Lecture 8: Multimodal data, multimodal models, weakly and self-supervised learning

Announcements

- A2 due next Tue Nov 1
- Midterm Mon Nov 7 **in-class**
    - 80 minutes
    - 1 page 8.5'' x 11'' of notes allowed (back and front)
    - No calculators allowed or needed
    - Covers material through "Genomics: Introduction"
    - Practice midterm will be released about a week before the midterm

Today

- Multimodal data and models
- Weakly and self-supervised learning

Multimodal data

Can be very similar, e.g. different image acquisition variants

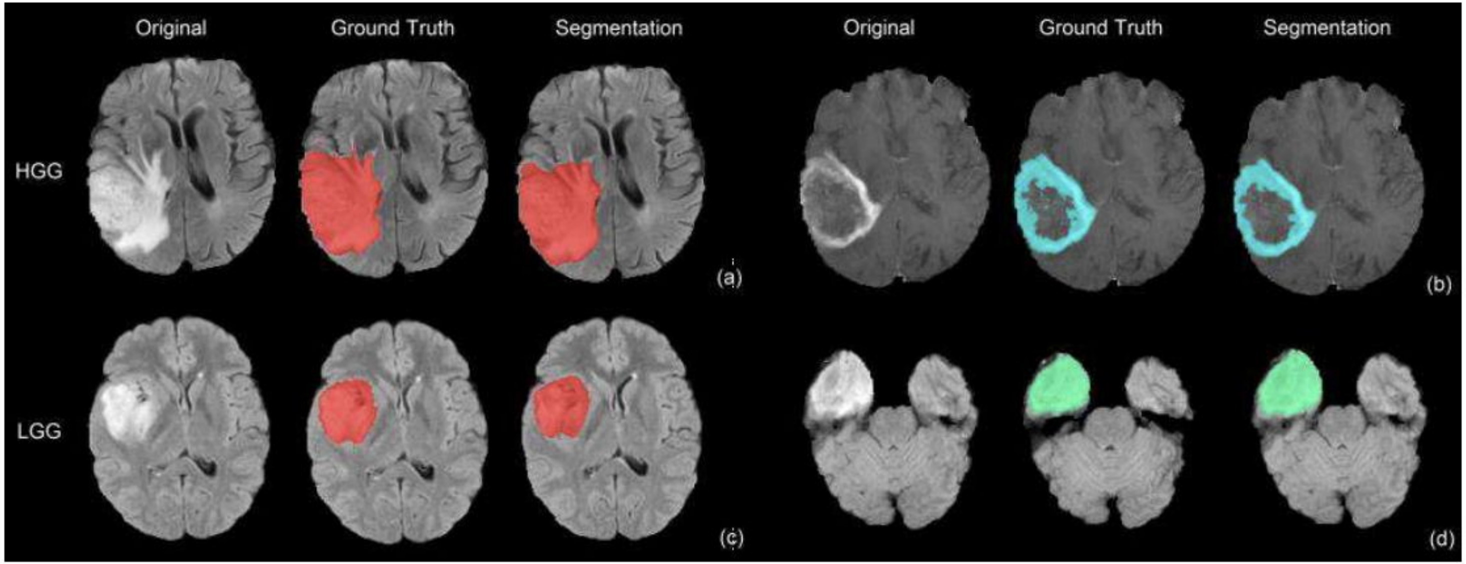


Figure credit: Dong et al. MIUA, 2017.

Multimodal data

Or very different, e.g. different types of clinical data

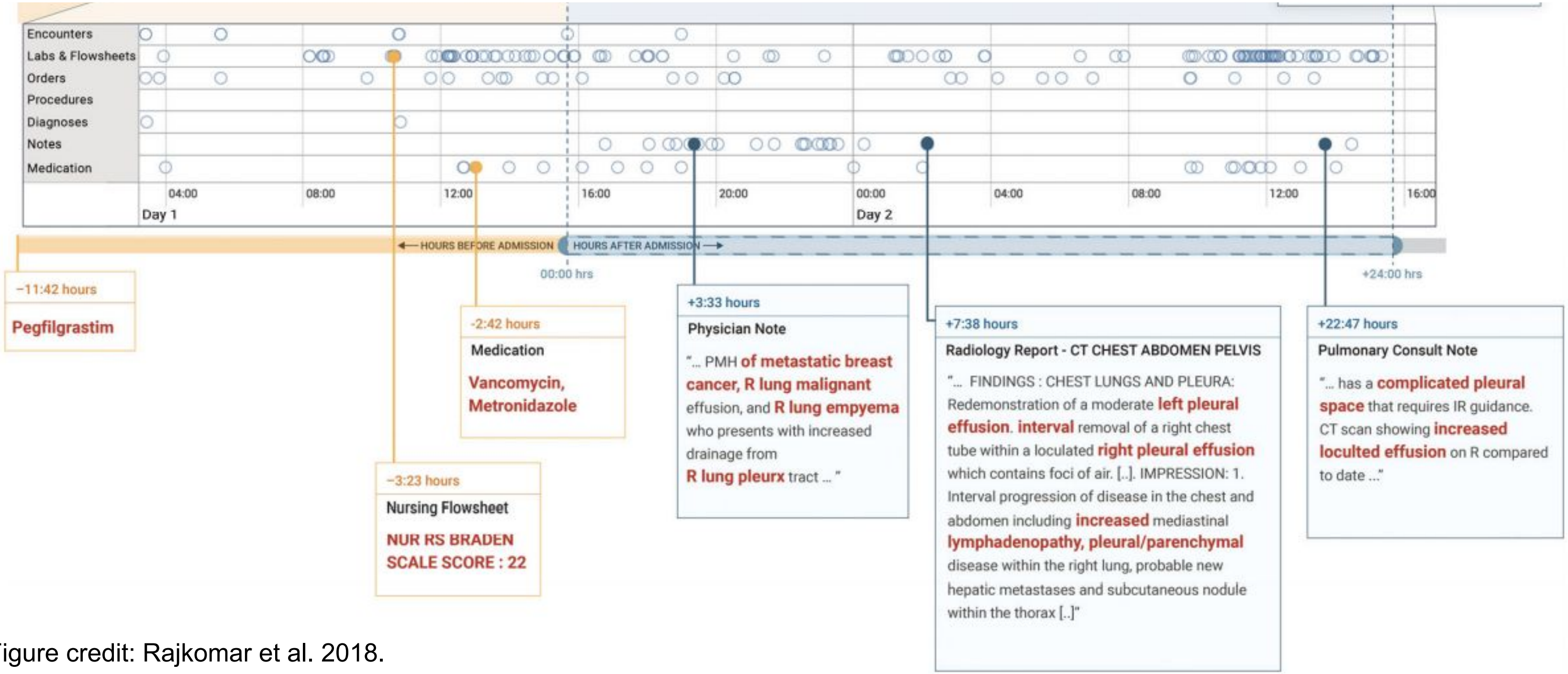
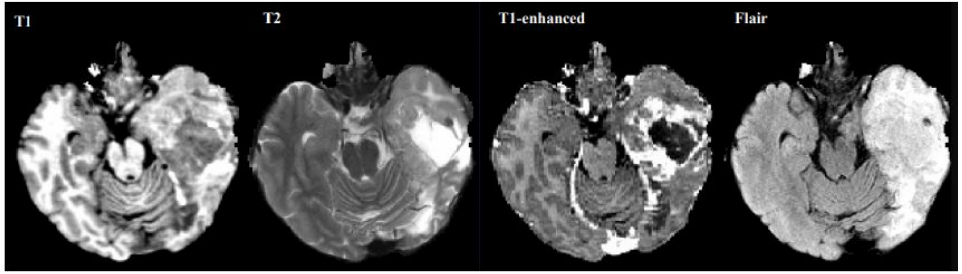


Figure credit: Rajkomar et al. 2018.

Similar data: can fuse at input

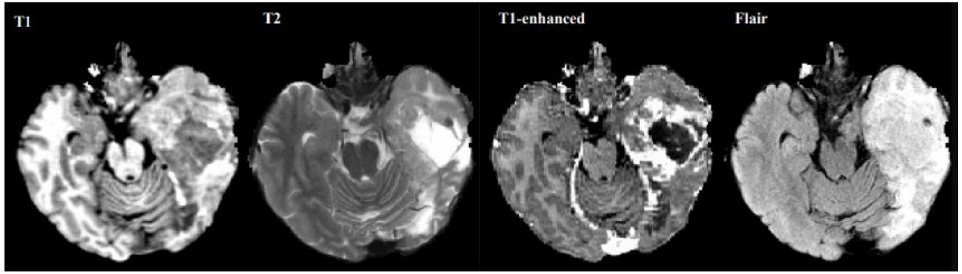
- Havaei et al.: brain tumor segmentation from multimodal MR images



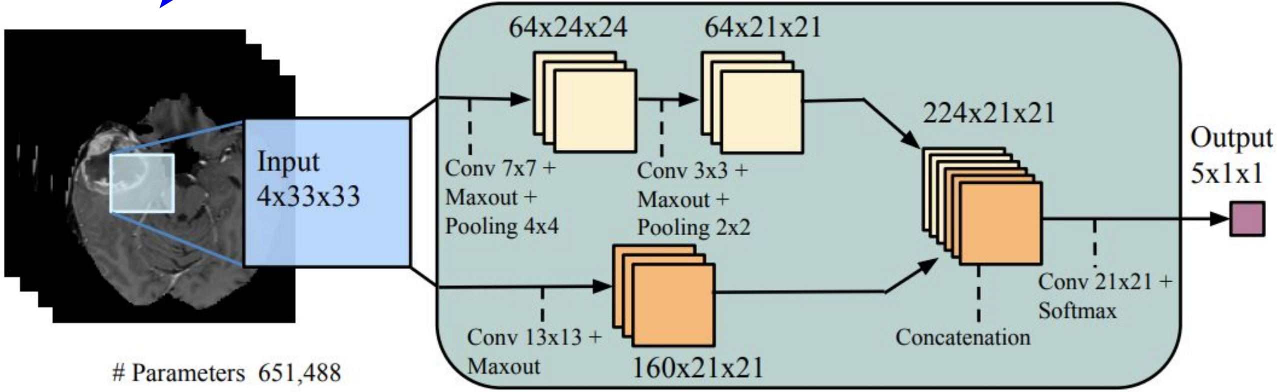
Havaei et al. Brain Tumor Segmentation with Deep Neural Networks. Medical Image Analysis, 2016.

Similar data: can fuse at input

- Havaei et al.: brain tumor segmentation from multimodal MR images



Stack modalities such that each channel of input is a different modality.

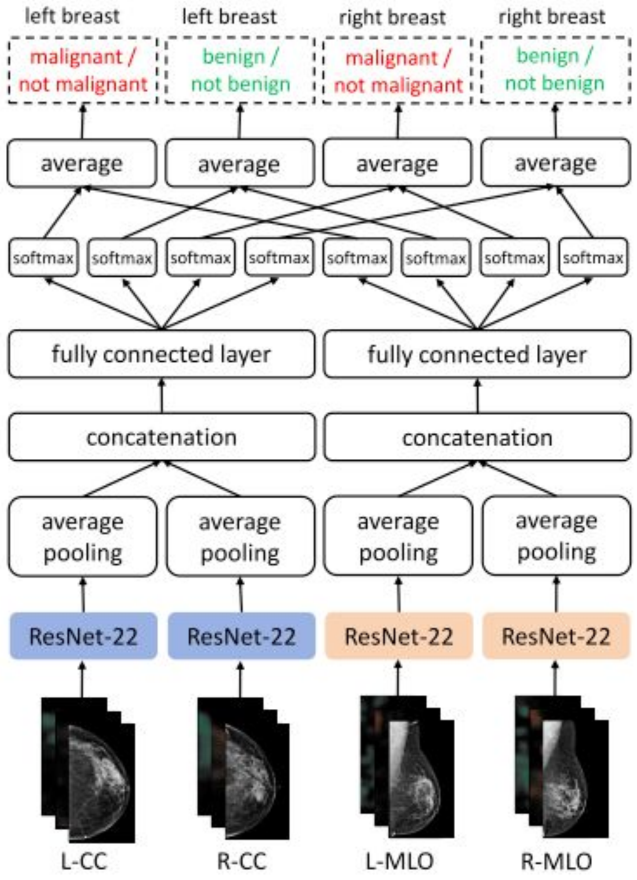


Havaei et al. Brain Tumor Segmentation with Deep Neural Networks. Medical Image Analysis, 2016.

