



CS 329P : Practical Machine Learning (2021 Fall)

13.1 Multimodal Data

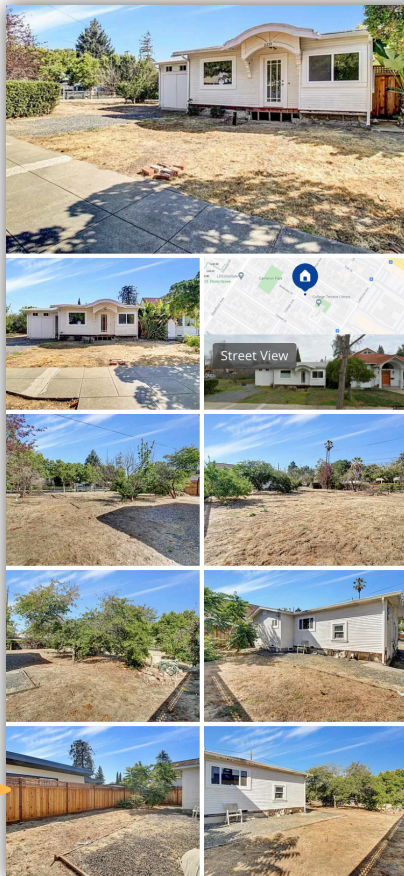
Qingqing Huang, Mu Li, Alex Smola

<https://c.d2l.ai/stanford-cs329p>

Multimodal Data

- Data is naturally multimodal in industry applications
- The raw data contain tables, texts, images, audios, graphs, ...
- E.g. house sales

Images



8,873 Square Feet
2239 Wellesley St, Palo Alto, CA 94306
● **Sold: \$3,395,000** | Sold on 08/30/21 | Zestima
Est. refi payment: \$14,616/mo [Refinance your](#)
Home value Owner tools [Home details](#) Neigh

Tabular

Overview

Note: This property is not currently for sale or for rent on Zillow. The description and property data below may've been provided by a third party, the homeowner or public records.

Huge 8,874 sq ft lot in college terrace close to Stanford Campus, California Ave shops and Library. Existing dwelling has foundation issues and "Red Tagged" by the city. Per the city, one can build a single family home with a 1,000SF ADU and a 500SF Jr. ADU. Yes three separate units that can accommodate street access for each. Or, repair the existing 1,000 sf home as an ADU and build a new home next to it.

Facts and fea

[Edit](#)

Type: Vacan **ng:** Forced air, Electric, Gas
 Year built: 1922 **Cooling:** None
 Parking: Garage - Attached

Text

Utilities / Green Energy Details

Amazon Product



Images + text +
tabular

Graph

Text, images,
videos



Dog Man: Grime and Punishment: A Graphic Novel
(Dog Man #9): From the Creator of Captain
Underpants (9) Hardcover – Illustrated, September 1,
2020

by Dav Pilkey (Author, Illustrator)
★★★★★ 36,089 ratings

#1 Best Seller in Children's Animal Comics & Graphic Novels

See all formats and editions

Hardcover
\$6.48 -prime
71 Used from \$2.39
65 New from \$3.45
3 Collectible from \$3.99

Frequently bought together



+



+



Total price: **\$19.53**

Add all three to Cart

- ☒ **This item:** Dog Man: Grime and Punishment: A Graphic Novel (Dog Man #9): From the Creator of Captain Underpant...
- ☒ Dog Man: Mothering Heights: A Graphic Novel (Dog Man #10): From the Creator of Captain Underpants (10) by Dav Pilkey
- ☒ Dog Man: For Whom the Ball Rolls: From the Creator of Captain Underpants (Dog Man #7) by Dav Pilkey Hardcover **\$6.43**



B. A. Taylor

★★★★★ **Christmas present**

Reviewed in the United Kingdom on September 3, 2020

Verified Purchase



This is a Christmas present for my grandson, he loves the author, pity he has to wait another 4 months to read it.

Nice hard back book, with comic strip type stories.

I don't mind that it's in comic book format, just as long as he's enjoying reading.

