CommerceMM: Large-Scale Commerce MultiModal Representation Learning with Omni Retrieval

Licheng Yu, Jun Chen, Animesh Sinha, Mengjiao Wang, Yu Chen, Tamara L. Berg, Ning Zhang

Introduction

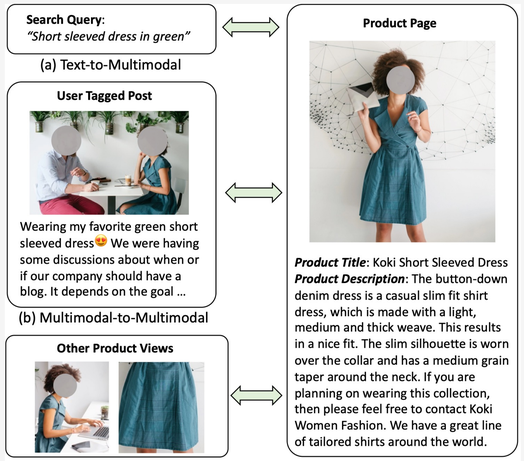
At Meta, nearly every post related to commerce is multimodal.
- Marketplace post is made of one or several views of a product with its product description
- Shop product listing is made of the product images and detailed specifics describing the product, e.g., title, attribute, size, material, etc.
- Influencers upload their fashion posts to Instagram with captions and hashtags.

We introduce **Commerce** **M**ulti**M**odal Representation (**CommerceMM**), a large-scale pre-trained model for joint multimodal commerce embedding at Facebook.

We follow the pre-training + fine-tuning training regime. Our pre-training data is made of:
- 50M Catalog Posts
- 52M Marketplace Posts
- 50M *Cross-Modal Cross-Pair* Data
  - IG and FB Shops text search queries with clicked product.
  - IG and FB posts where a product is tagged on the post.

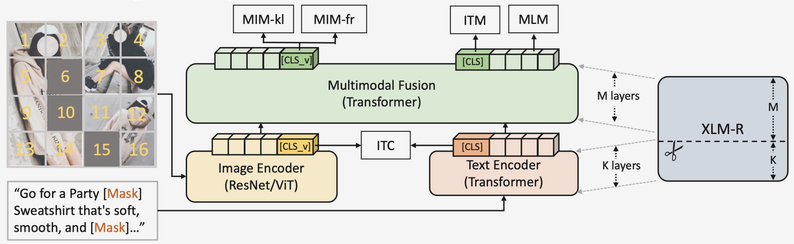
Cross-modal and Cross-pair Data

Users could use text query to do product search. Some users tag the relevant products when uploading their multimodal media. On product page, there could be multiple views of products. While those medias are of different type (text, multimodal, image), they are linked with the same product.



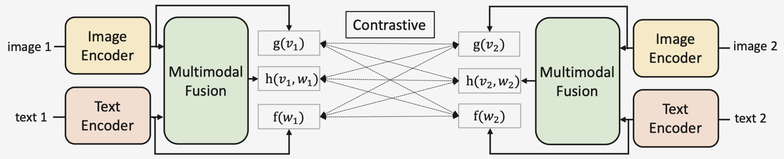
Pre-training and Fine-tuning

Image-Text Pre-training + Modality Randomization



1) Masked Language Modeling (MLM)
2) Masked Image Modeling – Feature Regression (MIM-fr)
3) Masked Image Modeling – KL-Divergence (MIM-kl)
4) Image-Text Contrastive Learning (ITC)
5) Image-Text Matching (ITM)

Omni Retrieval



Assumption: If a source pair is linked with a target pair, we assume any existing modality from the source would be highly correlated with any existing modality from the target.

We formulate the ***Cross-Modal Cross-Pair*** data as Source pair: (image 1, text 1) and Target pair: (image 2, text 2).

$$\text{Text: } \begin{cases} s_{ij}^{T\leftrightarrow T} = f(\mathbf{w}_i)^T f(\mathbf{w}_j) \\ s_{ij}^{T\leftrightarrow I} = f(\mathbf{w}_i)^T g(\mathbf{v}_j) \\ s_{ij}^{T\leftrightarrow M} = f(\mathbf{w}_i)^T h(\mathbf{w}_j, \mathbf{v}_j), \end{cases} \quad \text{Image: } \begin{cases} s_{ij}^{I\leftrightarrow T} = g(\mathbf{v}_i)^T f(\mathbf{w}_j) \\ s_{ij}^{I\leftrightarrow I} = g(\mathbf{v}_i)^T g(\mathbf{v}_j) \\ s_{ij}^{I\leftrightarrow M} = g(\mathbf{v}_i)^T h(\mathbf{w}_j, \mathbf{v}_j), \end{cases}$$