More different data: may want some layers of modality-specific processing

Wu et al. 2019:

- Binary classification of breast malignant and benign findings
- Model based on ResNet architecture
- Multi-view network (different views can be considered different modalities)



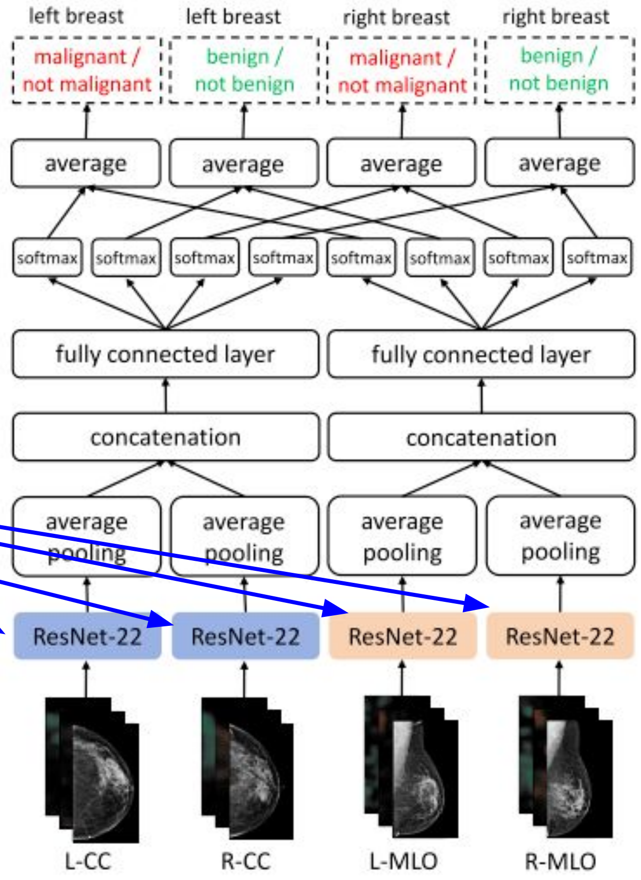
Wu et al. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. IEEE Trans Med Imaging, 2019.

More different data: may want some layers of modality-specific processing

Wu et al. 2019:

- Binary classification of breast malignant and benign findings
- Model based on ResNet architecture
- Multi-view network (different views can be considered different modalities)

Separate initial processing for different mammogram views



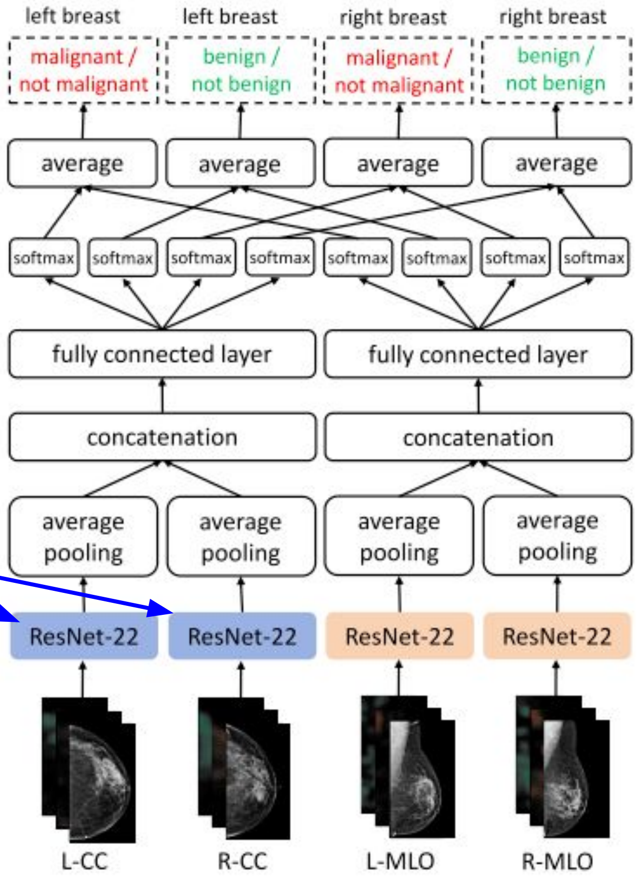
Wu et al. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. IEEE Trans Med Imaging, 2019.

More different data: may want some layers of modality-specific processing

Wu et al. 2019:

- Binary classification of breast malignant and benign findings
- Model based on ResNet architecture
- Multi-view network (different views can be considered different modalities)

Shared weights across the two networks



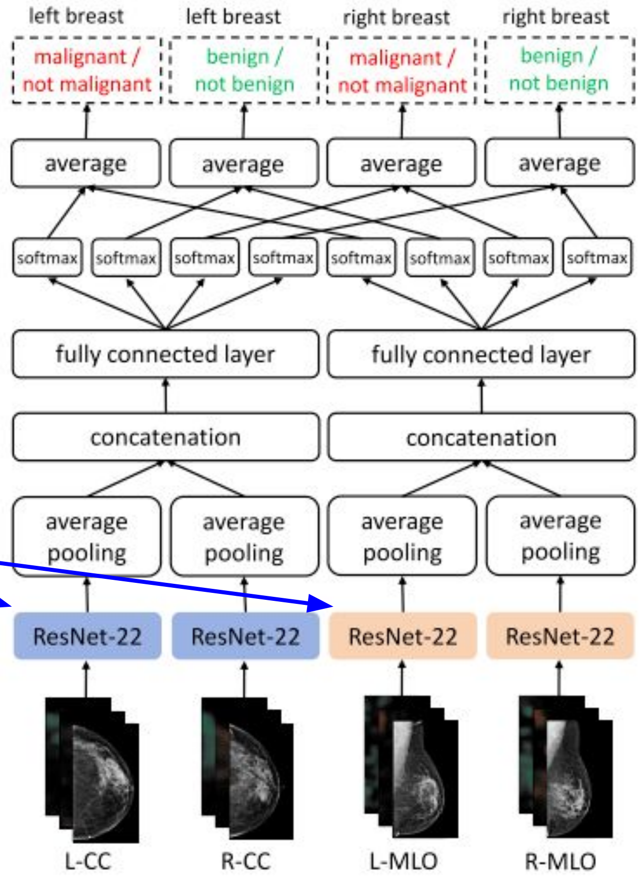
Wu et al. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. IEEE Trans Med Imaging, 2019.

More different data: may want some layers of modality-specific processing

Wu et al. 2019:

- Binary classification of breast malignant and benign findings
- Model based on ResNet architecture
- Multi-view network (different views can be considered different modalities)

More different views have separately learned parameters



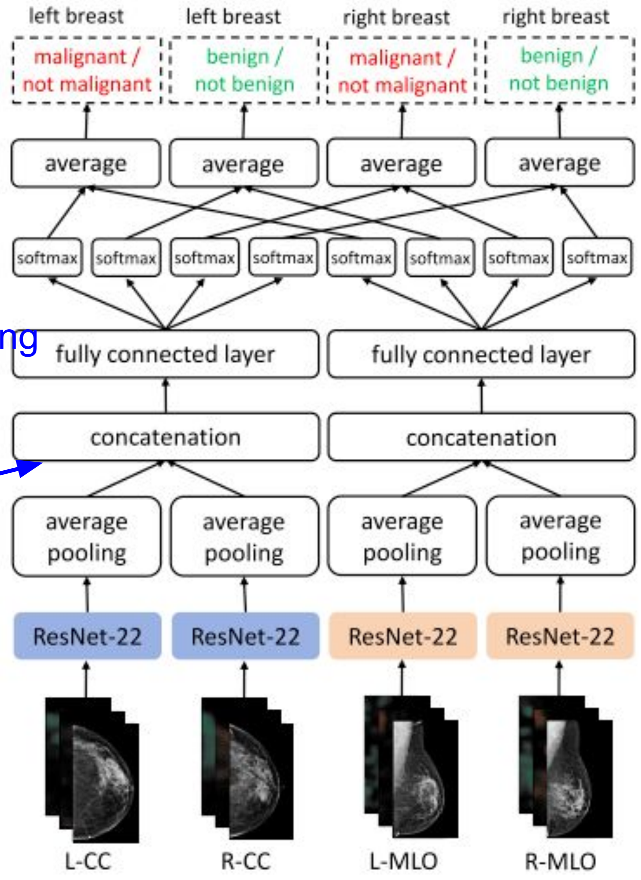
Wu et al. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. IEEE Trans Med Imaging, 2019.

More different data: may want some layers of modality-specific processing

Wu et al. 2019:

- Binary classification of breast malignant and benign findings
- Model based on ResNet architecture
- Multi-view network (different views can be considered different modalities)

Multimodal fusion at intermediate part of processing (very common): concatenate outputs of modality-specific processing into one feature vector.



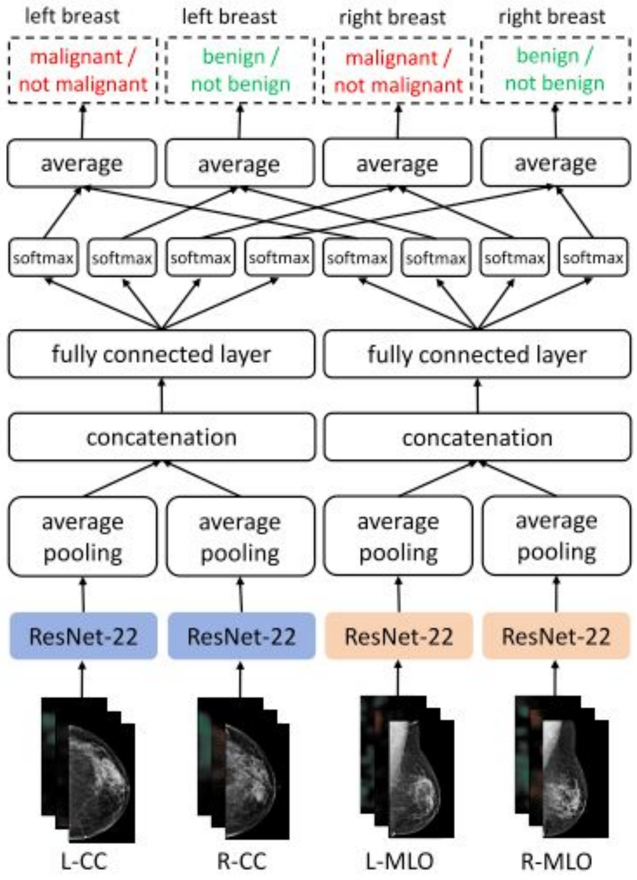
Wu et al. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. IEEE Trans Med Imaging, 2019.

More different data: may want some layers of modality-specific processing

Wu et al. 2019:

- Binary classification of breast malignant and benign findings
- Model based on ResNet architecture
- Multi-view network (different views can be considered different modalities)

Fully connected layer (or several) afterwards. Concatenated feature vector no longer contains spatial relationships suitable for conv layers.



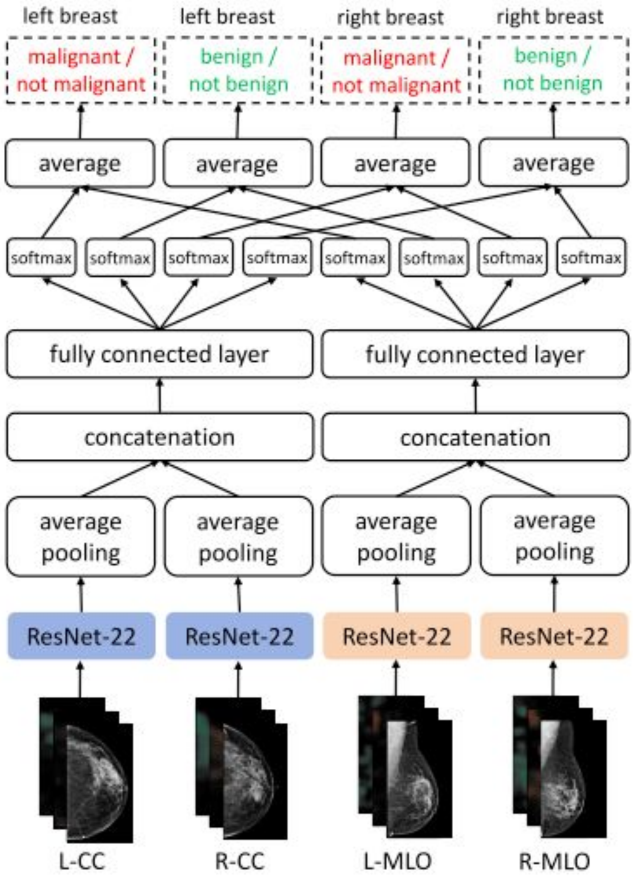
Wu et al. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. IEEE Trans Med Imaging, 2019.

