Multimodal Machine Learning

Kristýna Němcová

FI MU

Multimodal machine learning

Multimodal machine learning seeks to understand the problem in terms of multiple modalities. Modality is the way in which something happens or is experienced. Humans can distinguish for example the modality of touch, sight, hearing, smell, taste. Nonetheless, the modalities used in machine learning are not just sensory modalities. Most used are text, speech, images and videos.

There are multiple challenges in the field of multimodal learning. The main ones are: representation, translation, alignment, fusion and co-learning. The focus of our research is fusion.

Fusion

Fusion is most commonly used in the field of multimodal machine learning. It focuses on prediction while joining data from multiple modalities. It provides more robust predictions as it does not rely on a single point of view. Another advantage is that fusion can deal with missing data in modalities.

Early fusion: In early fusion, the features are combined into a single vector and classified together.

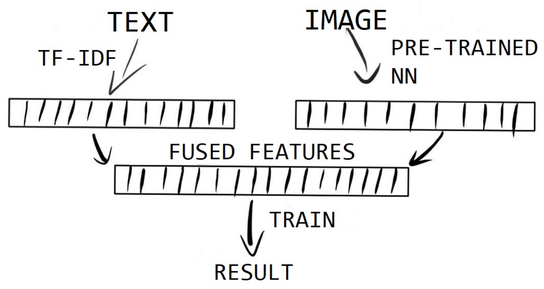

Figure 1:Early fusion

Late fusion: In late fusion, the model is trained for each modality and the results are combined. We can combine them by averaging, voting or a learned model.

Hybrid fusion: The hybrid fusion model combines the benefits of early and late fusion. It is trained for every modality plus early fusion. And the results are then combined.

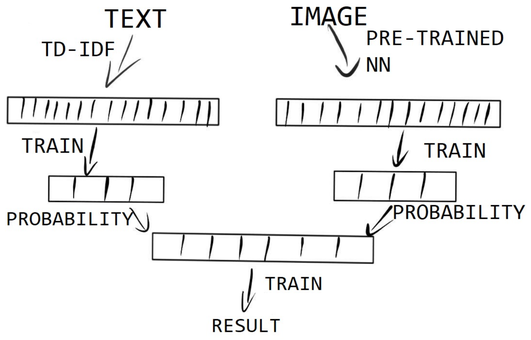

Figure 2:Late fusion

Experiments

We processed the text and image data as described in the dataset section, resulting in vector of size 1108 for text and 1000 for image. The adoption speed class is 1, 2, 3, 4. Train test split of 70/30.

So far, the best results for single modality are random forests. The accuracy of 0.389 was achieved for text and 0.358 for image. The issue is overfitting as it learns the entire training domain. Another promising results are from simple neural networks. Image data accuracy 0.315. For text data it proved to be too weak, reaching accuracy of 0.28 with poor distribution.

Early fusion with random forest did not exceed single modality text prediction (acc: 0.367). There was visible improvement with neural network. Accuracy of 0.325 surpassed both predictions of single modality.

Late fusion is the next logical step as we can see that different methods of learning are suitable for different modalities. With late fusion, we can take a model for text as random forest and model for image as NN. The late fusion model was able to achieve accuracy of 0.328. Which both outperforms the early fusion with NN and does not have issues with overfitting as random forest has.

	text	image	text&image
Random forest	0.389	0.358	0.367
Neural Network	0.28	0.315	0.325
Late fusion	-	-	0.328

Datset - Adoption Prediction

For the experimentation, we found dataset PetFinder.my Adoption Prediction. It consists of real data from Malaysia's adoption webside. There is 15 000 examples with text, tabular, and image data. The goal is to predict the speed at which the pet is adopted.

There is a lot of information provided in the dataset. However, for the purposes of trying out methods of multimodal machine learning we chose to work with texts and images only.

In this classification task is predicted adoption speed. It shows how quickly is the animal adopted. The 5 classes were somewhat balanced except one with almost no examples. We merged class 0 with class 1 to avoid the issue, creating 4 classes.

Textual data were processed, stemmed and vectorized via tf-idf. The resulting vector had over 100 000 items so only 1108 were selected via variance threshold.

Images were put into pre-trained neural network. The resulting vector of 1000 features is our input.

Data features

Some of the features listed:

- Adoption speed - class 0-4, the higher the longer the pet had to wait
- Description - short text describing the pet
- Photo - how many photos were uploaded
- Type - dog or cat
- Age
- Breed
- Color
- Size
- Health
- Fur length

Adoption speed

- 0 - adopted the same day
- 1 - adopted in the 1st week
- 2 - adopted in the 1st month
- 3 - adopted 2nd or 3rd month
- 4 - didn't get adopted

Example

Description:
Curly is a little trooper! A little shy at first she will soon be clambering all over you to play. She loves running around the house as all kittens do and has the sweetest personality. Curly and her siblings will be ready for adoption in one month. To reserve or more info please whatsapp.
Adoption Speed: **3** - Adopted between 31 and 90 days after being listed
Images: