# A multi-modal deep neural network for predicting clothing demand

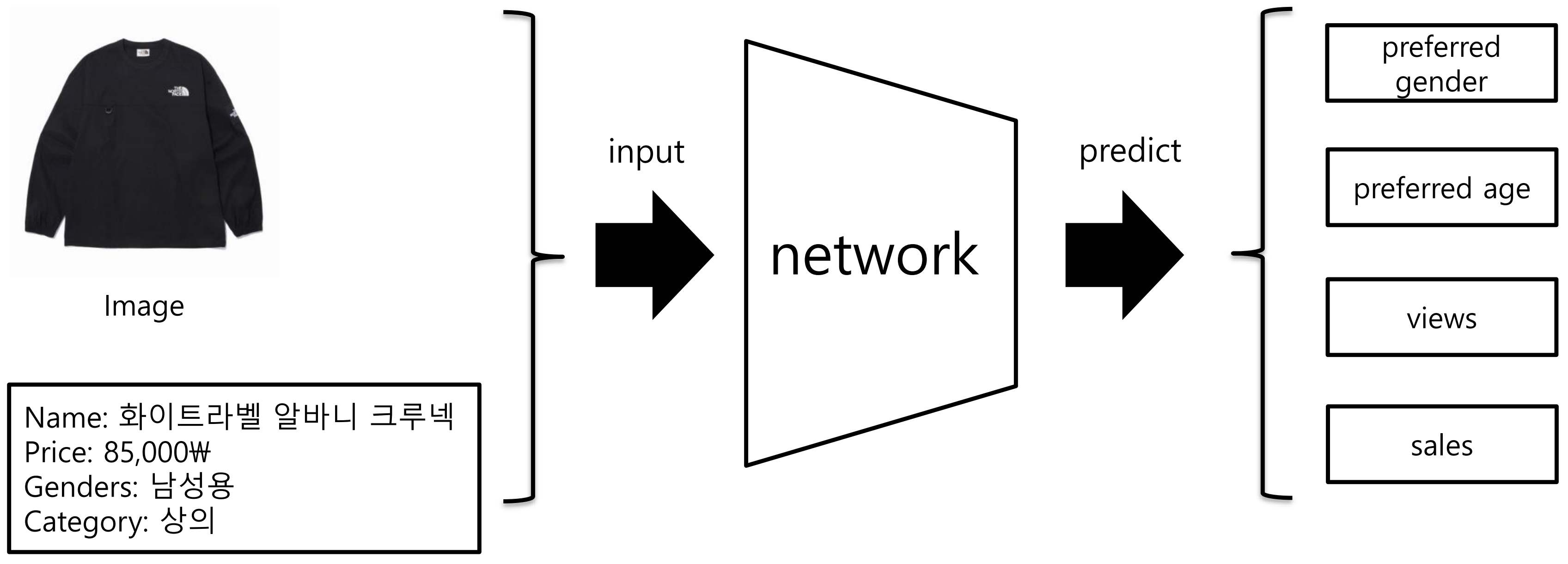
VISION MODELING LAB

Dongjoo Kim and Minsik Lee

Department of Applied Artificial Intelligence,
Department of Electrical Engineering,
Hanyang University