More different data: may want some layers of modality-specific processing

Wu et al. 2019:

- Binary classification of breast malignant and benign findings
- Model based on ResNet architecture
- Multi-view network (different views can be considered different modalities)

Predict all 4 binary outputs from each view



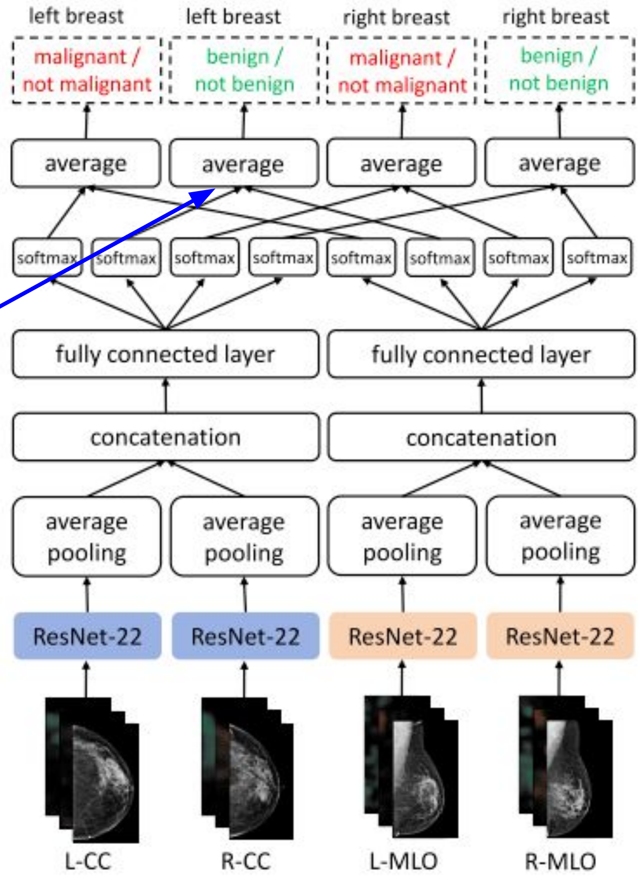
Wu et al. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. IEEE Trans Med Imaging, 2019.

More different data: may want some layers of modality-specific processing

Wu et al. 2019:

- Binary classification of breast malignant and benign findings
- Model based on ResNet architecture
- Multi-view network (different views can be considered different modalities)

This model also uses a second type of fusion for the CC vs. MLO views: late fusion of predictions through averaging.

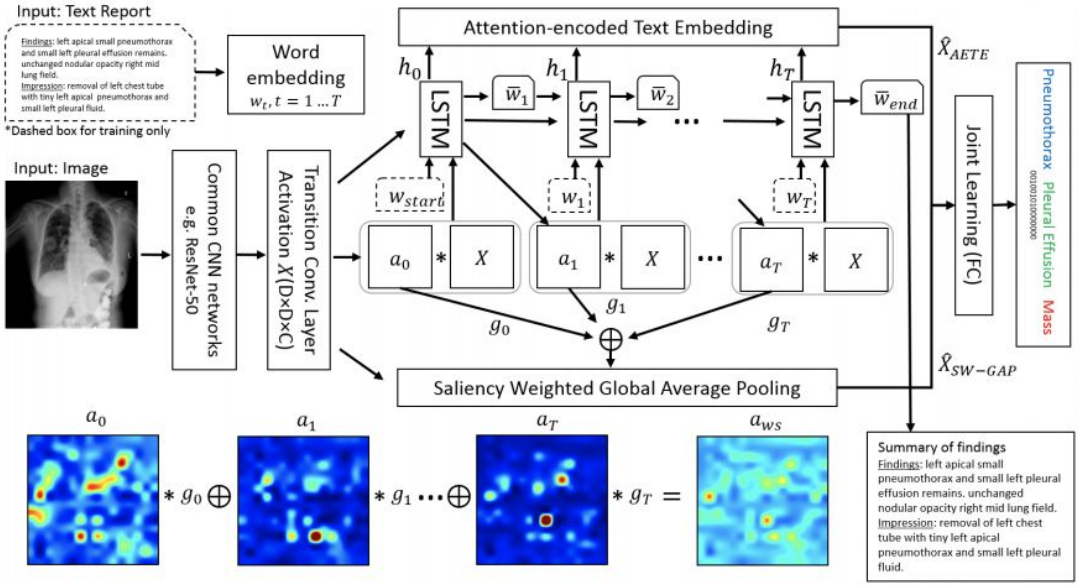


Wu et al. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. IEEE Trans Med Imaging, 2019.

A recurrent network approach for combining multimodal data

Wang et al. 2018:

- Jointly process chest x-rays and associated reports to produce disease labels that can be used to produce auto-annotation disease labels



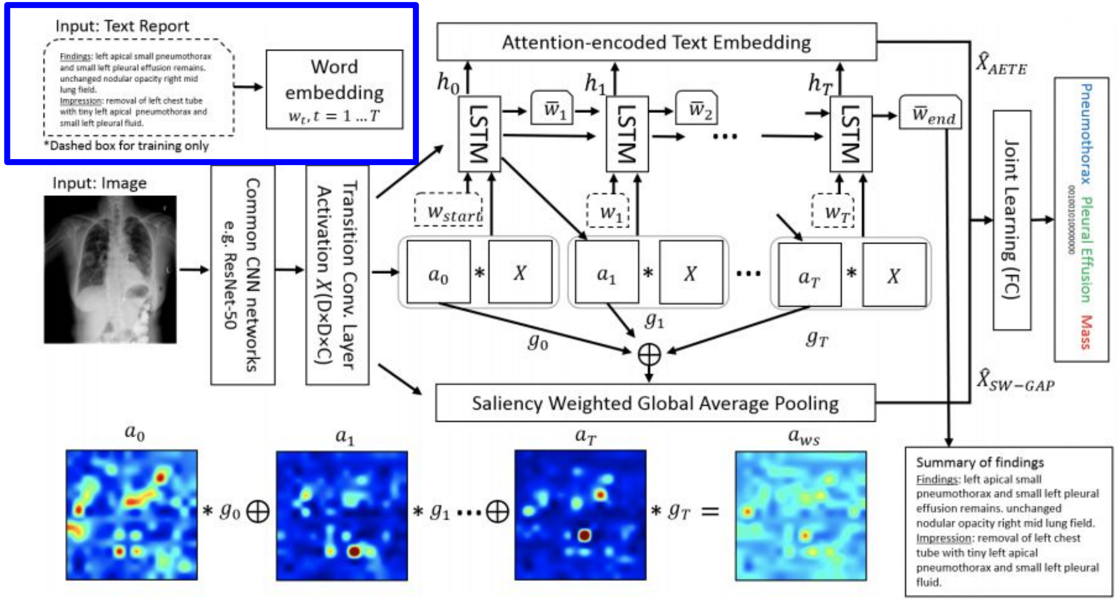
Wang et al. TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays. CVPR, 2018.

A recurrent network approach for combining multimodal data

Wang et al. 2018:

- Jointly process chest x-rays and associated reports to produce disease labels that can be used to produce auto-annotation disease labels

Use NLP approaches to generate word embedding representations of words in text



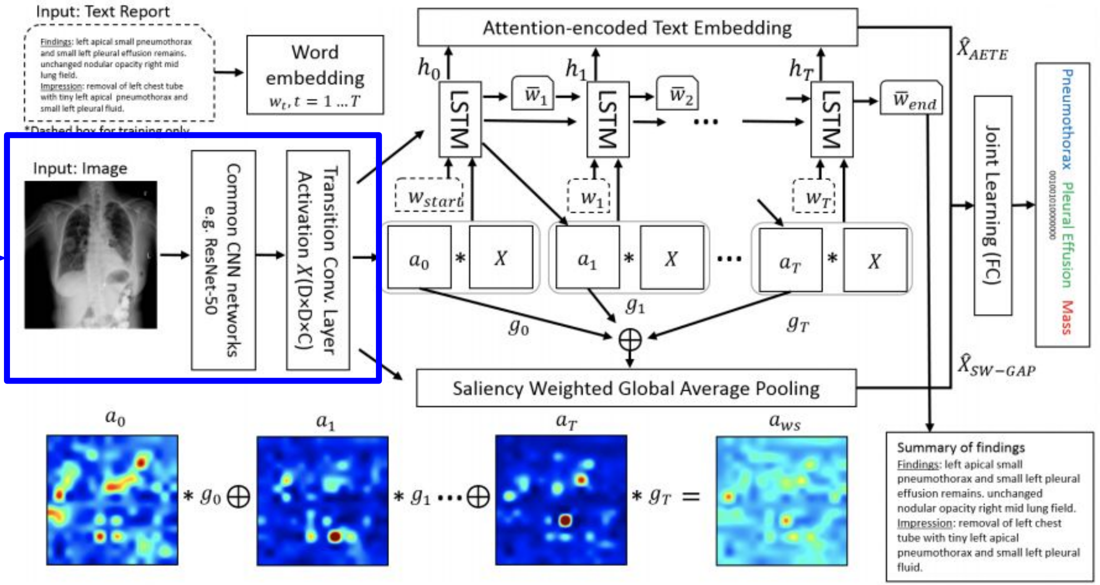
Wang et al. TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays. CVPR, 2018.

A recurrent network approach for combining multimodal data

Wang et al. 2018:

- Jointly process chest x-rays and associated reports to produce disease labels that can be used to produce auto-annotation disease labels

Use common CNN networks to generate feature representation of image data



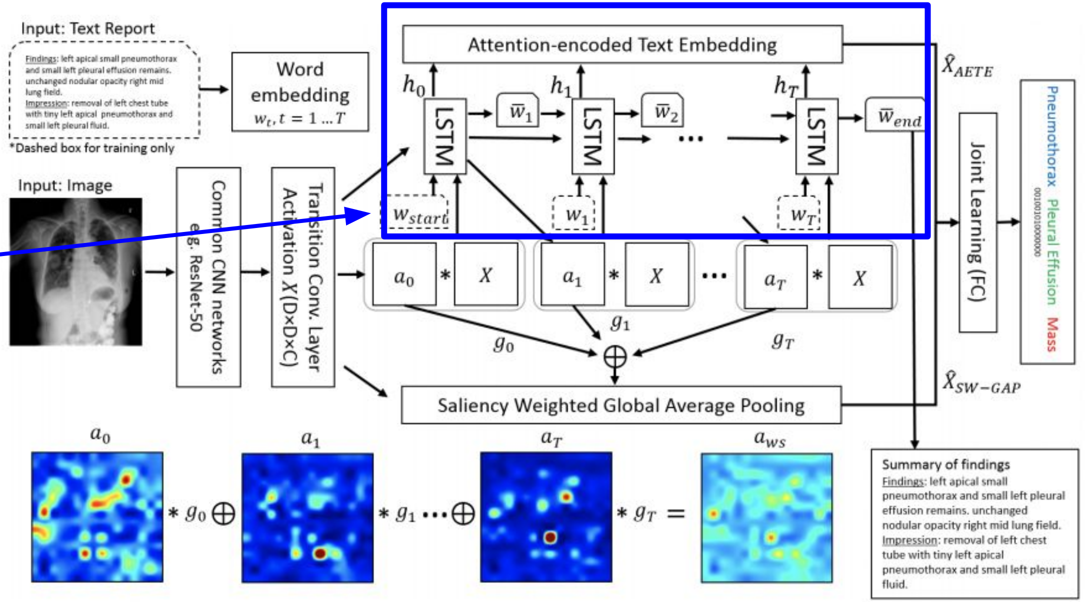
Wang et al. TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays. CVPR, 2018.

A recurrent network approach for combining multimodal data

Wang et al. 2018:

- Jointly process chest x-rays and associated reports to produce disease labels that can be used to produce auto-annotation disease labels

Use LSTM to process sequence of text data embedding representations



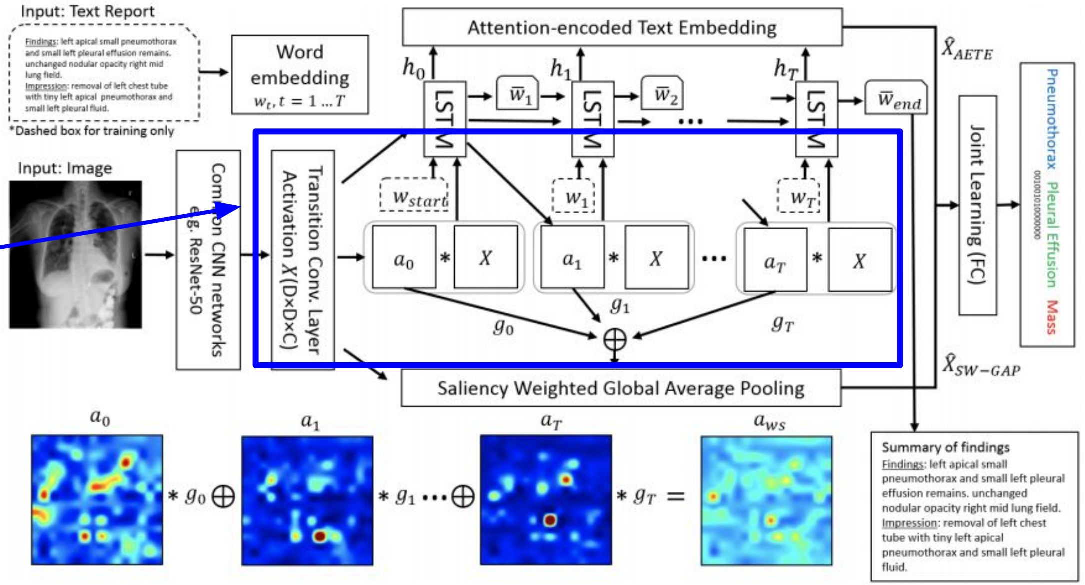
Wang et al. TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays. CVPR, 2018.

A recurrent network approach for combining multimodal data

Wang et al. 2018:

- Jointly process chest x-rays and associated reports to produce disease labels that can be used to produce auto-annotation disease labels

Image data is an additional input to the LSTM at each time step (with soft-attention weighting)



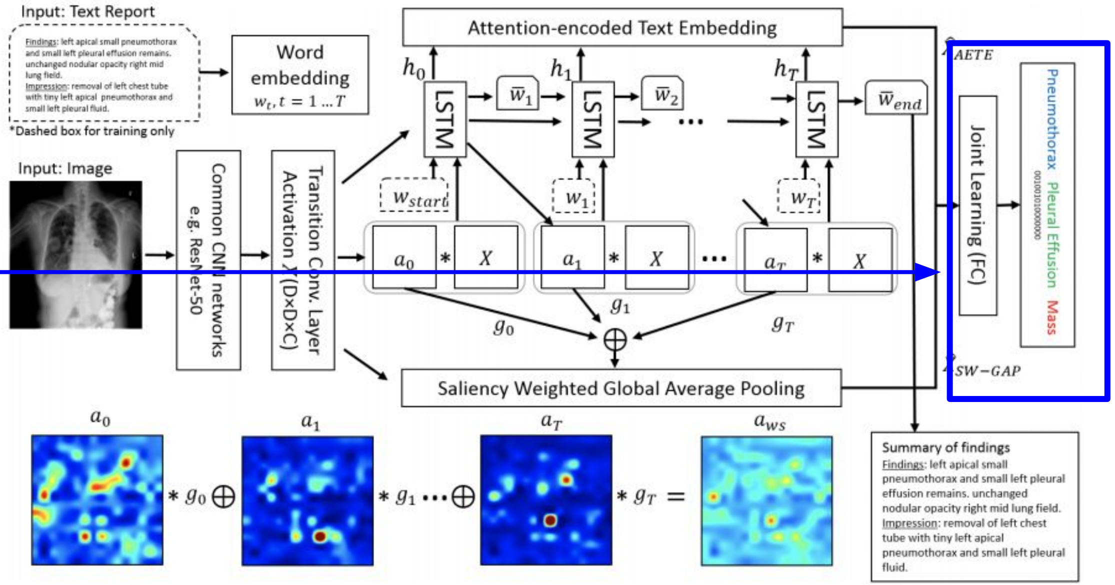
Wang et al. TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays. CVPR, 2018.

A recurrent network approach for combining multimodal data

Wang et al. 2018:

- Jointly process chest x-rays and associated reports to produce disease labels that can be used to produce auto-annotation disease labels

Final fully-connected layer fusion and prediction of disease labels

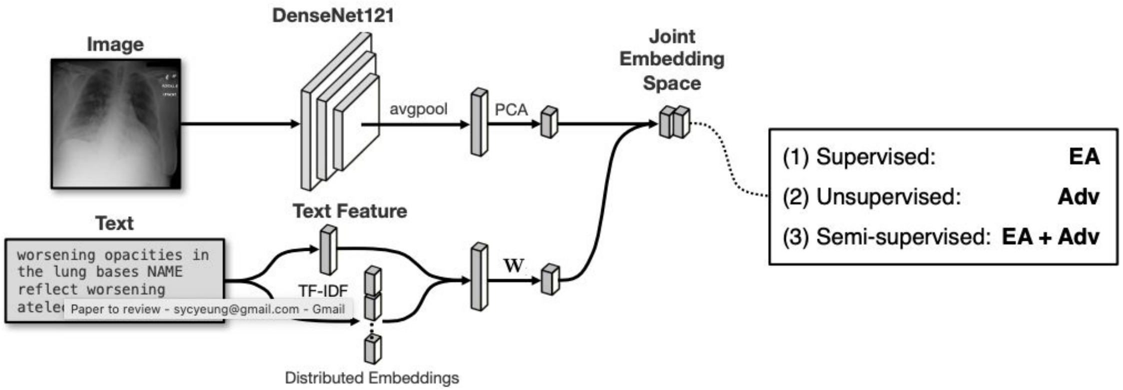


Wang et al. TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays. CVPR, 2018.

Another direction of research: learning multimodal embedding spaces

Hsu et al. 2018:

- Learn mapping from images and text to vectors in the same embedding space, such that images are embedded closer to their corresponding reports than other reports, and vice versa.
- Can be used for e.g. cross-domain retrieval



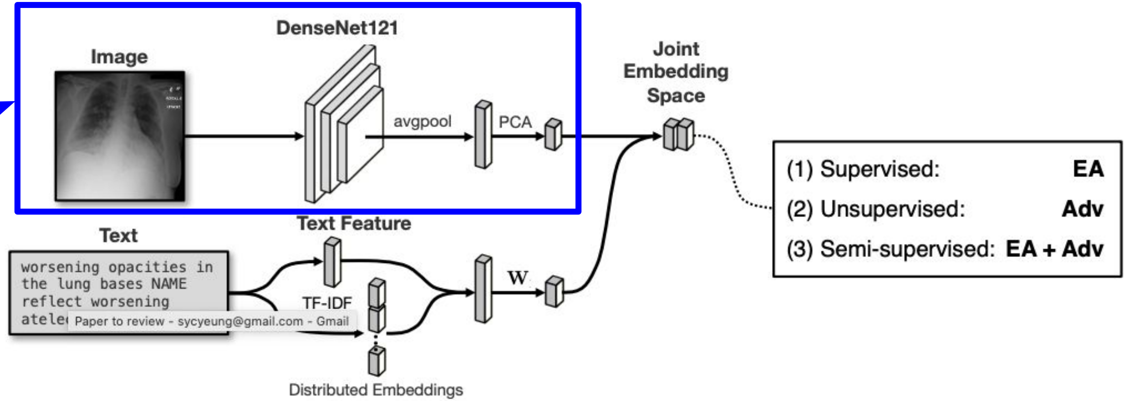
Hsu et al. Unsupervised Multimodal Representation Learning across Medical Images and Reports. NeurIPS ML4H, 2018.

Another direction of research: learning multimodal embedding spaces

Hsu et al. 2018:

- Learn mapping from images and text to vectors in the same embedding space, such that images are embedded closer to their corresponding reports than other reports, and vice versa.
- Can be used for e.g. cross-domain retrieval

Image-specific processing



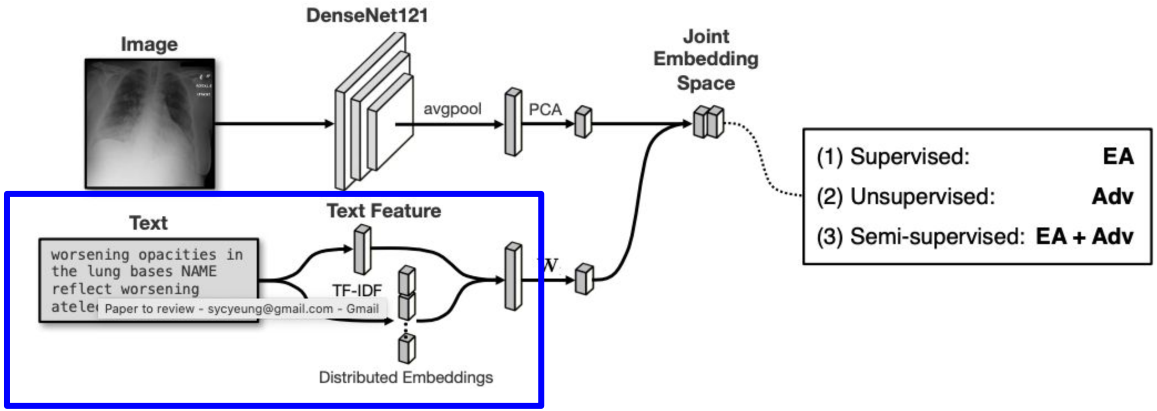
Hsu et al. Unsupervised Multimodal Representation Learning across Medical Images and Reports. NeurIPS ML4H, 2018.

Another direction of research: learning multimodal embedding spaces

Hsu et al. 2018:

- Learn mapping from images and text to vectors in the same embedding space, such that images are embedded closer to their corresponding reports than other reports, and vice versa.
- Can be used for e.g. cross-domain retrieval

Text-specific processing



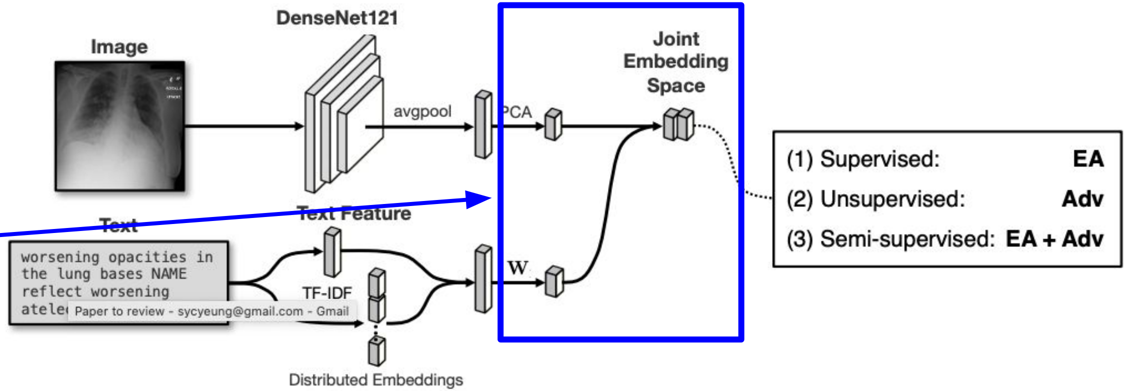
Hsu et al. Unsupervised Multimodal Representation Learning across Medical Images and Reports. NeurIPS ML4H, 2018.