$$\text{Multimodal: } \begin{cases} s_{ij}^{M\leftrightarrow T} = h(\mathbf{w}_i, \mathbf{v}_i)^T f(\mathbf{w}_j) \\ s_{ij}^{M\leftrightarrow I} = h(\mathbf{w}_i, \mathbf{v}_i)^T g(\mathbf{v}_j) \\ s_{ij}^{M\leftrightarrow M} = h(\mathbf{w}_i, \mathbf{v}_i)^T h(\mathbf{w}_j, \mathbf{v}_j). \end{cases}$$

$$\mathcal{L}_{omni} = \sum_{u,v\in\{I,T,M\}} \sum_i \delta_i^u [\log \frac{\exp(s_{ii}^{u\leftrightarrow v}/\tau)}{\sum_j \exp(s_{ij}^{u\leftrightarrow v}/\tau)} + \log \frac{\exp(s_{ii}^{u\leftrightarrow v}/\tau)}{\sum_j \exp(s_{ji}^{u\leftrightarrow v}/\tau)}],$$

Downstream Tasks

1) Catalog Product Categorization
2) Marketplace Product Categorization
3) Image-to-Text Retrieval
4) Text-to-Image Retrieval
5) Query-to-Product Retrieval
6) Image-to-Product Retrieval
7) Image-to-Image Retrieval

Task 1&2: Product Categorization

Task 4: Text-to-Image Retrieval

Task 6: Image-to-Product Retrieval

Experiments

Ablation Study on Pre-training

	Pre-training Tasks	Meta Avg.	CC	MPC	T2I	I2T	Q2P	I2P	I2P^i
1	None	50.39	72.08	63.75	22.70	23.88	48.43	55.80	66.10
2	MLM	52.77	73.10	67.94	24.87	25.84	51.59	60.46	65.59
3	MIM-kl	53.59	73.26	69.04	26.31	26.91	53.89	59.18	66.54
4	MIM-kl + MIM-fr	54.18	73.27	69.12	27.88	28.61	54.05	59.48	66.83
5	MLM + MIM-kl + MIM-fr	54.19	73.64	69.55	26.66	26.98	53.47	61.66	67.30
6	MLM + MIM-kl + MIM-fr + ITM	54.62	73.64	69.45	27.82	28.33	54.63	61.63	66.87
7	MLM + MIM-kl + MIM-fr + ITM + ITC	57.87	73.76	69.61	39.11	40.30	55.60	60.03	66.65
8	Omni Retrieval (Omni)	56.27	72.98	67.81	29.69	30.78	57.34	67.98	67.31
9	MLM + MIM-kl + MIM-fr + ITM + ITC + Omni	60.64	73.77	69.73	42.05	43.06	58.48	69.20	68.16

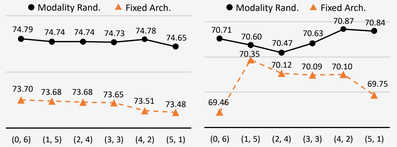
Ablation Study of different pre-training tasks on different downstream tasks, including Catalog Categorization (CC), Marketplace Categorization (MPC), Text-to-Image Retrieval (T2I), Image-to-Text Retrieval (I2T), Query-to-Product Retrieval (Q2P), Image-to-Product Retrieval (I2P), and Image-to-Product-Image Retrieval (I2P^i). Meta Average is the average score of 7 tasks.

Transferability

	FashionGen T2I			FashionGen I2T		
	R1	R5	R10	R1	R5	R10
FashionBERT [11]	26.8	46.5	55.7	24.0	46.3	52.1
KaleidoBERT [51]	33.9	60.6	68.6	28.0	60.1	68.4
CommerceMM (small)	**39.6**	**61.5**	**72.7**	**41.6**	**64.0**	**72.8**

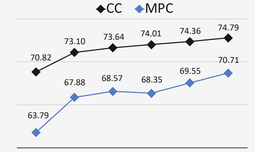
Image-Text Retrieval on the academic dataset – FashionGen.

Modality Randomization



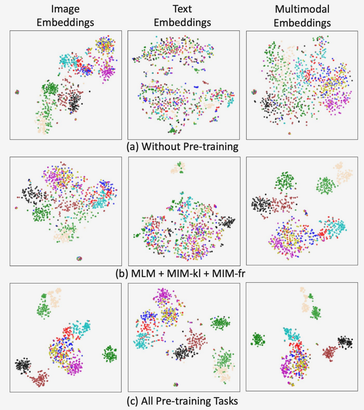
Effect of (K, M) from a fixed-architecture (3, 3) pre-trained model and modality-randomized pre-trained model. (K, M) stands for (#text layers, #multimodal layers).

Model Size



Effect of total number of layers from a modality randomized model evaluated on CC and MPC after fine- tuning. (K, M) stands for (#text layers, #multimodal layers). K=0 means early-fusion model without text encoder.

Visualization



T-SNE of different pre-trained models' embeddings.

Attention

Visualization of text-to-image attention from different pre-trained models.