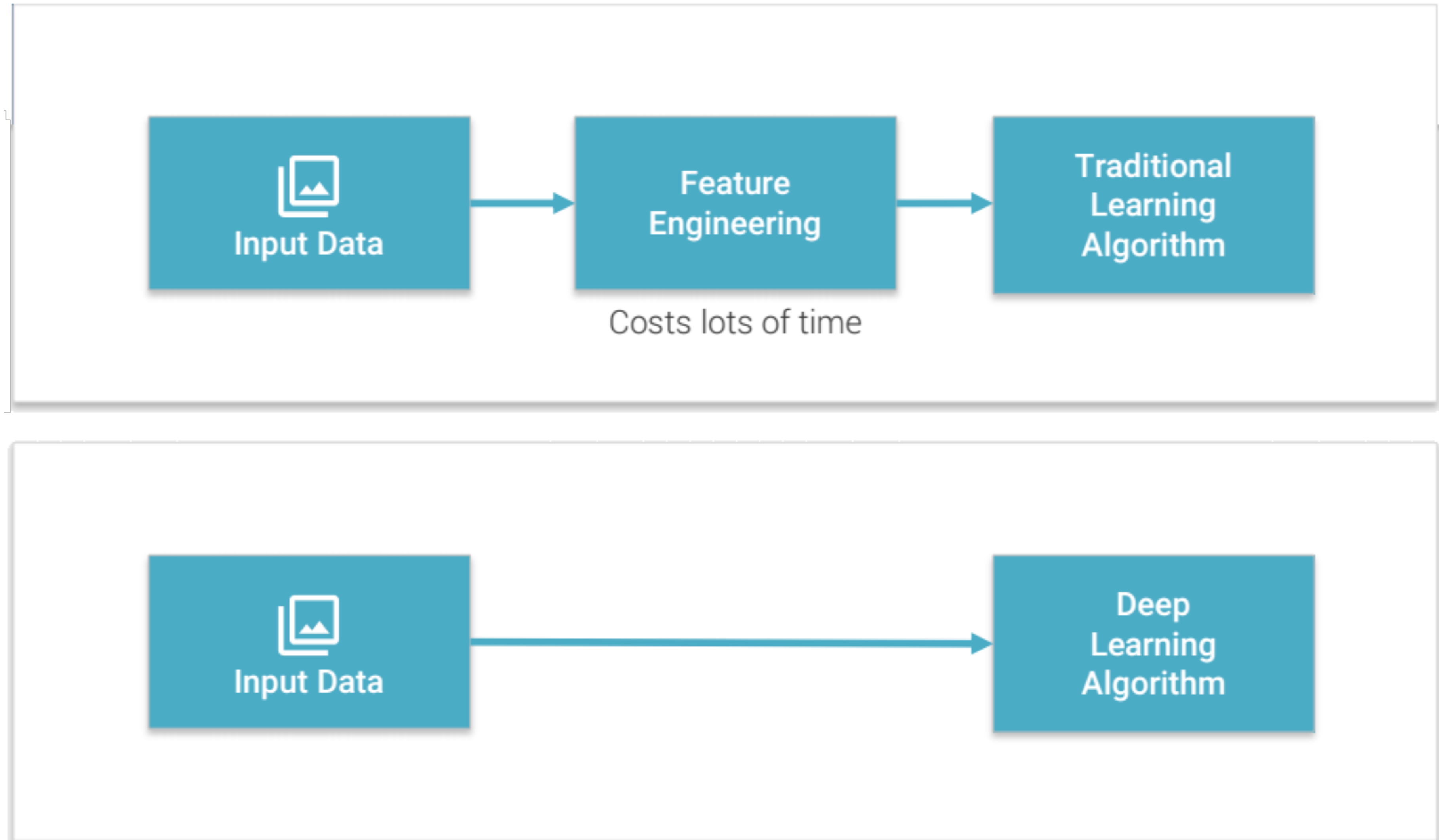
- Nos concentramos en contenido de texto y en técnicas para representarlo de forma estructurada:
  - Modelo de espacio vectorial (TF-IDF)
  - Modelos de tópicos: LSA (LSI) y LDA
  - Embeddings de palabras (W2V, GloVe) y de texto (ELMO, BERT)

Contenido Visual y Musical

- La representación del contenido visual y musical no es tan intuitiva como en el caso de texto.

- Si bien hay investigación madura en como representar música y texto, los modelos de Deep Learning de los años recientes han modificado profundamente esta área:
  - Modelos anteriores hacían feature (características) engineering
  - Modelos modernos usan Deep Learning (DL) para aprender las características.

DL para extracción de características



Temas de hoy

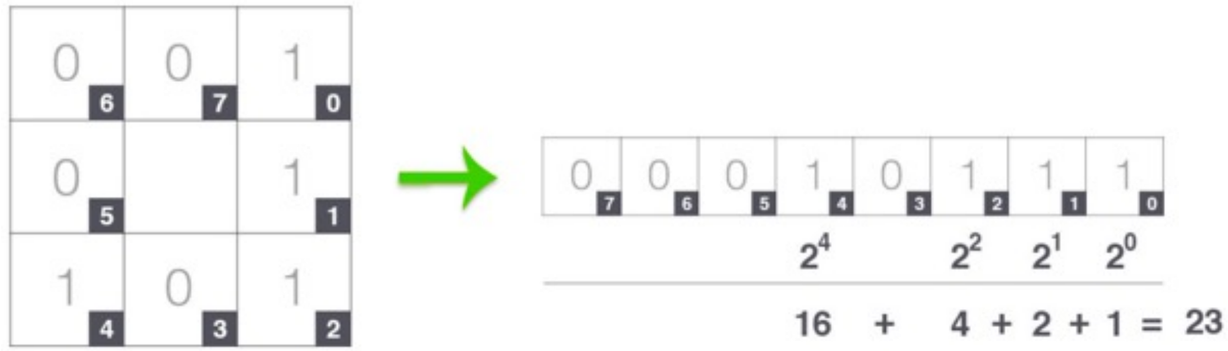
- Recomendación visual basada en contenido

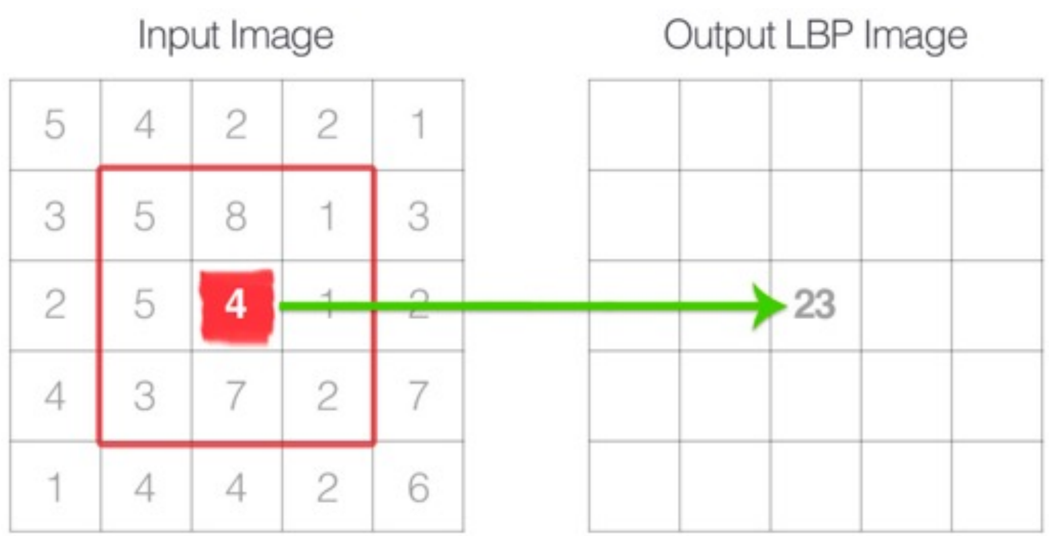
- Recomendación musical basada en contenido

Recomendación visual basada en contenido

- En los modelos tradicionales la extracción de características a partir de imágenes se realized por diferentes técnicas. Algunas de ellas:
  - Local Binary patterns (LBP): Método “manual” usado tradicionalmente como referencia de comparación en tareas de Visión por Computador. Obtiene un histograma de 59 patrones encontrados en una imagen.
  - Attractiveness : serie de 7 métricas

LBP





Fuente: https://www.pyimagesearch.com/2015/12/07/local-binary-patterns-with-python-opencv/

LBP

- Finalmente se calcula un histograma que tabula el número de ocasiones en que cada patron LBP ocurrió.
- Podemos pensar en este histograma como un vector de features.