Another direction of research: learning multimodal embedding spaces

Hsu et al. 2018:

- Learn mapping from images and text to vectors in the same embedding space, such that images are embedded closer to their corresponding reports than other reports, and vice versa.
- Can be used for e.g. cross-domain retrieval

Mapping to joint embedding space

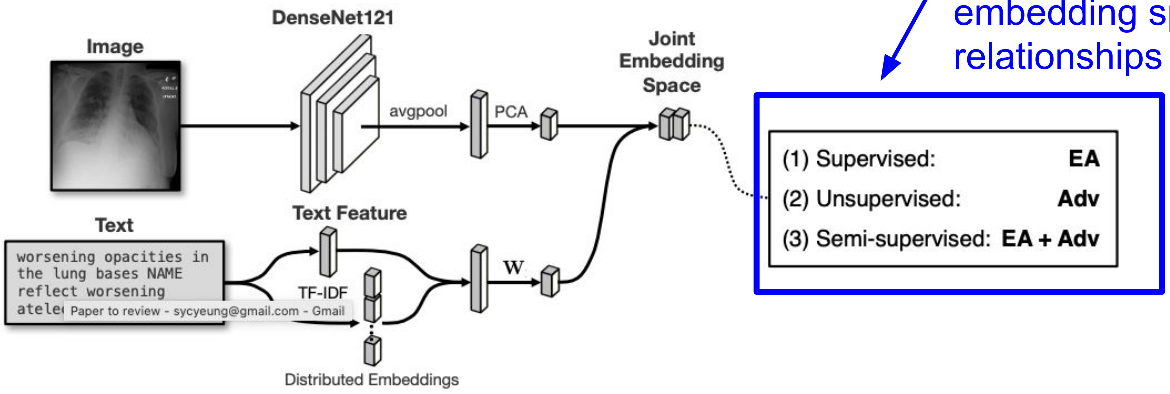


Hsu et al. Unsupervised Multimodal Representation Learning across Medical Images and Reports. NeurIPS ML4H, 2018.

Another direction of research: learning multimodal embedding spaces

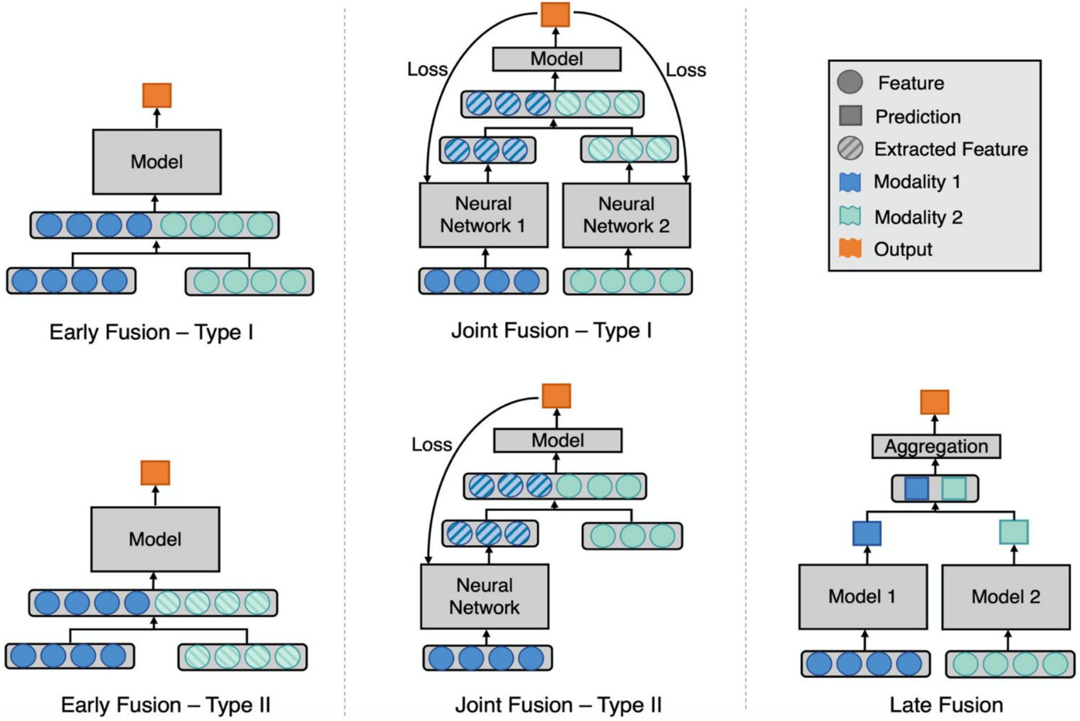
Hsu et al. 2018:
- Learn mapping from images and text to vectors in the same embedding space, such that images are embedded closer to their corresponding reports than other reports, and vice versa.
- Can be used for e.g. cross-domain retrieval

Different loss objectives can be used to encourage desired embedding space relationships



Hsu et al. Unsupervised Multimodal Representation Learning across Medical Images and Reports. NeurIPS ML4H, 2018.

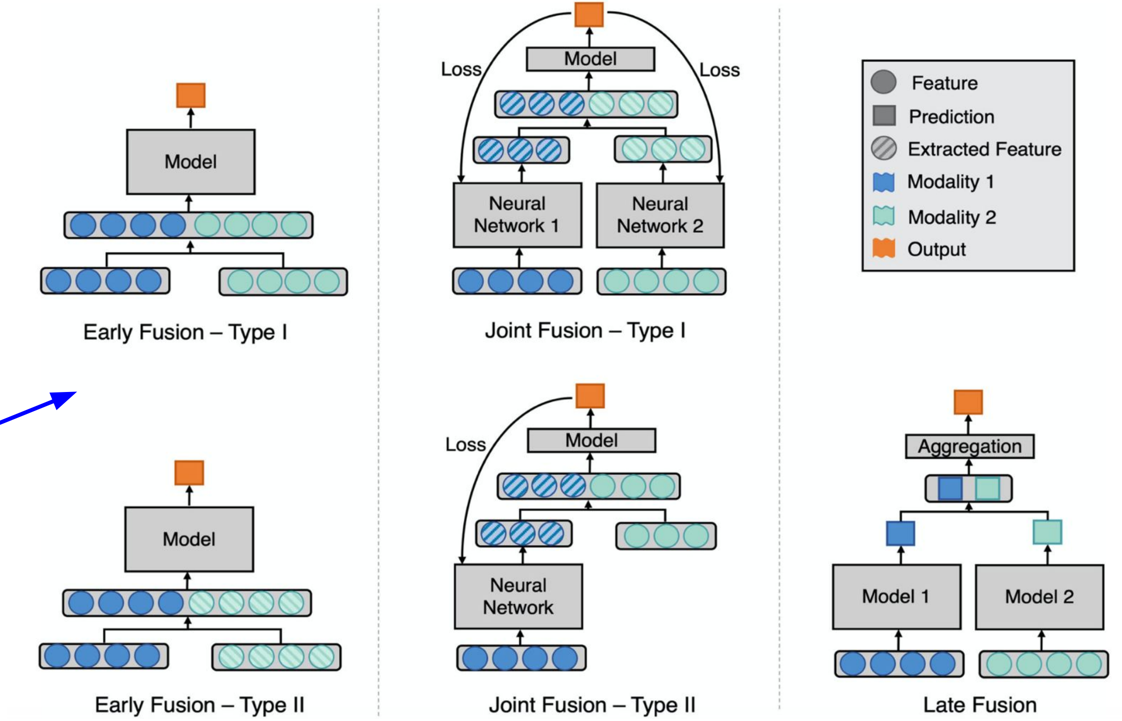
Categorizations of multimodal models



Huang et al. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines, 2020.

Categorizations of multimodal models

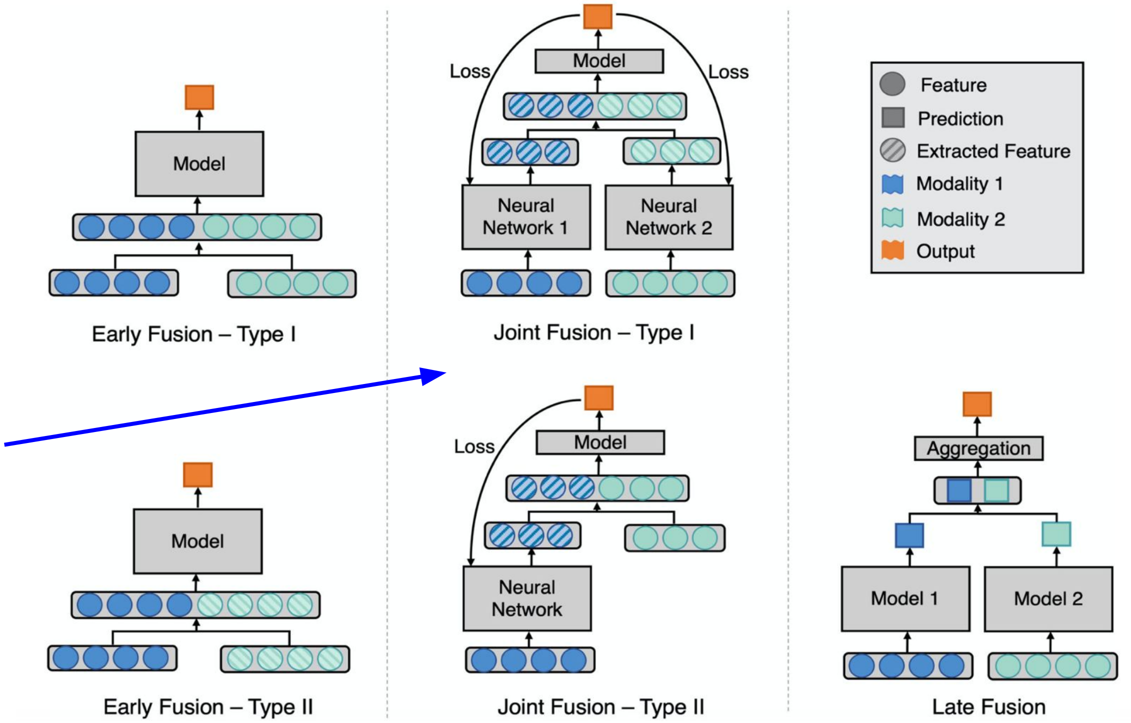
Early fusion: concatenate / combine data before any model processing. Includes using extracted features as input, if model gradients are not backpropagated to update feature extractor parameters



Huang et al. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines, 2020.

Categorizations of multimodal models

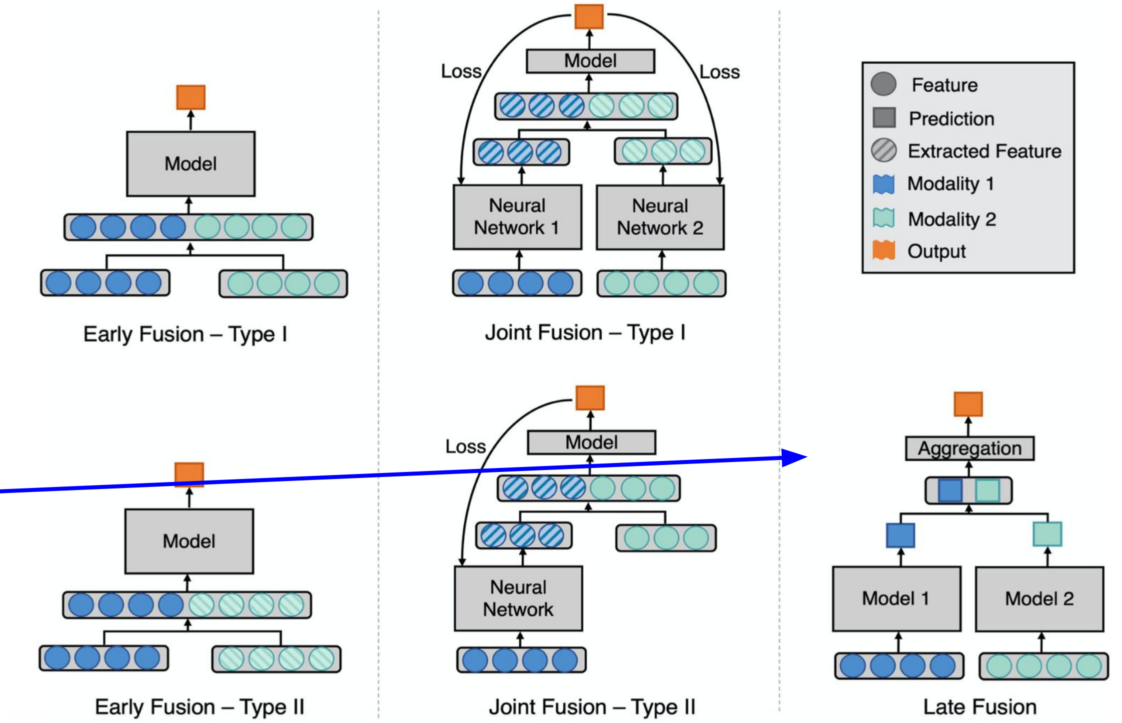
Joint fusion: Both modality-specific components (with learnable parameters) and combined-modality components within the model, that are updated during model training



Huang et al. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines, 2020.