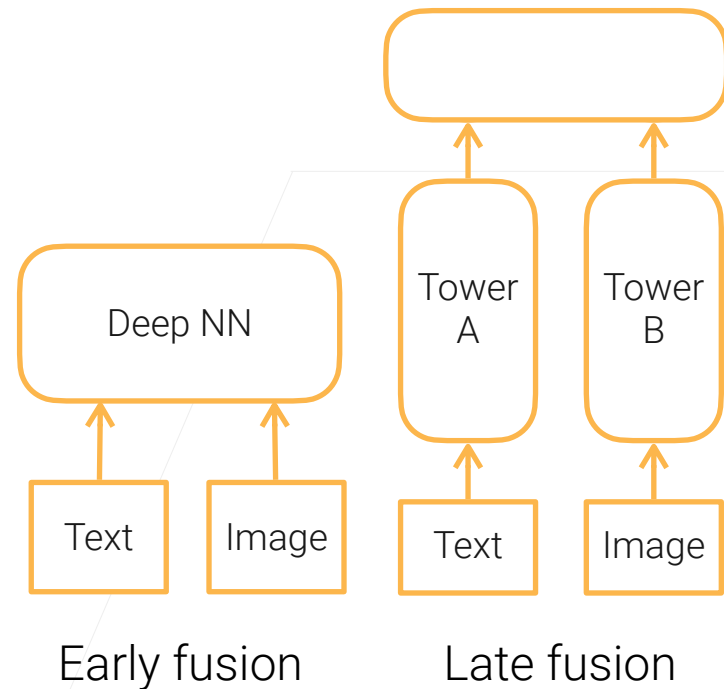
Self-Driving Cars



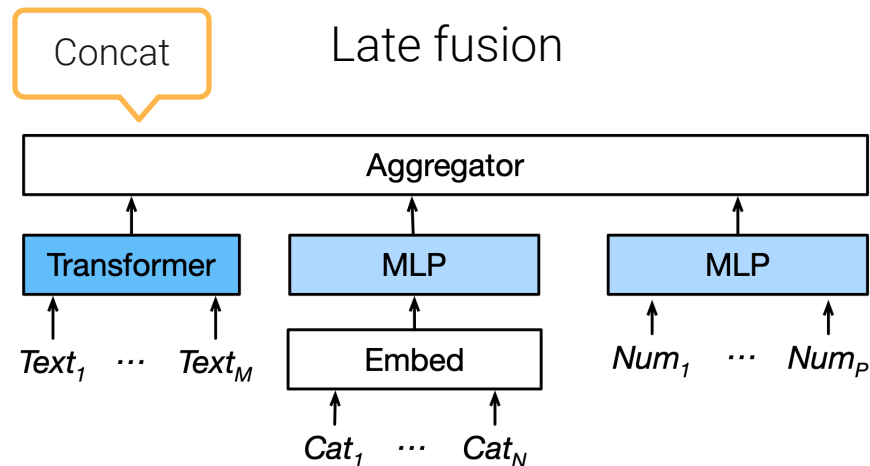
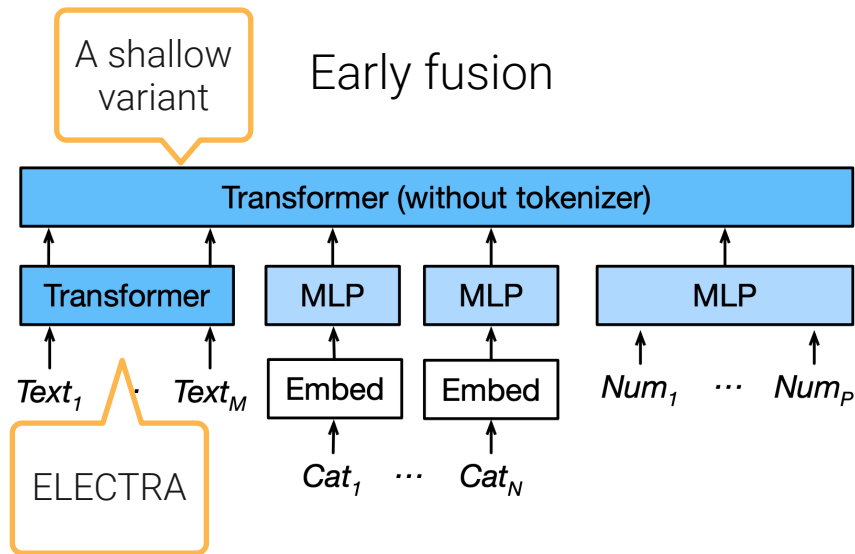
Multimodal Learning