Fuente: https://www.pyimagesearch.com/2015/12/07/local-binary-patterns-with-python-opencv/

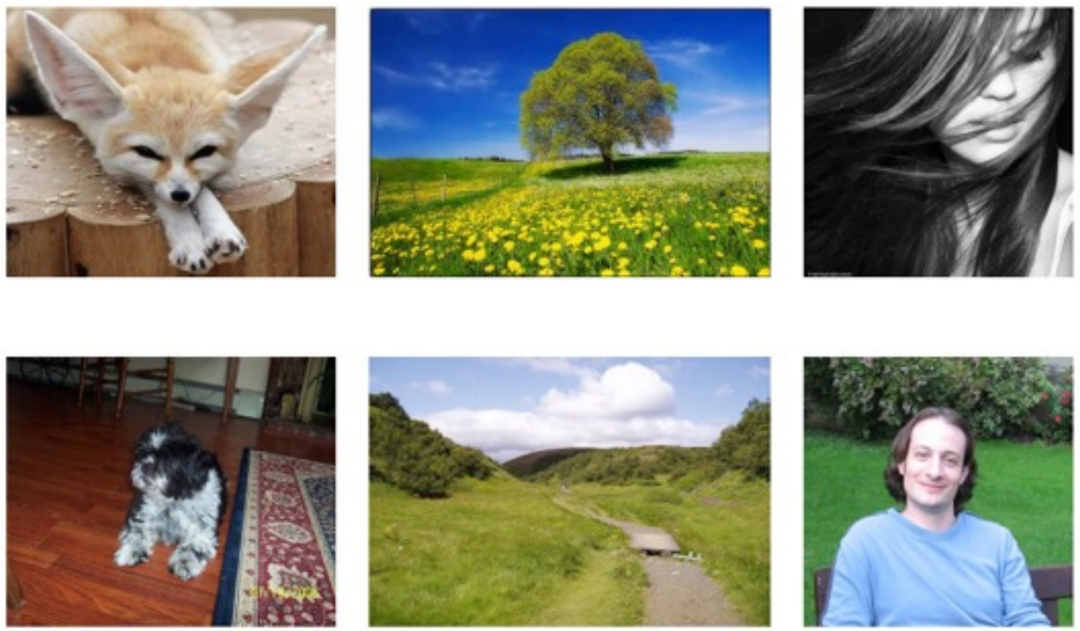
Características visuales atractivas

- San Pedro y Sierdorfer (2009) estudiaron características para caracterizar imágenes por su atractivo visual:
  - Brightness (brillo)
  - Saturation (saturación)
  - Sharpness (nitidez)
  - RMS-contrast (contraste RMS)
  - Colorfulness (colorido)
  - Naturalness (naturalidad)
  - Entropy (entropía)

San Pedro, J., & Siersdorfer, S. (2009). Ranking and classifying attractiveness of photos in folksonomies. In *Proceedings of the 18th international conference on World wide web* (pp. 771-780).

Ranking: texto vs. features visuales (Flickr)

Atractivas



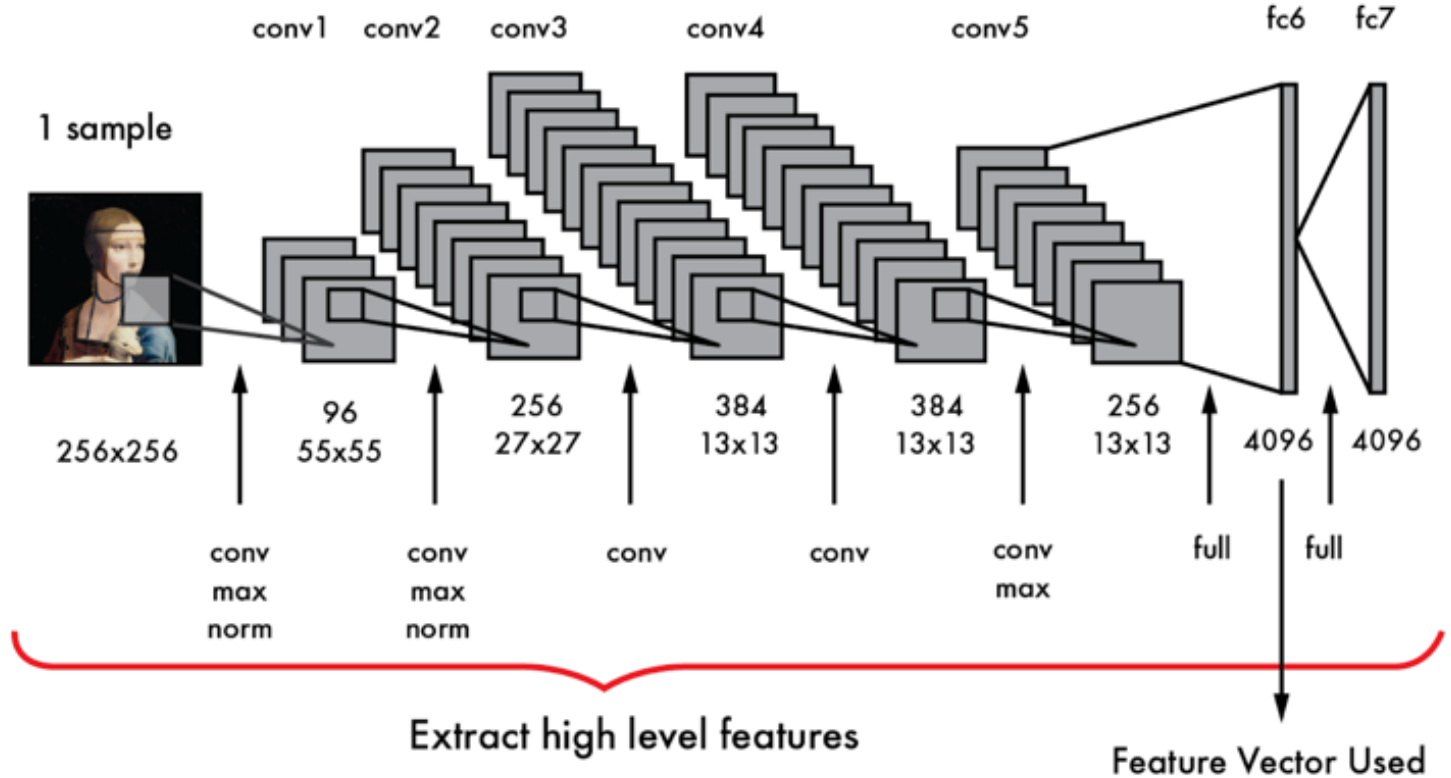
No Atractivas

Table 6: Ranking using Regression (Kendall's Tau-b): 40000 training photos

Method	Kendall's Tau-b
brightness	0.0006
contrast	-0.0172
RGB contrast	0.0288
saturation	0.1064
saturation variation	0.0472
colorfulness	-0.0497
sharpness	0.0007
sharpness variation	-0.0914
naturalness	0.0143
text	0.3629
visual	0.2523
text+visual	0.4841

Features manuales versus Deep Learning

- Con DL podemos usar features aprendidas automáticamente con una red neuronal pre-entrenada para otra tarea: clasificación de objetos del dataset Imagenet.