Categorizations of multimodal models

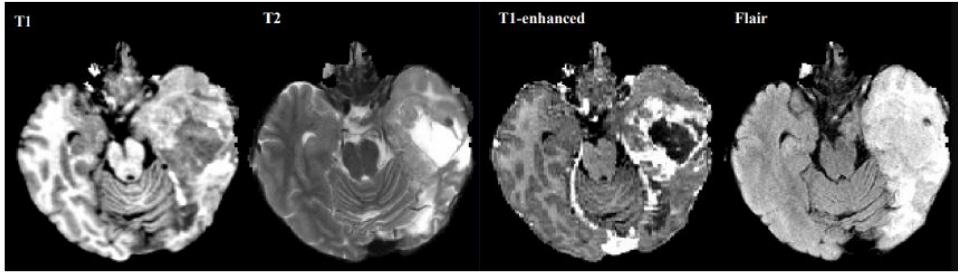
Late fusion: Main learnable model components are only model specific. Individual modality outputs are then aggregated.

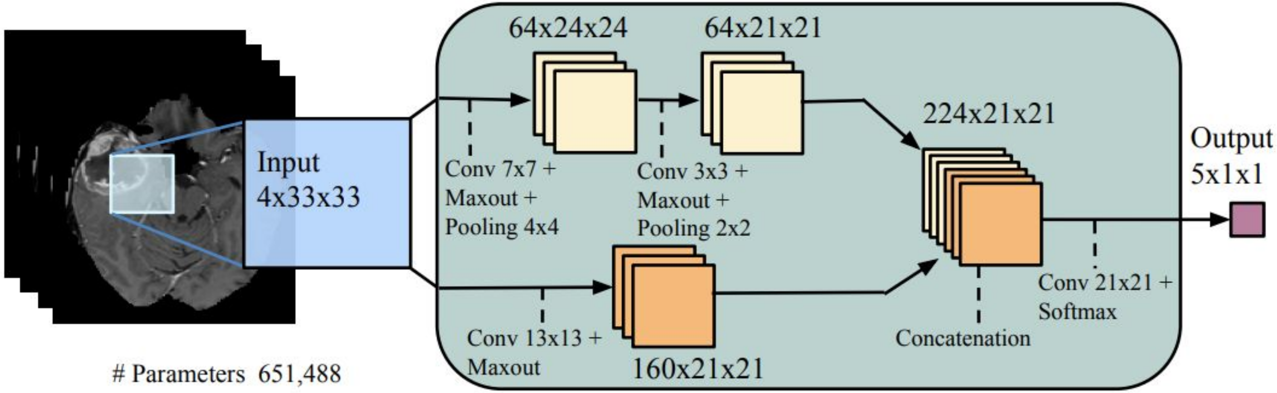


Huang et al. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines, 2020.

Q: What kind of fusion was this model?

- Havaei et al.: brain tumor segmentation from multimodal MR images



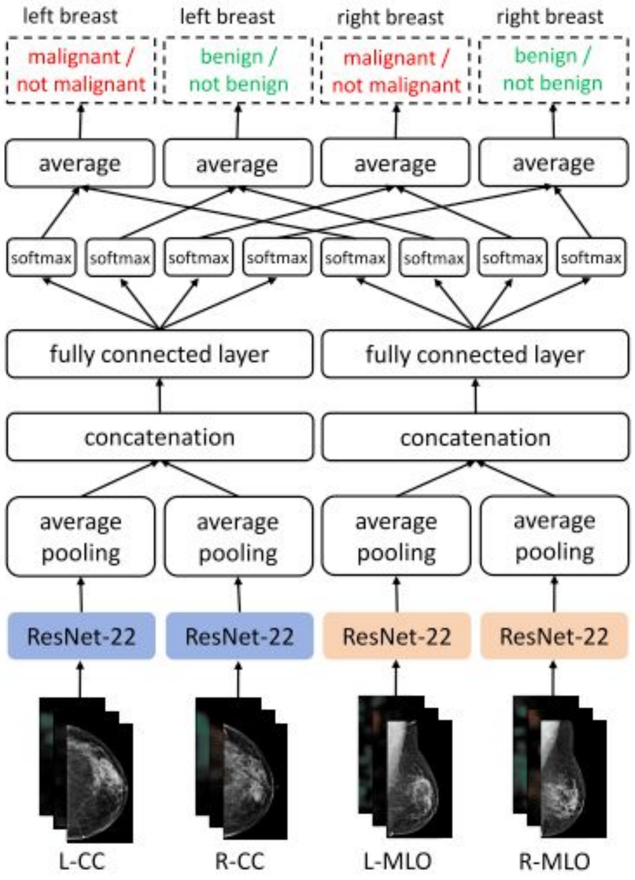


Havaei et al. Brain Tumor Segmentation with Deep Neural Networks. Medical Image Analysis, 2016.

Q: What kind of fusion was this model?

Wu et al. 2019:

- Binary classification of breast malignant and benign findings
- Model based on ResNet architecture
- Multi-view network (different views can be considered different modalities)

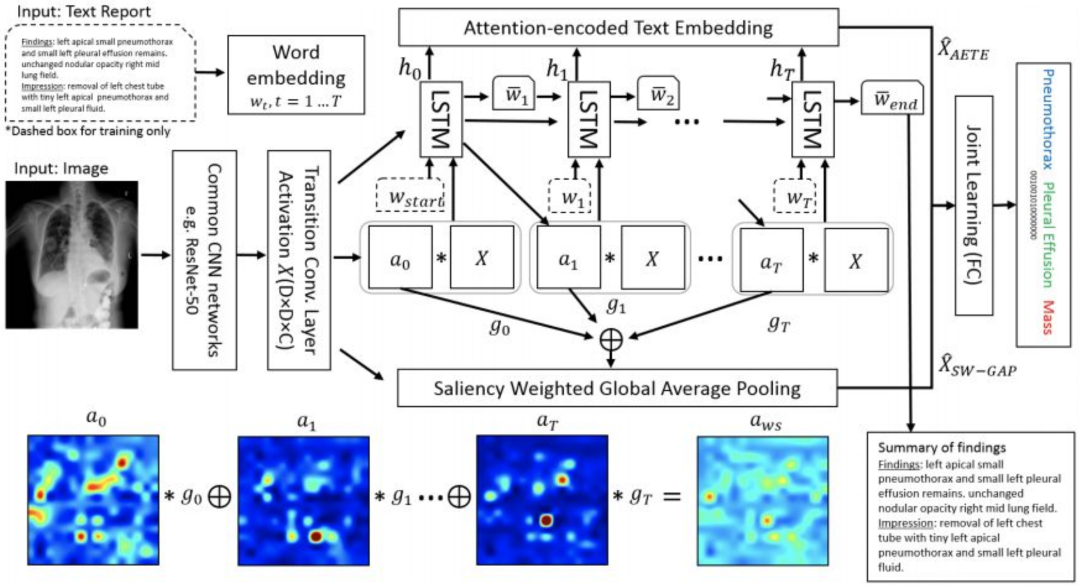


Wu et al. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. IEEE Trans Med Imaging, 2019.

Q: What kind of fusion was this model?

Wang et al. 2018:

- Jointly process chest x-rays and associated reports to produce disease labels that can be used to produce auto-annotation disease labels



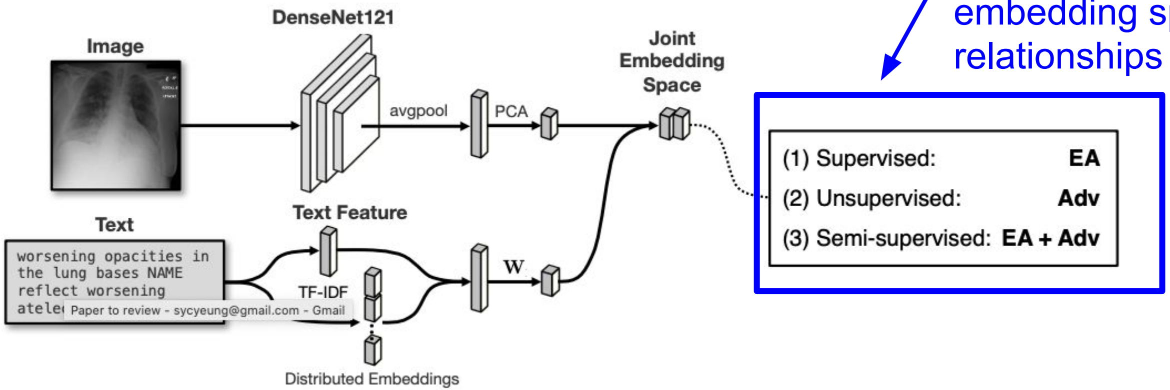
Wang et al. TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays. CVPR, 2018.

Back to learning multimodal embedding spaces

Hsu et al. 2018:

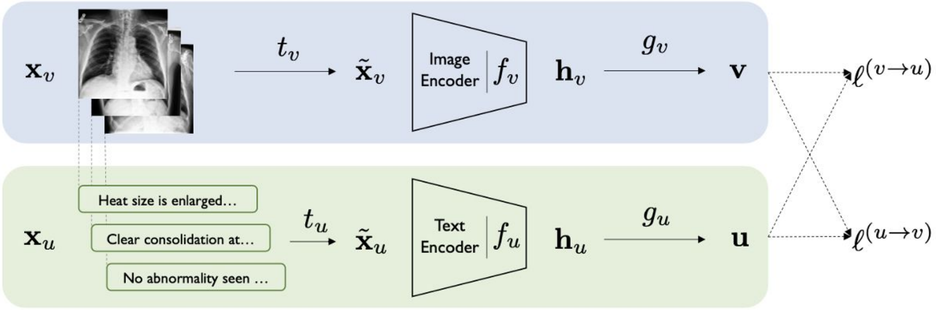
- Learn mapping from images and text to vectors in the same embedding space, such that images are embedded closer to their corresponding reports than other reports, and vice versa.
- Can be used for e.g. cross-domain retrieval

Different loss objectives can be used to encourage desired embedding space relationships

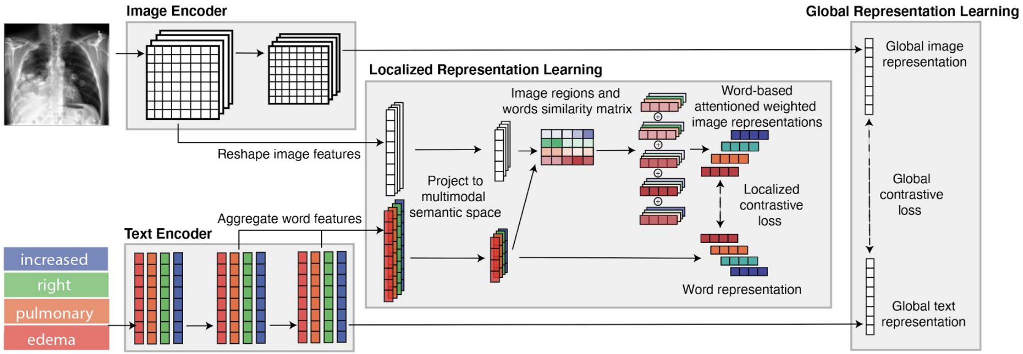


Hsu et al. Unsupervised Multimodal Representation Learning across Medical Images and Reports. NeurIPS ML4H, 2018.

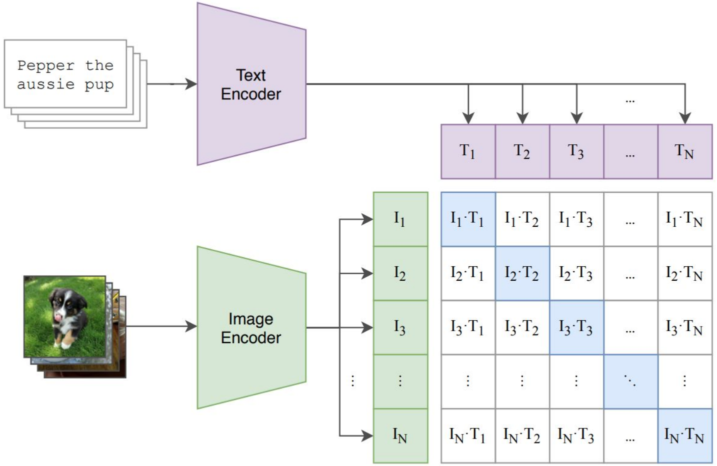
A little more: learning multimodal embedding spaces through **contrastive learning**



Zhang et al. 2020.



