* Sds
* Dsds
* [Modern Recommender Systems. - A Deep Dive into the AI algorithms… | by Maximilian Beckers | Towards Data Science](https://towardsdatascience.com/modern-recommender-systems-a0c727609aa8)

DLRM (Deep Learning Recommendation Model) is a neural network architecture designed for building recommendation systems. It was introduced by Facebook AI Research (FAIR) and has gained popularity for its effectiveness in handling large-scale recommendation tasks. DLRM combines both deep learning and factorization machine techniques to capture complex user-item interactions and make accurate recommendations.

Here is a high-level overview of the DLRM architecture:

Input Layer: DLRM takes as input the user and item features. These features could include categorical variables like user demographics, item genres, or numerical variables like ratings, timestamps, etc.

Embedding Layer: Categorical features are typically embedded into low-dimensional continuous vectors using embedding tables. Embeddings capture the latent representation of categorical variables, allowing the model to learn the interactions between different feature combinations.

Bottom MLP: The embedded features are passed through a stack of fully connected layers known as the bottom Multi-Layer Perceptron (MLP). The bottom MLP learns non-linear transformations of the input embeddings, introducing flexibility in modeling complex feature interactions.

Top MLP: The outputs from the bottom MLP are concatenated and passed through another stack of fully connected layers called the top MLP. The top MLP captures high-level interactions between the user and item features and generates a joint representation.

Interaction Layer: The output of the bottom MLP is also combined with the dot product of the embedding vectors to capture both deep and shallow interactions. This interaction layer allows DLRM to model second-order feature interactions using factorization machines.

Prediction Layer: Finally, the joint representation obtained from the top MLP and the interaction layer is fed into the prediction layer. The prediction layer typically consists of a single output unit with a sigmoid activation function, which predicts the probability of a user engaging with an item (e.g., clicking, purchasing, etc.).

During training, DLRM uses binary cross-entropy loss to compare the predicted probabilities with the actual user-item interactions. The model parameters are optimized using backpropagation and stochastic gradient descent (SGD) or other optimization algorithms.

DLRM offers a flexible and scalable architecture for recommendation systems. It can handle large-scale datasets, capture complex feature interactions, and effectively learn user preferences to provide personalized recommendations. The specific implementation details and hyperparameters of DLRM can vary based on the requirements and characteristics of the recommendation task at hand.