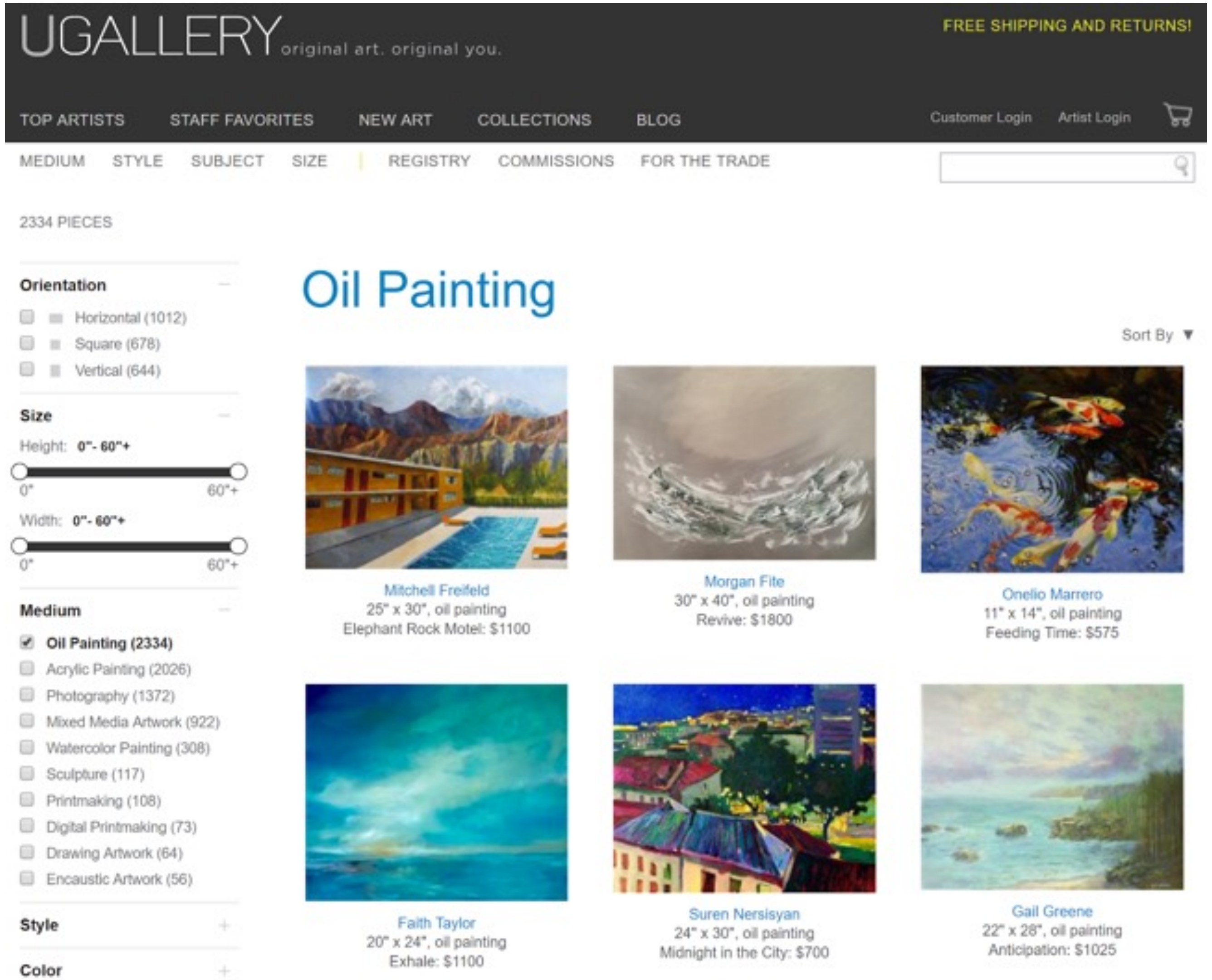
Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information pro*

Ejemplo

- Messina, P., Dominguez, V., Parra, D., Trattner, C., & Soto, A. (2019). Content-based artwork recommendation: integrating painting metadata with neural and manually-engineered visual features. *User Modeling and User-Adapted Interaction*, *29*(2), 251-290.

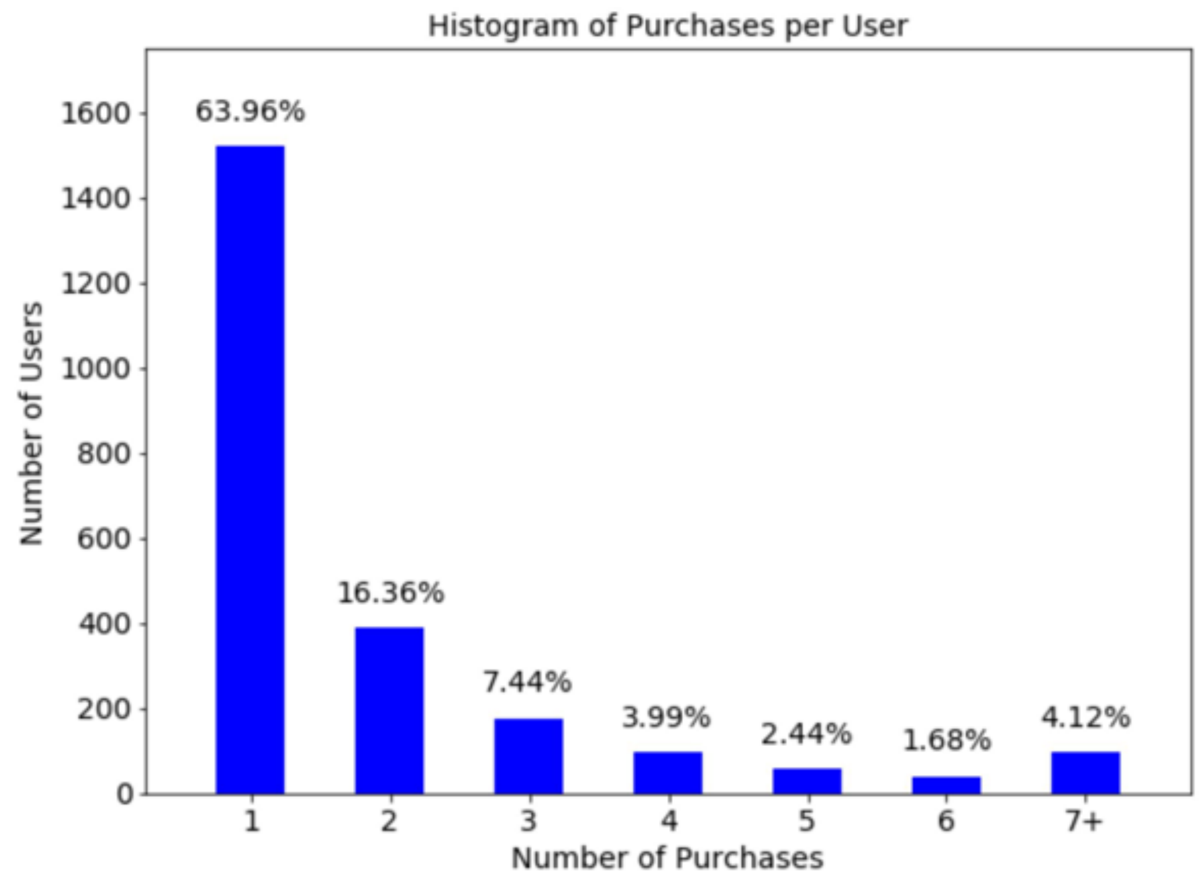
Problema: Recomendación de obras de arte

- Datos provistos por empresa Ugallery
- User feedback: transacciones (compras)
- Problema: una vez que un usuario compra una obra, sale del inventario (one of a kind)