- How to match different modal data into the same semantic space
 - Early vs late fusion
- How to construct loss
 - Combined to learn predict labels
 - Contrastive learning: embedding space in which similar sample pairs stay close to each other while dissimilar ones are far apart



Early vs Late fusion on Tabular + Text



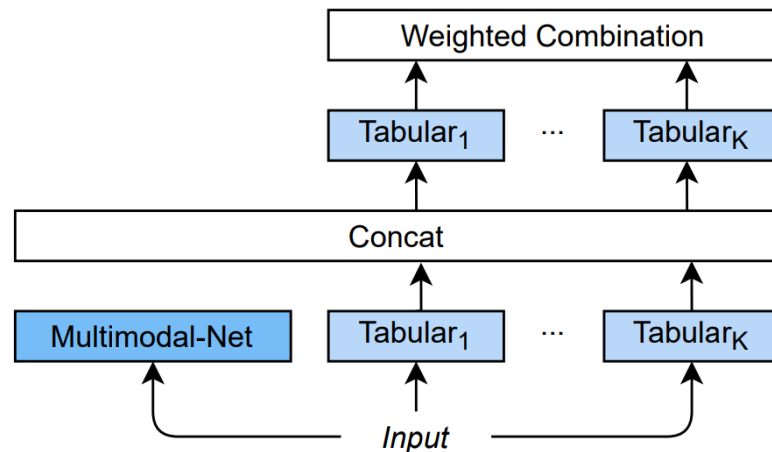
Shi et.al., NeurIPS'21

- Averaged scores on 13 datasets
 - Early fusion: 0.662, late fusion: **0.667** (the larger the better)

Model Ensemble

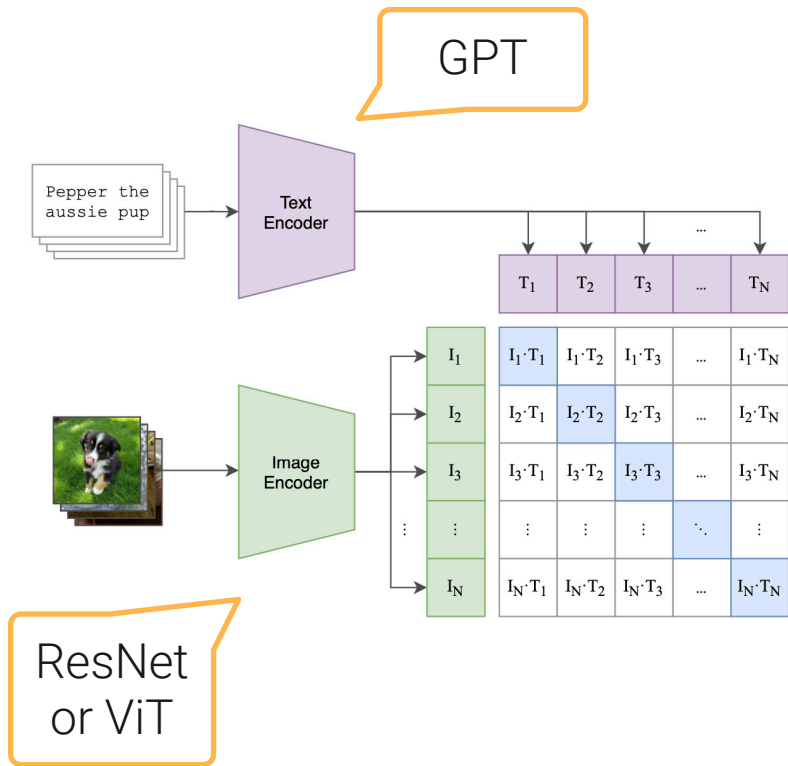


- Stack with other base models
- Averaged scores on 13 datasets
 - Multimodel-Net alone: 0.667
 - Stacked with Multimodel-Net: **0.683**
 - AutoGluon (N-gram for text): 0.659
 - H2O (Word2vec for text): 0.600