Huang et al. 2021.



Radford et al. 2021.

To understand contrastive learning, first understand self-supervised learning

Traditional supervised learning trains a model to perform a prediction task, using paired training data of inputs with corresponding ground truth labels for the desired task (e.g., manual class labels or EHR-obtained labels).

To understand contrastive learning, first understand self-supervised learning

Traditional supervised learning trains a model to perform a prediction task, using paired training data of inputs with corresponding ground truth labels for the desired task (e.g., manual class labels or EHR-obtained labels).

Self-supervised learning does not directly train a model to perform the desired prediction task. Instead, it generates supervisory training signal from raw data itself to learn a good feature encoder for the data type. No external labels (e.g, manual class labels) are used during self-supervised training. Then, this feature encoder can be useful for downstream tasks, such as initializing and fine-tuning a prediction model with much less labeled data needed.

To understand contrastive learning, first understand self-supervised learning

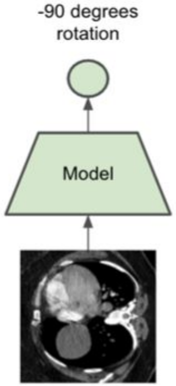
Traditional supervised learning trains a model to perform a prediction task, using paired training data of inputs with corresponding ground truth labels for the desired task (e.g., manual class labels or EHR-obtained labels).

Self-supervised learning does not directly train a model to perform the desired prediction task. Instead, it generates supervisory training signal from raw data itself to learn a good feature encoder for the data type. No external labels (e.g, manual class labels) are used during self-supervised training. Then, this feature encoder can be useful for downstream tasks, such as initializing and fine-tuning a prediction model with much less labeled data needed.

Effective way to tackle challenges of limited labeled data! Related to earlier discussion on pre-training on larger datasets and transfer learning, now we can also use self-supervised learning to pre-train on larger amounts of ***unlabeled*** data from the same domain.

To understand contrastive learning, first understand self-supervised learning

Some common types of self-supervised learning objectives:

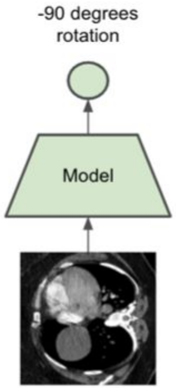


Innate relationship objective
E.g., predict rotation angle (or some other innate property) of an image

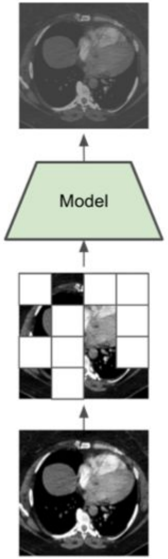
Figure credit: Mars Huang

To understand contrastive learning, first understand self-supervised learning

Some common types of self-supervised learning objectives:



Innate relationship objective
E.g., predict rotation angle (or some other innate property) of an image

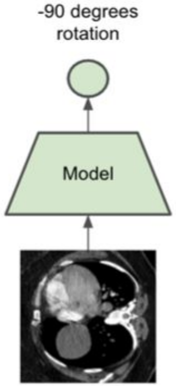


Self-prediction objective
Mask parts of input data and predict these parts

Figure credit: Mars Huang

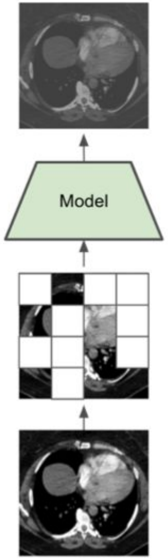
To understand contrastive learning, first understand self-supervised learning

Some common types of self-supervised learning objectives:

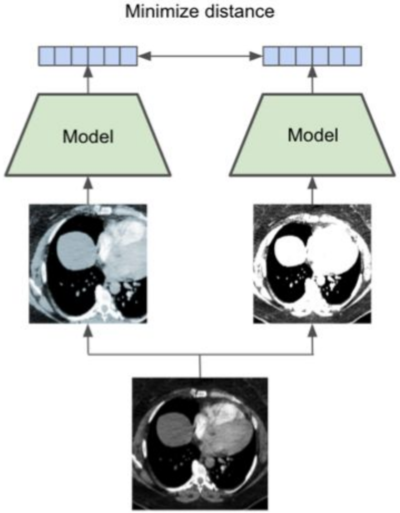


Innate relationship objective
E.g., predict rotation angle (or some other innate property) of an image

Figure credit: Mars Huang



Self-prediction objective
Mask parts of input data and predict these parts

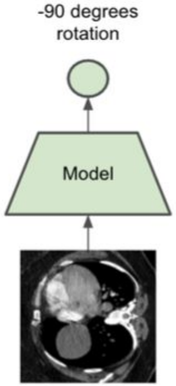


Contrastive objective
Different views of the same input should have more similar representation to each other than with a different input

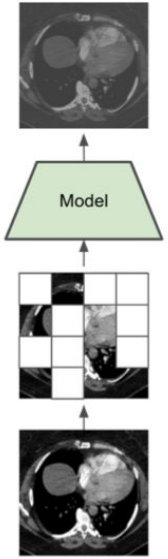
To understand contrastive learning, first understand self-supervised learning

Some common types of self-supervised learning objectives:

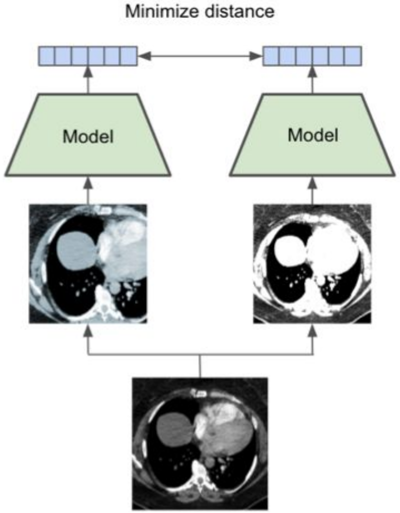
Can have varied formulations of these objectives within each type



Innate relationship objective
E.g., predict rotation angle (or some other innate property) of an image



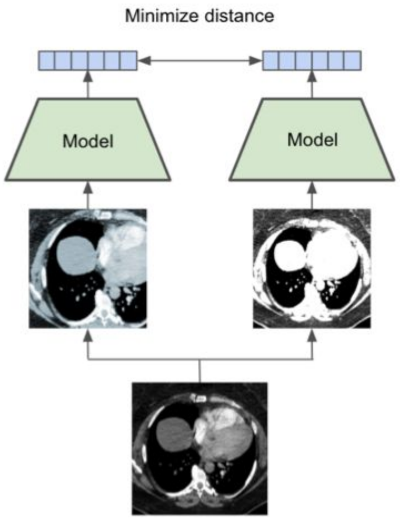
Self-prediction objective
Mask parts of input data and predict these parts



Contrastive objective
Different views of the same input should have more similar representation to each other than with a different input

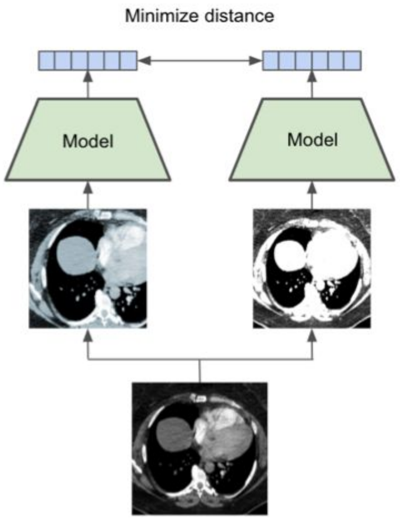
Figure credit: Mars Huang

SimCLR: a common approach for contrastive self-supervised learning



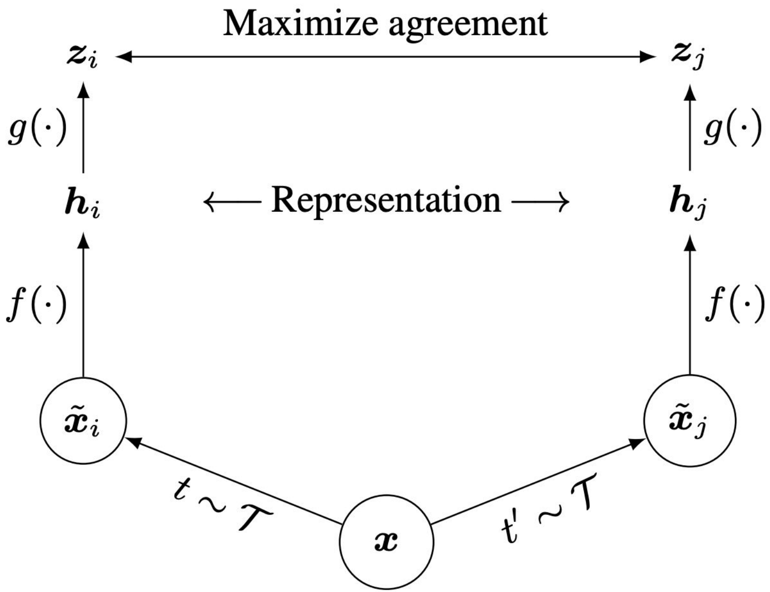
Contrastive objective
Different views of the same input should have more similar representation to each other than with a different input

SimCLR: a common approach for contrastive self-supervised learning



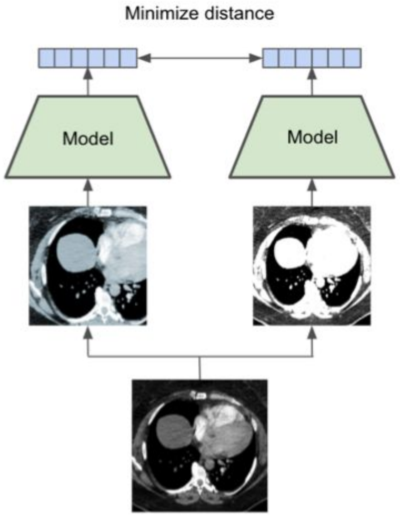
Contrastive objective
Different views of the same input should have more similar representation to each other than with a different input

SimCLR formulation



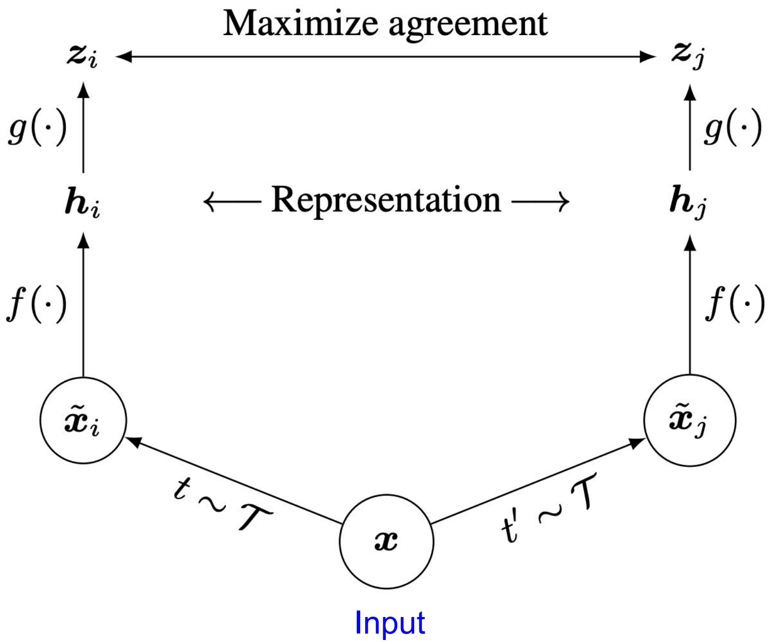
Chen et al. 2020

SimCLR: a common approach for contrastive self-supervised learning



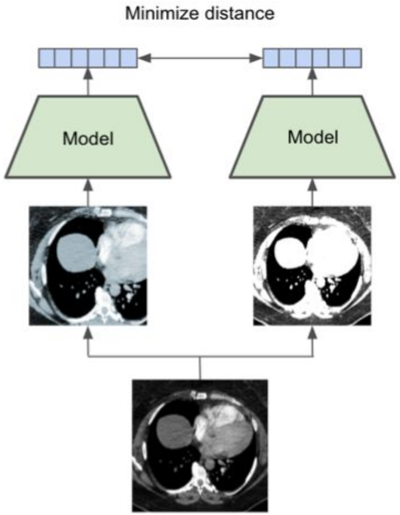
Contrastive objective
Different views of the same input should have more similar representation to each other than with a different input