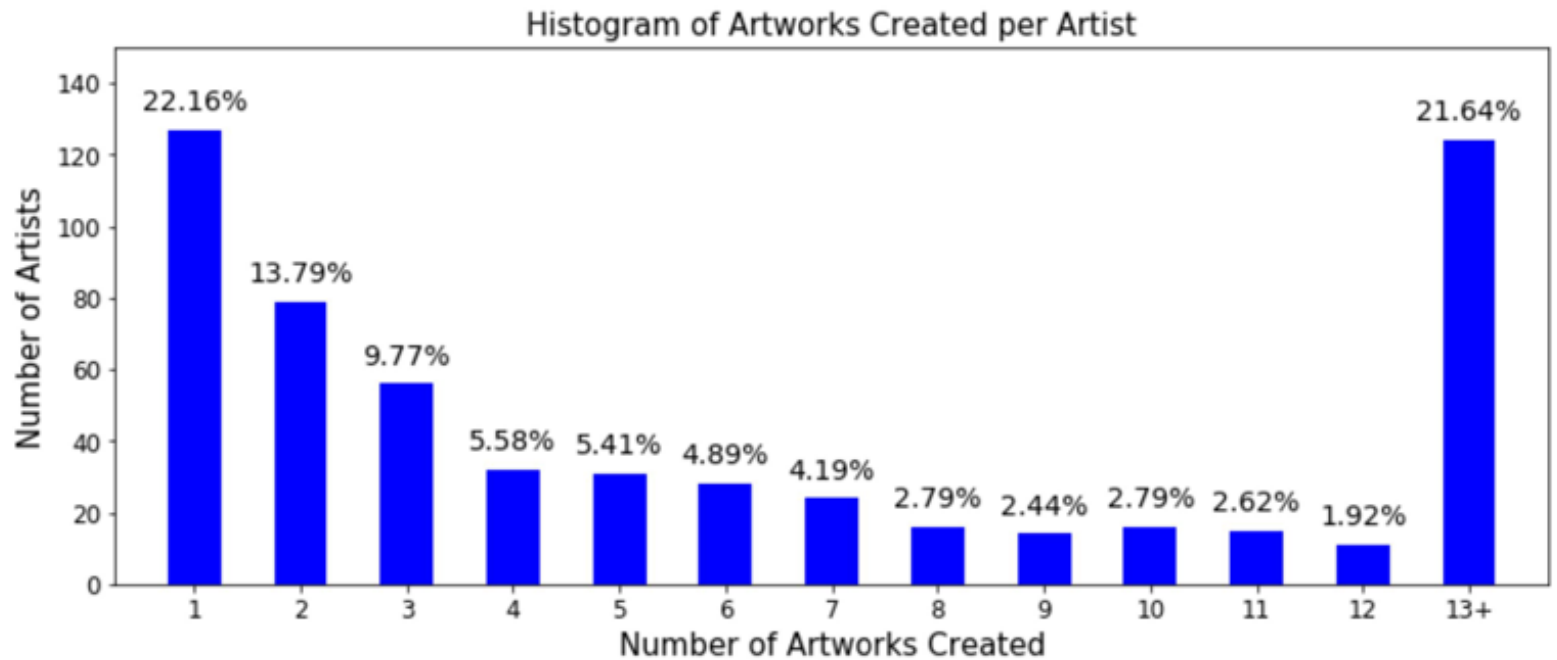
Dataset

- 5336 transacciones (compras)
- 2378 usuarios
- 6040 obras de arte

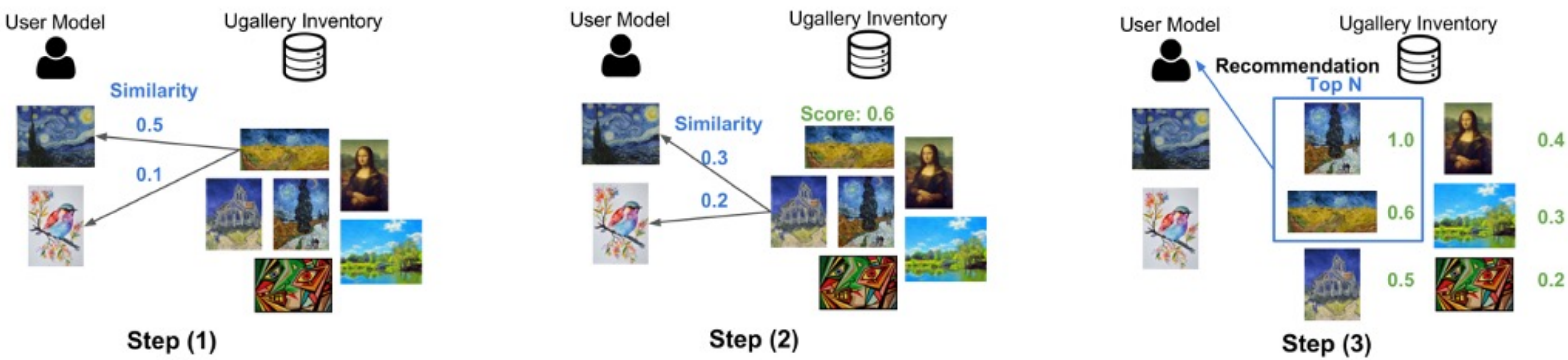


Dataset

- 573 artistas en total
- Un artista por obra
- 10,54 obras por artista en promedio



Recomendación basada en contenido



Métodos utilizados

1. Most Popular Curated Attribute Value (MPCAV)
2. Personalized Most Popular Curated Attribute Value (PMPCAV)
3. Personalized Favorite Artist (FA)
4. **Learned Visual Features: Deep Convolutional Neural Networks (CNN)**
5. **Handcrafted Visual Features (HVF)**
6. Hybrid Recommendations (Hybrid)

Score con HVF (manuals)

- Análogo a las CNNs: similaridad coseno + agregaciones (max, average, average-top-k)

$$sim(V_i^{Attract}, V_j^{Attract}) = cos(V_i^{Attract}, V_j^{Attract}) \qquad sim(V_i^{LBP}, V_j^{LBP}) = cos(V_i^{LBP}, V_j^{LBP})$$

- También probamos un híbrido Attractiveness + LBP:

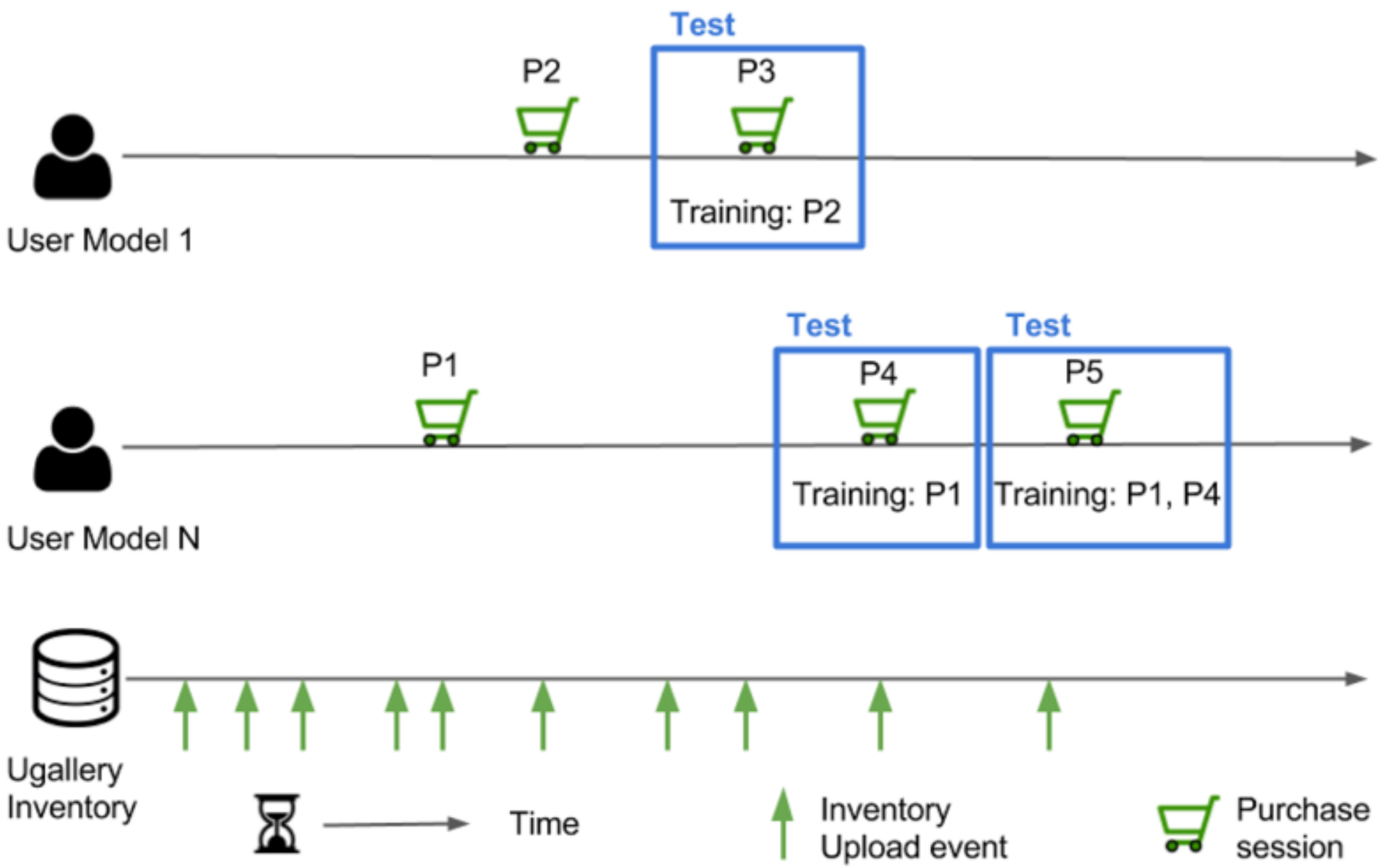
$$\begin{aligned} score(u,i)_{HVF} = {} & \alpha_1 \cdot score(u,i)_{Attractiveness} \\ & + \alpha_2 \cdot score(u,i)_{LBP} \end{aligned}$$