Shi et.al., NeurIPS'21

Image Representations from Text Supervision



[CLIP. Radford et.al. ICML'21](#)

```
# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

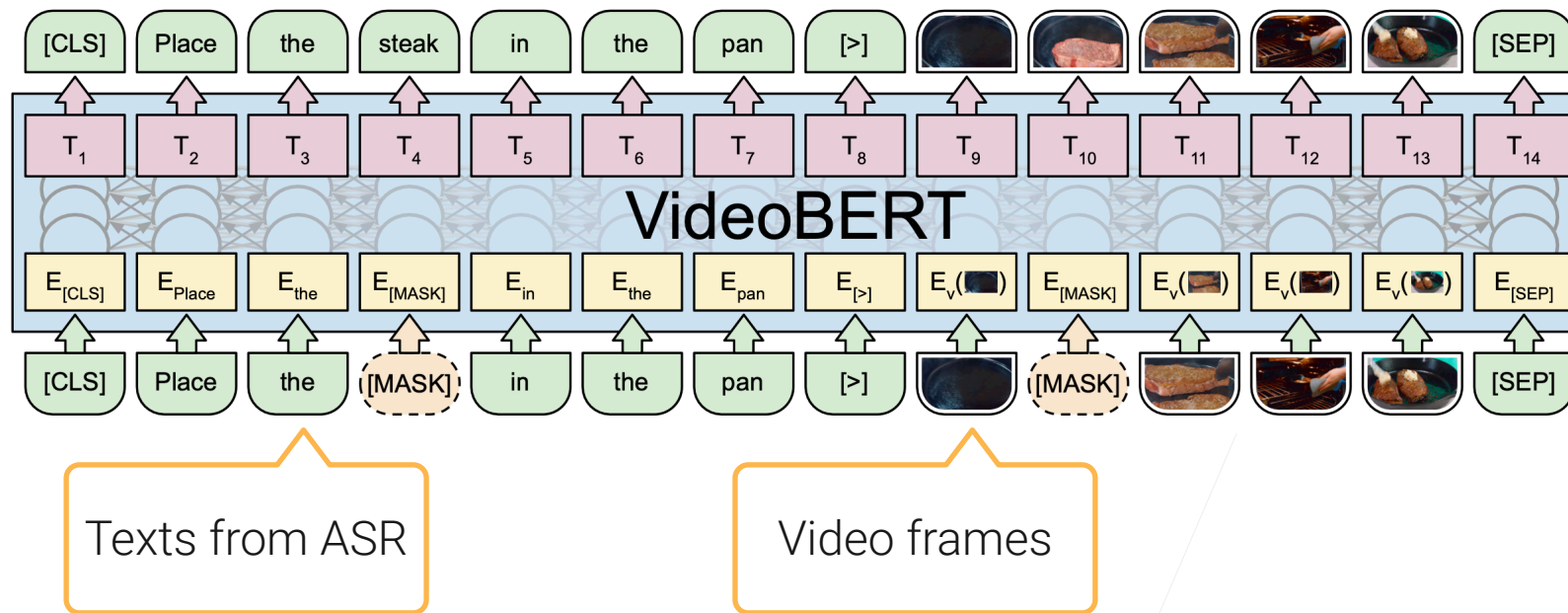
# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

- Trained on 300M (image, text) pairs
- Comparative/better features compared to trained on ImageNet

VideoBERT: Video + Audio



[Sun et.al. ICCV'19](#)

- Trained on cooking/recipe 23K hours Youtube videos

Summary



- Real data is often multi-modal
- Project each modal data into a common space via early or late fusion
- Joint learning labels or contrastive learning for self-supervised training