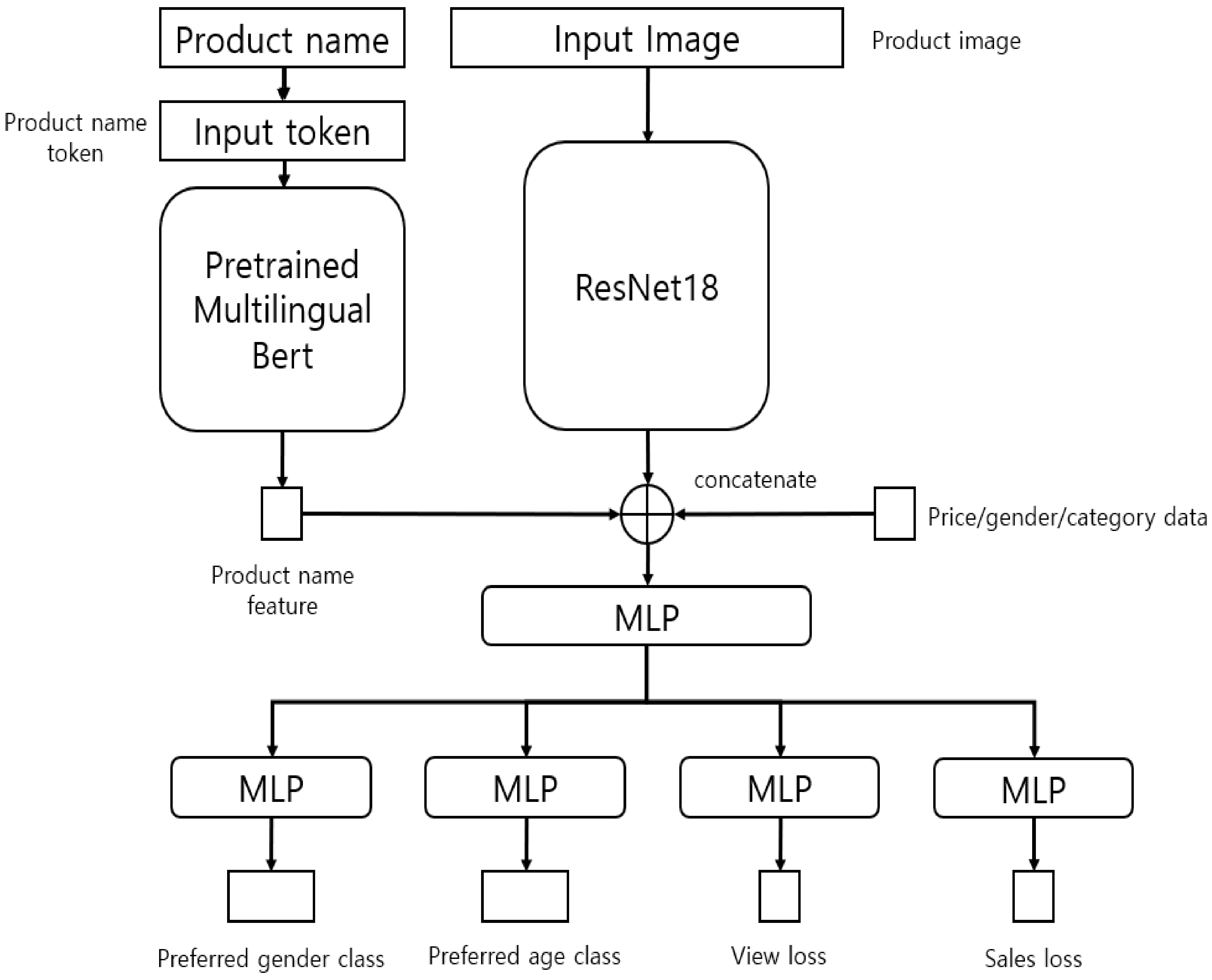
## INTRODUCTION

This paper proposes a deep neural network to predict the quantitative demand information of clothes displayed in an online shopping mall, using images and other miscellaneous information. In this paper, we construct a clothing demand evaluation database with clothing images collected from an online shopping mall, along with product names, prices, genders, preferred genders, preferred age groups, product numbers, cumulative sales, views, and product categories. Various modalities are combined in the proposed deep neural network to predict the demand information. To process the images and the text information (product names), we employ ResNet18 and the Multilingual BERT, respectively, in the proposed network. The experimental results show that the proposed multi-modal network shows reliable performance in predicting the demand of clothing.



Name: 화이트라벨 알바니 크루넥
Price: 85,000₩
Genders: 남성용
Category: 상의

## Clothing demand information predicting network



- Network predicts preferred genders, preferred age groups, views, cumulative sales using clothing image, product name, product price, genders, product numbers, and product categories.
- It has a Resnet18, pretrained Multilingual Bert, and two fully connected layers.
- Resnet is trained until the end of network training, and Bert is not additionally trained after performing feature extraction.
- The four pieces of numerical information except for clothing image are concatenated immediately before fully connected layer.
- Data augmentation was performed on the clothing image.
- Due to the imbalanced class problem, focal loss was applied on the preferred age class.

## Clothing demand information predicting database

- Constructed a database using web scraping from MUSINSA web site.
- Inputs: clothing image, product name, product price, genders, product numbers, product categories
  - Scrapped from MUSINSA web site.
- Labels: preferred genders, preferred age groups, views, cumulative sales
  - Scrapped from MUSINSA web site also.
  - Network predicts clothing demand information.

## EXPERIMENTS

| Network | Preferred gender | Preferred age | Views | Sales |
|---|---|---|---|---|
| Proposed | 84.5% | 73.7% | 0.092 | 0.054 |
| No word | 83.9% | 73.5% | 0.09 | 0.058 |

- In order to solve overfitting problem, we applied dropout to each fully connected layers with a probability of 0.5.
- Data augmentation was performed by rotation, changing brightness, contrast, and random saturation, and rotation was 90 or 180 degrees.
- The optimizer used for network learning was Adam optimizer.
- The loss was the cross-entropy loss for preferred gender and the focal loss for preferred age class, and the mse loss for views and sales classes.
- The following total loss was defined as

$$L_{total} = CE_{gender} + FL_{age} + \alpha MSE_{view} + \beta MSE_{sales}$$

- $\alpha$ and $\beta$ are hyperparameters that determine the effect of each term on the total loss.($\alpha = 0.01, \beta = 0.01$)
- The learning rate was set to 0.0001 for 3,000 epoch.
- The experimental results show that the proposed network is useful for analyzing demand information for clothing by using the image and miscellaneous information.
- The “No word” method indicates that the word features (product name) were not used.
- The ablation study shows that the proposed network using word features shows better performance in other classes except for the views class.
- Although the proposed method achieved slightly lower performance in the views class, the proposed method is more effective because it achieves relatively higher performance improvement in other indicators.