Score con features de red neuronal CNN

$$score(u,i)_X = \begin{cases} \max\limits_{j \epsilon P_u}\{sim(V_i^X, V_j^X)\} & (maximum) \\ \frac{\sum_{j \in P_u} sim(V_i^X, V_j^X)}{|P_u|} & (average) \\ \frac{\sum_{r=1}^{\min\{K,|P_u|\}} \max_{j \epsilon P_u}{}^{(r)}\{sim(V_i^X, V_j^X)\}}{\min\{K,|P_u|\}} & (average\ top\ K) \end{cases}$$

$$sim(V_i, V_j) = cos(V_i, V_j) = \frac{V_i \cdot V_j}{||V_i||\,||V_j||}$$

Evaluación offline



Resultados

ID	Method	nD@20	R@20	P@20	F1@20
1	CNN (All)	**.1295**^{3}	**.1702**^{4}	**.0151**^{2}	**.0248**^{2}
2	CNN (ResNet50)	.1247^{3}	.1628^{4}	.0145^{5}	.0236^{4}
3	CNN (AlexNet)	.1081^{4}	.1461^{4}	.0135^{5}	.0216^{4}
4	CNN (VGG19)	.1008^{5}	.1398^{8}	.0124^{6}	.0205^{5}
5	CNN (InceptionV3)	.1007^{6}	.1332^{8}	.0125^{4}	.0201^{6}
6	CNN (NASNet Large)	.0998^{8}	.1379^{8}	.0120^{8}	.0197^{7}
7	CNN (InceptionResNetV2)	.0932^{8}	.1300^{8}	.0119^{8}	.0192^{8}
8	HVF (LBP)	.0507^{9}	.0736^{11}	.0068^{9}	.0107^{9}
9	HVF (LBP + Attr.)	.0493^{11}	.0728^{11}	.0064^{10}	.0103^{11}
10	HVF (Attractiveness)	.0407^{11}	.0628^{11}	.0059^{11}	.0095^{11}
11	Random	.0097	.0200	.0015	.0025

Stat. significance by multiple t-tests, Bonferroni corr.
$\alpha_{bonf} = \alpha/n = 0.05/55 = .00091$.

Diversidad

ID	Method	F1@20	D@10 visual cluster	D@10 visual pairwise	D@10 artist	D@10 jaccard pairwise	D@10 color	D@10 medium
1	Hybrid$_1$ (FA+CNN+PMPCAV)	**.0333**^{2}	10.0697^{4}	.3952^{2}	8.4375^{3}	.7433^{2}	11.7362^{12}	2.2719^{3}
2	Hybrid$_2$ (FA+CNN)	**.0325**^{5}	9.1883	.3803	7.6165^{4}	.7730^{1}	12.0959^{3}	2.7902^{4}
3	Hybrid$_3$ (FA+PMPCAV)	**.0312**^{4}	11.8327^{9}	.4297^{9}	7.8472^{2}	.7214^{4}	11.8309^{12}	2.0459^{12}
4	FA	.0295^{5}	9.7124^{2}	.4092^{8}	2.8809	.7068	11.9983^{3}	2.3864^{1}
5	CNN (All)	.0248^{6}	9.6688^{2}	.3913^{2}	12.6822^{1}	.8488^{16}	12.6514^{7}	3.3951^{2}
6	CNN (ResNet50)	.0236^{8}	10.1429^{4}	.3968^{7}	12.6804^{1}	.8524^{5}	12.7164^{7}	3.4399^{2}
7	CNN (AlexNet)	.0216^{8}	10.1732^{4}	.3923^{2}	13.0314^{5}	.8502^{16}	12.4317^{2}	3.5119^{5}
8	CNN (VGG19)	.0205^{9}	10.6845^{7}	.4016^{6}	14.3341^{11}	.8648^{6}	13.0546^{15}	3.5386^{6}
9	CNN (InceptionV3)	.0201^{10}	11.2208^{8}	.4195^{11}	13.8768^{7}	.8712^{8}	13.1360^{15}	3.6926^{11}
10	CNN (NASNet Large)	.0197^{12}	11.0767^{8}	.4144^{8}	14.0180^{16}	.8697^{8}	13.1435^{15}	3.6827^{8}
11	CNN (InceptionResNetV2)	.0192^{13}	11.1313^{8}	.4151^{4}	14.0232^{16}	.8703^{8}	13.1871^{15}	3.6072^{6}
12	PMPCAV(All)	.0156^{13}	13.6607^{3}	.4498^{3}	14.4608^{11}	.7429^{3}	11.0691	1.8303^{16}
13	HVF (LBP)	.0107^{14}	14.6874^{12}	.4667^{12}	15.8733^{12}	**.8949**^{15}	13.9820^{11}	**4.1296**^{9}
14	HVF (LBP + Attr.)	.0103^{16}	15.3969^{13}	.4732^{13}	**16.3359**^{13}	**.8961**^{15}	**14.0628**^{11}	**4.0633**^{9}
15	HVF (Attractiveness)	.0095^{17}	**15.4358**^{13}	**.4743**^{13}	**16.5584**^{13}	.8850^{9}	12.8210^{7}	4.0569^{9}
16	MPCAV(Medium)	.0081^{17}	**15.4375**^{13}	**.4829**^{15}	13.7440^{7}	.7844^{2}	**14.3841**^{14}	1.0017
17	Random	.0025	**17.4006**^{16}	**.4972**^{16}	**18.4069**^{15}	**.9123**^{14}	**14.2869**^{14}	**4.5804**^{13}

Statistical significance was obtained using multiple pairwise t-tests with Bonferroni correction, $\alpha_{bonf} = \alpha/n = 0.05/136 = .00037$.

Evaluación online (8 curadores de UGallery)

Evaluación online (8 curadores de UGallery)

Name	**nD@5**	**nD@10**	**P@5**	**P@10**
Hybrid(FA+CNN+HVF)	**0.9042**	**0.8913**	**0.7500**	**0.6750**
Hybrid(CNN+HVF)	0.6747	0.6638	0.5000	0.4250
CNN	0.7176	0.6947	0.5000	0.4000
FA	0.4276	0.5662	0.3000	0.4000
HVF	0.5498	0.5314	0.3500	0.2625

Otro método: Visual BPR

- VBPR = Visual Bayesian Personalized Ranking (R. He & McAuley, 2016)

Vector de features desde AlexNet CNN

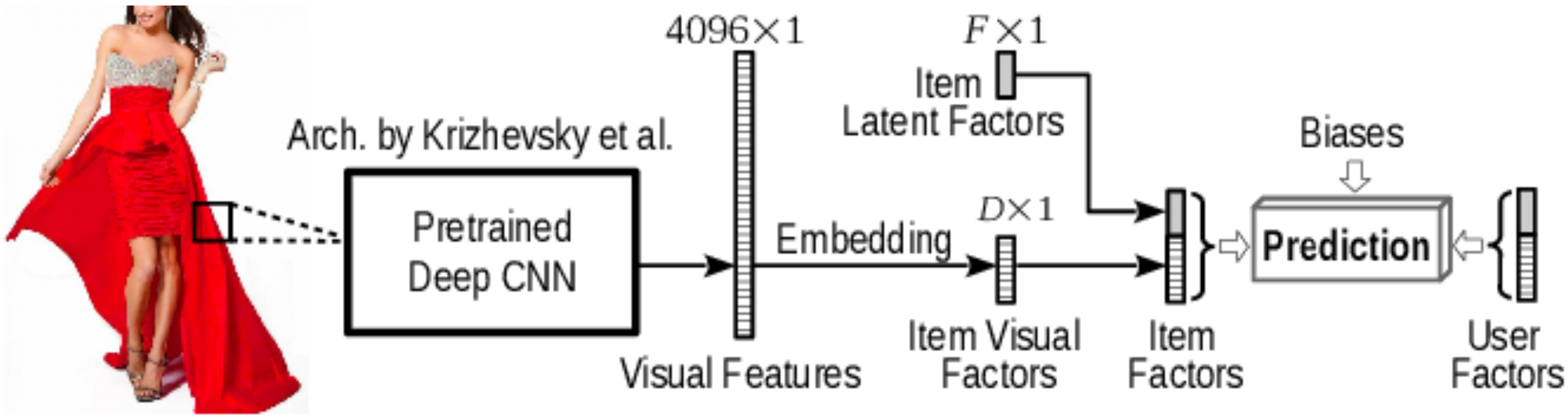
$$\hat{x}_{u,i} = \beta_i + \gamma_u^T \gamma_i + \theta_u^T (E f_i) + \beta'^T f_i$$

- Variables se aprenden con BPR-OPT (Rendle et al., 2009)

$$D_S = \{(u,i,j) | u \in U \wedge i \in I_u^+ \wedge j \in I \setminus I_u^+\}$$

$$\sum_{(u,i,j) \in D_S} ln(\sigma(\hat{x}_{uij}(\Theta))) - \lambda_\Theta ||\Theta||^2 \qquad \hat{x}_{uij}(\Theta) = \hat{x}_{u,i} - \hat{x}_{u,j}$$

VBPR



He, R., & McAuley, J. (2016). VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. ***AAAI*****.**

VBPR resultados

Table 3: AUC on the test set T (#factors = 20). The best performing method on each dataset is boldfaced.

Dataset	Setting	(a) RAND	(b) MP	(c) IBR	(d) MM-MF	(e) BPR-MF	(f) VBPR	improvement f vs. best	f vs. e
Amazon Women	All Items	0.4997	0.5772	0.7163	0.7127	0.7020	**0.7834**	9.4%	11.6%
	Cold Start	0.5031	0.3159	0.6673	0.5489	0.5281	**0.6813**	2.1%	29.0%
Amazon Men	All Items	0.4992	0.5726	0.7185	0.7179	0.7100	**0.7841**	9.1%	10.4%
	Cold Start	0.4986	0.3214	0.6787	0.5666	0.5512	**0.6898**	1.6%	25.1%
Amazon Phones	All Items	0.5063	0.7163	0.7397	0.7956	0.7918	**0.8052**	1.2%	1.7%
	Cold Start	0.5014	0.3393	**0.6319**	0.5570	0.5346	0.6056	-4.2%	13.3%
Tradesy.com	All Items	0.5003	0.5085	N/A	0.6097	0.6198	**0.7829**	26.3%	26.3%
	Cold Start	0.4972	0.3721	N/A	0.5172	0.5241	**0.7594**	44.9%	44.9%

Fine-tuning: ¿es necesario?

- del Rio, F., Messina, P., Dominguez, V., & Parra, D. (2018). Do Better ImageNet Models Transfer Better… for Image Recommendation?. *arXiv preprint arXiv:1807.09870*.

Table 1: Results of different pre-trained embeddings at the artwork image recommendation task to the left (R:Recall, P:Precision), and their performance at the ILSVRC Challenge trained on ImageNet dataset (Acc: Accuracy). The top methods in both tasks do not correlate.

CNN	Artwork Image Recommendation				ILSVRC-2012-CLS	
	R@20	P@20	MRR@20	nDCG@20	Top-1 Acc. (%)	Top-5 Acc. (%)
ResNet50	**.1632**	**.0141**	**.0979**	**.1253**	75.2	92.2
VGG19	**.1398**	**.0124**	**.0750**	**.1008**	71.1	89.8
NASNet Large	.1379	.0120	.0743	.0998	**82.7**	**96.2**
InceptionV3	.1332	.0125	.0744	.1007	78.0	93.9
InceptionResNetV2	.1302	.0117	.0692	.0936	**80.4**	**95.3**
Random	.0172	.0013	.0051	.0093	-	-

Fine-tuning: ¿es necesario?

- !Sí! Ayuda y mucho:

CNN	R@20	P@20	F1@20	MAP@20	MRR@20	nDCG@20
ResNet-deep-fine-tune-ugallery	.1954	.0164	.0276	.0294	.1155	.1476
ResNet-deep-fine-tune-ugallery-only-artist	.1943	.0166	.0279	.0300	.1166	.1493
Omniart-deep-fine-tune-ugallery	.1900	.0159	.0266	.0267	.0973	.1330
ResNet	.1632	.0141	.0235	.0246	.0979	.1253
Omniart-shallow-with-task-weights	.1609	.0134	.0224	.0227	.0879	.1147
ResNet-shallow-fine-tune-ugallery-only-artist	.1501	.0137	.0230	.0242	.0936	.1202
ResNet-shallow-fine-tune-ugallery	.1541	.0138	.0229	.0238	.0942	.1196
ResNet-shallow-fine-tune-ugallery-only-medium	.1541	.0138	.0225	.0238	.0894	.1165
Omniart-shallow-only-type	.1510	.0127	.0212	.0217	.0831	.1092
Omniart-shallow-no-task-weights	.1473	.0129	.0214	.0234	.0906	.1150
Omniart-shallow-only-artist	.1442	.0129	.0213	.0235	.0908	.1153
ResNet-deep-fine-tune-ugallery-only-medium	.1374	.0124	.0204	.0218	.0856	.1101
Omniart-shallow-only-period	.0937	.0081	.0135	.0127	.0514	.0689
Random	.0172	.0013	.0022	.0014	.0051	.0093

Música

- Ejemplo de recomendación de Spotify (2014)
  - Muchos sistemas, incluso a la fecha, representan música con diferentes features manuales, siendo MFCC (Mel Frequency Cepstral Coefficients), los más populares. Se obtienen así:
    - Separar la señal en pequeños tramos.
    - A cada tramo aplicarle la Transformada de Fourier discreta y obtener la potencia espectral de la señal.
    - Aplicar el banco de filtros correspondientes a la Escala Mel al espectro obtenido en el paso anterior y sumar las energías en cada uno de ellos.
    - Tomar el logaritmo de todas las energías de cada frecuencia mel
    - Aplicarle la transformada de coseno discreta a estos logaritmos.

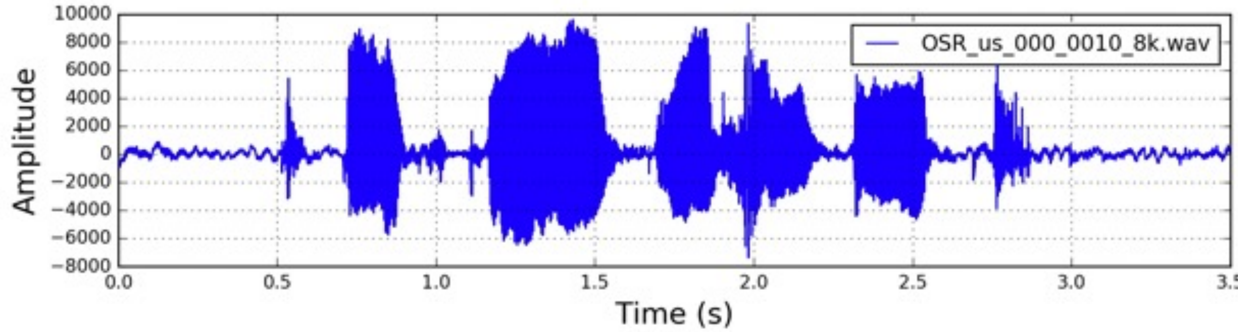
Escala Mel

- La escala Mel es una escala que relaciona la frecuencia percibida de un tono con la frecuencia medida real. Escala la frecuencia para que coincida más con lo que el oído humano puede escuchar (los humanos son mejores para identificar pequeños cambios en el habla a frecuencias más bajas.
- Una frecuencia en Hertz se convierte a escala Mel con:

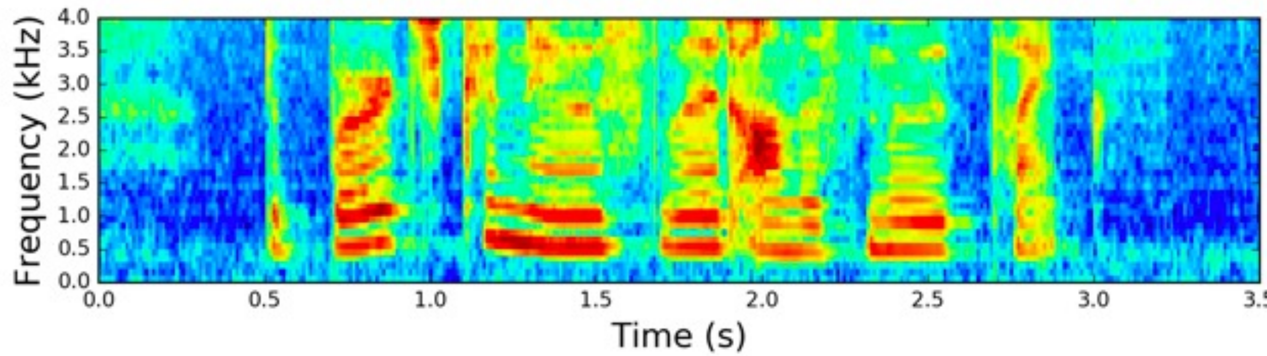
$$\mathrm{Mel}(f) = 2595 \log\left(1 + \frac{f}{700}\right)$$

https://medium.com/prathena/the-dummys-guide-to-mfcc-aceab2450fd

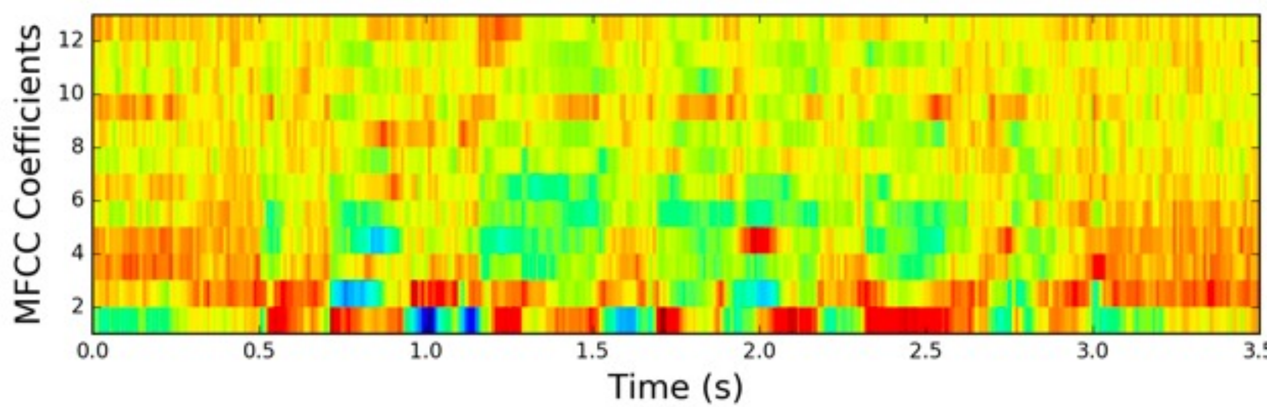
MFCC



Señal de audio



Espectrograma



MFCC

Approach con redes neuronales

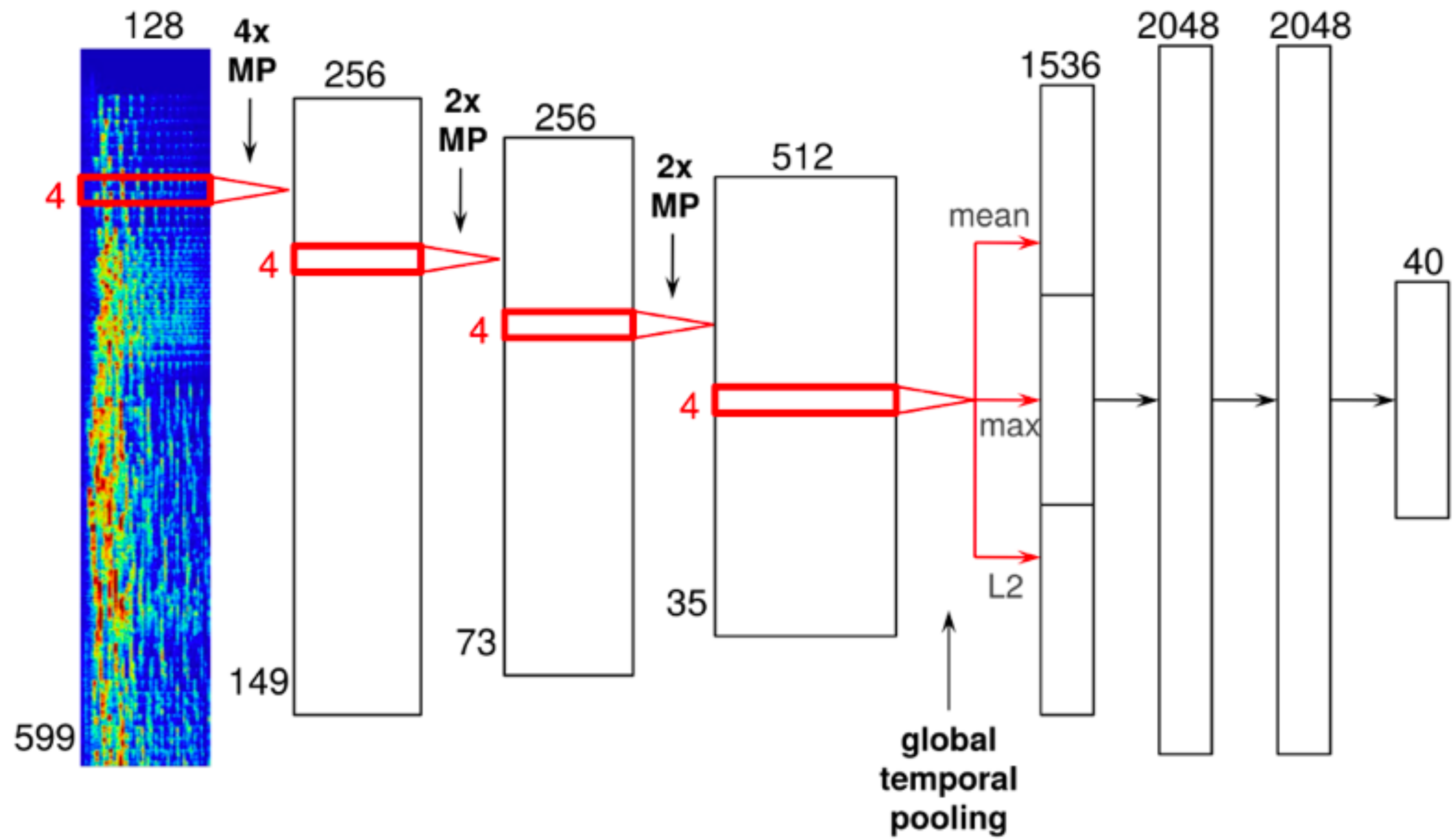
- En Van den Oord et al (2013) comparan un approach tradicional basado en MFCC con DL features.
- El approach tradicional:

- **Extract MFCCs from the audio signals.** We computed 13 MFCCs from windows of 1024 audio frames, corresponding to 23 ms at a sampling rate of 22050 Hz, and a hop size of 512 samples. We also computed first and second order differences, yielding 39 coefficients in total.
- **Vector quantize the MFCCs.** We learned a dictionary of 4000 elements with the K-means algorithm and assigned all MFCC vectors to the closest mean.
- **Aggregate them into a bag-of-words representation.** For every song, we counted how many times each mean was selected. The resulting vector of counts is a bag-of-words feature representation of the song.

Van den Oord, A., Dieleman, S., & Schrauwen, B. (2013). Deep content-based music recommendation. In Advances in neural information processing systems (pp. 2643-2651).

DL en Van den Oord et al (2013)

Clip de 3
segundos



Latent factor prediction

- Como baseline se obtienen factores latentes de usuarios e items usando WMF (ALS)
- La tarea es predecir los factores latentes directamente desde el audio, con los siguientes métodos:
  - Linear regression trained on the bag-of-words representation described in Section 4.1.
  - A multi-layer perceptron (MLP) trained on the same bag-of-words representation.
  - A convolutional neural network trained on log-scaled mel-spectrograms to minimize the mean squared error (MSE) of the predictions.
  - The same convolutional neural network, trained to minimize the weighted prediction error (WPE) from the WMF objective instead.

Latent factor prediction

- Dataset: Million Song Dataset (MSD)
- http://millionsongdataset.com/

Latent factor prediction: resultados

MFCC

Model	mAP	AUC
MLR	0.01801	0.60608
linear regression	0.02389	0.63518
MLP	0.02536	0.64611
CNN with MSE	0.05016	0.70987
CNN with WPE	0.04323	0.70101

Table 2: Results for all considered models on a subset of the dataset containing only the 9,330 most popular songs, and listening data for 20,000 users.

Model	mAP	AUC
random	0.00015	0.49935
linear regression	0.00101	0.64522
CNN with MSE	0.00672	0.77192
upper bound	0.23278	0.96070

Table 3: Results for linear regression on a bag-of-words representation of the audio signals, and a convolutional neural network trained with the MSE objective, on the full dataset (382,410 songs and 1 million users). Also shown are the scores achieved when the latent factor vectors are randomized, and when they are learned from usage data using WMF (upper bound).

Evaluación por muestras

Query	Most similar tracks (WMF)	Most similar tracks (predicted)
Jonas Brothers - Hold On	Jonas Brothers - Games Miley Cyrus - G.N.O. (Girl's Night Out) Miley Cyrus - Girls Just Wanna Have Fun Jonas Brothers - Year 3000 Jonas Brothers - BB Good	Jonas Brothers - Video Girl Jonas Brothers - Games New Found Glory - My Friends Over You My Chemical Romance - Thank You For The Venom My Chemical Romance - Teenagers
Beyoncé - Speechless	Beyoncé - Gift From Virgo Beyonce - Daddy Rihanna / J-Status - Crazy Little Thing Called Love Beyoncé - Dangerously In Love Rihanna - Haunted	Daniel Bedingfield - If You're Not The One Rihanna - Haunted Alejandro Sanz - Siempre Es De Noche Madonna - Miles Away Lil Wayne / Shanell - American Star
Coldplay - I Ran Away	Coldplay - Careful Where You Stand Coldplay - The Goldrush Coldplay - X & Y Coldplay - Square One Jonas Brothers - BB Good	Arcade Fire - Keep The Car Running M83 - You Appearing Angus & Julia Stone - Hollywood Bon Iver - Creature Fear Coldplay - The Goldrush
Daft Punk - Rock'n Roll	Daft Punk - Short Circuit Daft Punk - Nightvision Daft Punk - Too Long (Gonzales Version) Daft Punk - Aerodynamite Daft Punk - One More Time / Aerodynamic	Boys Noize - Shine Shine Boys Noize - Lava Lava Flying Lotus - Pet Monster Shotglass LCD Soundsystem - One Touch Justice - One Minute To Midnight

Table 4: A few songs and their closest matches in terms of usage patterns, using latent factors obtained with WMF and using latent factors predicted by a convolutional neural network.

Factores latentes predichos por CNN

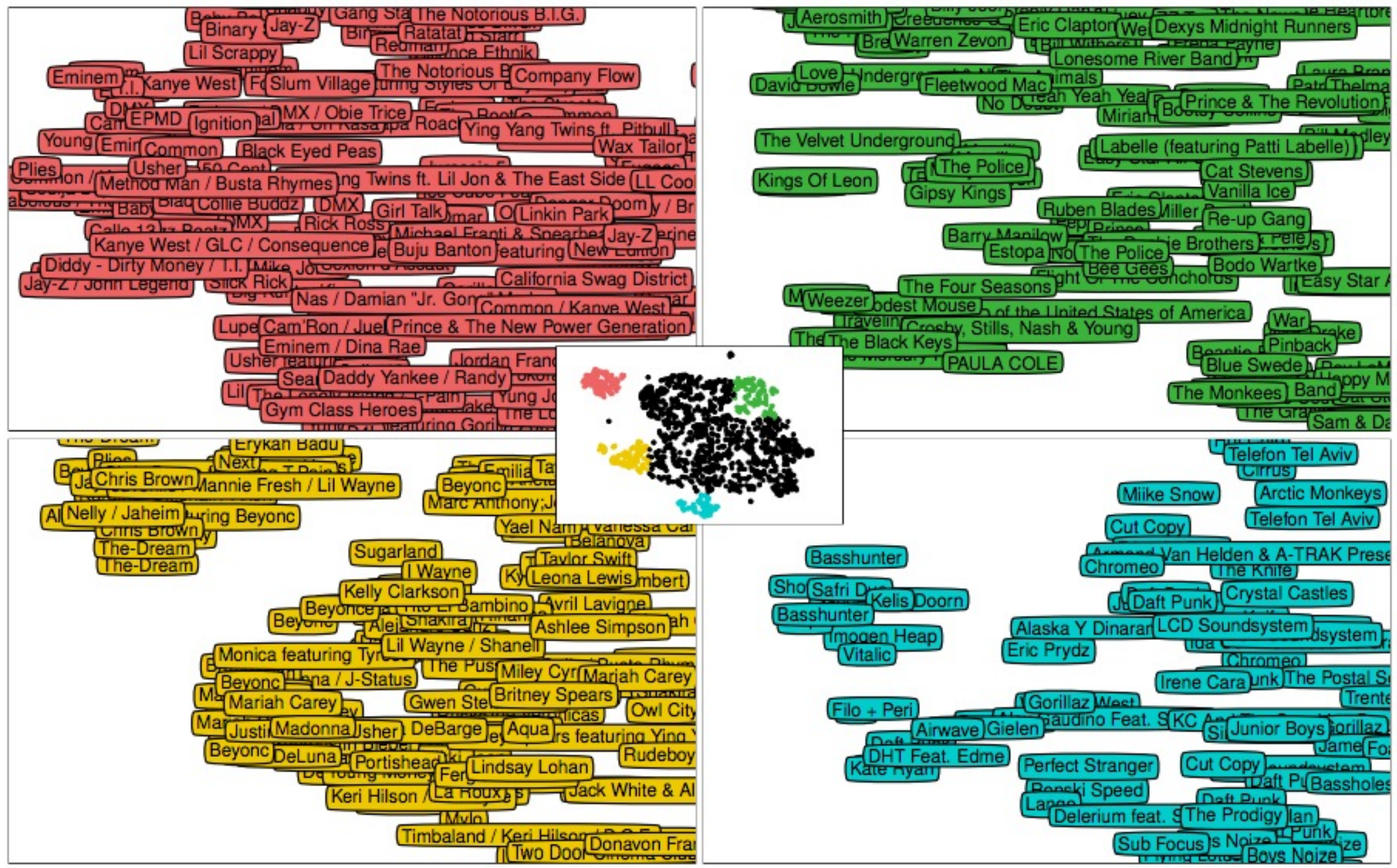


Figure 1: t-SNE visualization of the distribution of predicted usage patterns, using latent factors predicted from audio. A few close-ups show artists whose songs are projected in specific areas. We can discern hip-hop (red), rock (green), pop (yellow) and electronic music (blue). This figure is best viewed in color.

Resumen

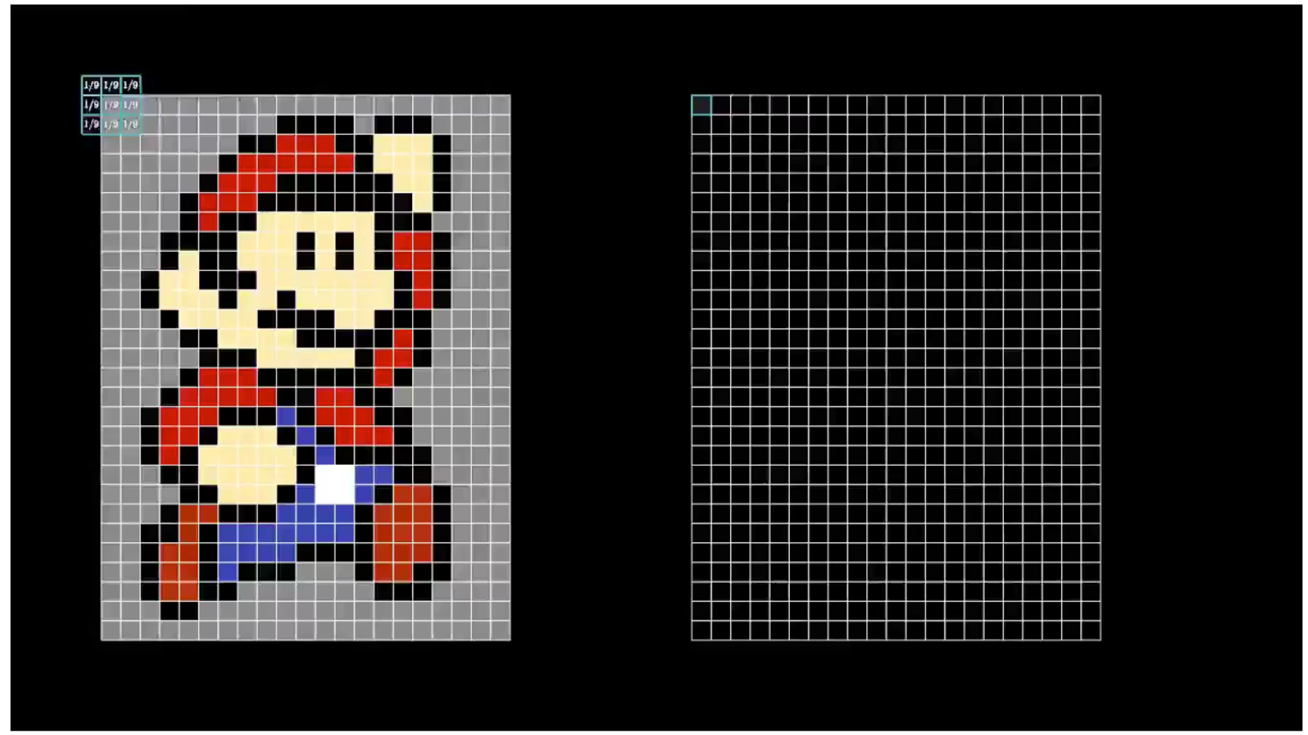
- Para imágenes y música, usar features aprendidas por modelos de deep learning puede ser muy útil y evita el costo de la “ingeniería de características manual”

- Una debilidad de este approach es que las características aprendidas son difícilmente explicables.

Gracias!

- dparra@ing.puc.cl

¿Qué hacen las convoluciones?



https://twitter.com/3blue1brown/status/1303489896519139328