SimCLR formulation



Chen et al. 2020

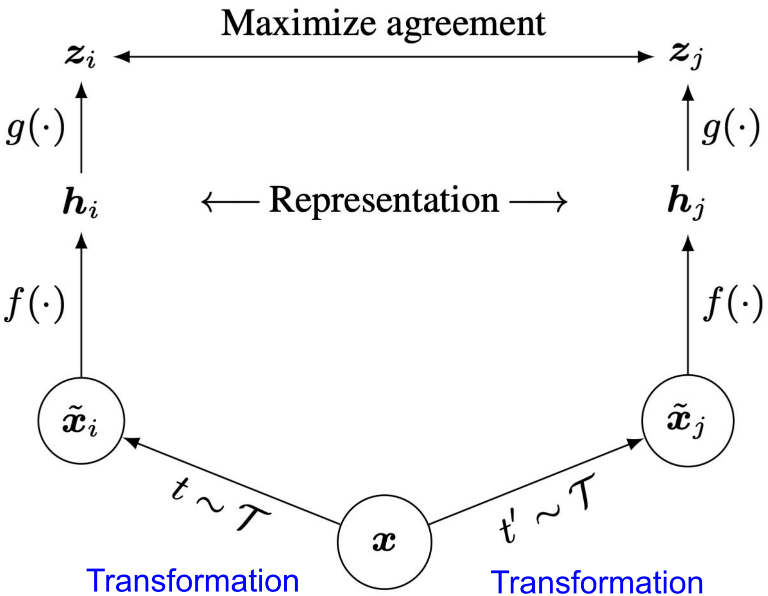
SimCLR: a common approach for contrastive self-supervised learning



Contrastive objective

Different views of the same input should have more similar representation to each other than with a different input

SimCLR formulation



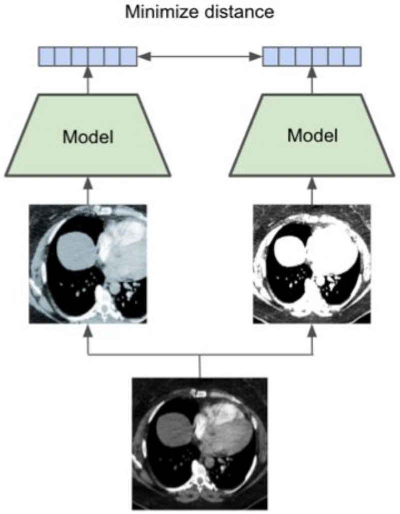
Transformation *t*

Input

Transformation *t'*

Chen et al. 2020

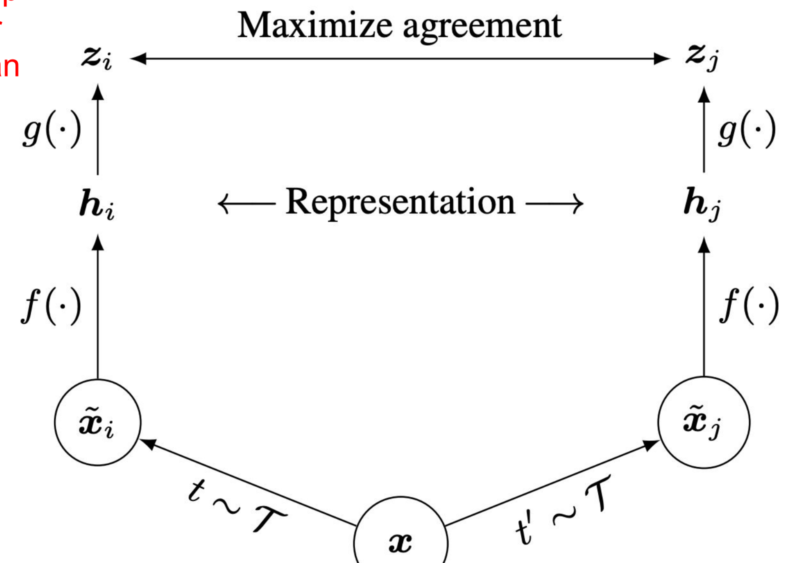
SimCLR: a common approach for contrastive self-supervised learning



Contrastive objective
Different views of the same input should have more similar representation to each other than with a different input

Transformation set: random crop (w/ flip and resize), color distortion, Gaussian blur

SimCLR formulation



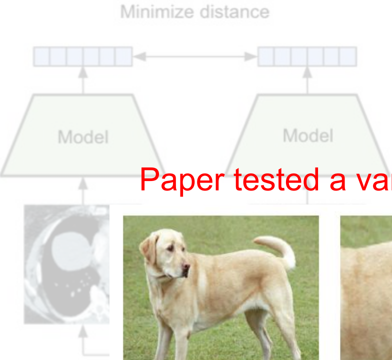
Transformation *t*

Input

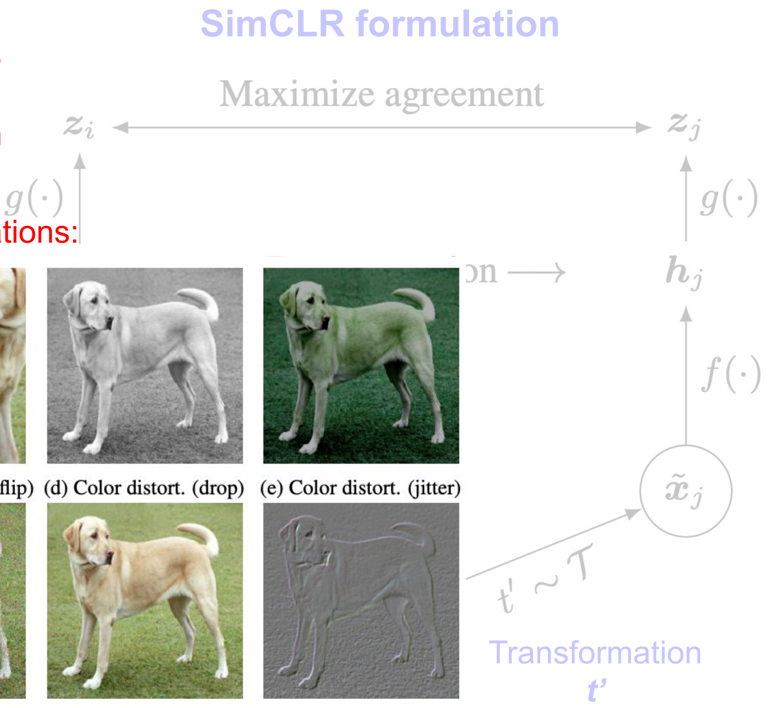
Transformation *t'*

Chen et al. 2020

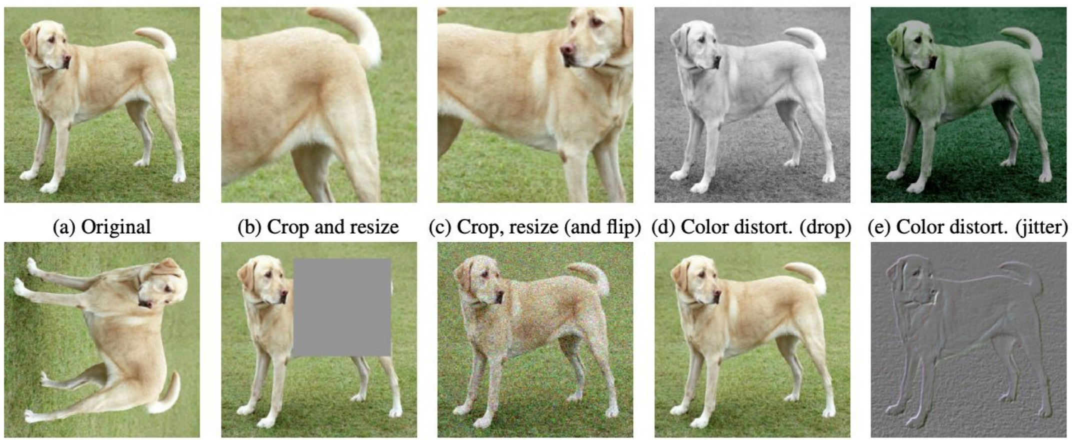
SimCLR: a common approach for contrastive self-supervised learning



Transformation set: random crop (w/ flip and resize), color distortion, Gaussian blur



Paper tested a variety of other transformations:



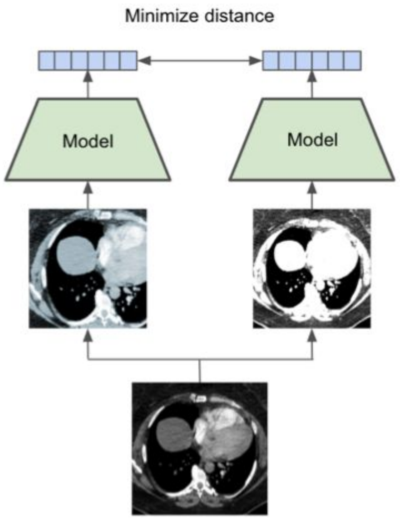
(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)

(f) Rotate $\{90°, 180°, 270°\}$ (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering

Contr
Different views
have more sim
other thar

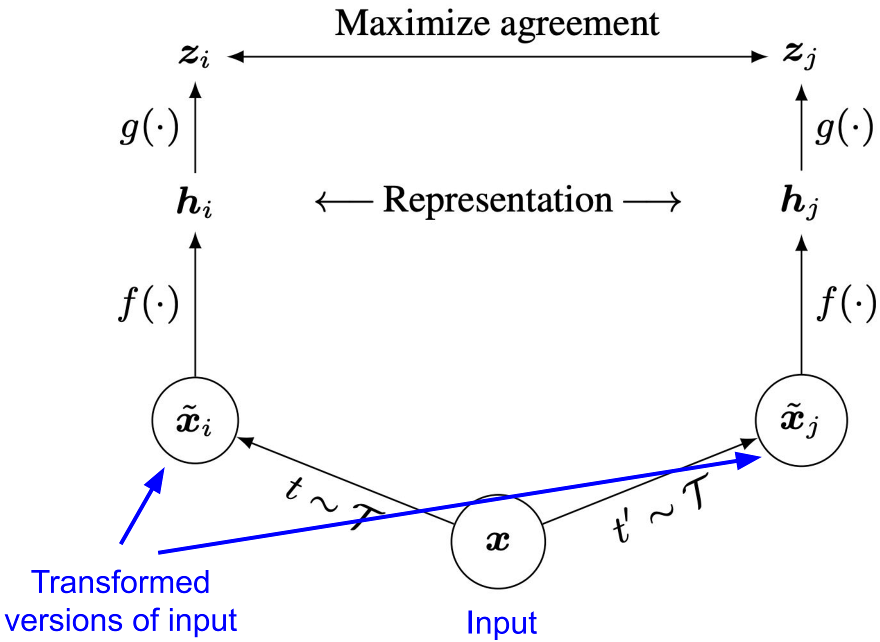
Chen et al. 2020

SimCLR: a common approach for contrastive self-supervised learning



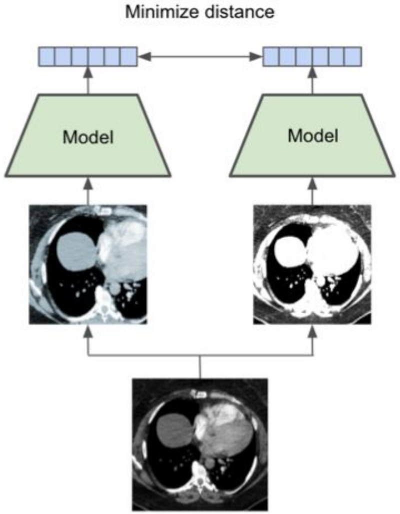
Contrastive objective
Different views of the same input should have more similar representation to each other than with a different input

SimCLR formulation



Chen et al. 2020

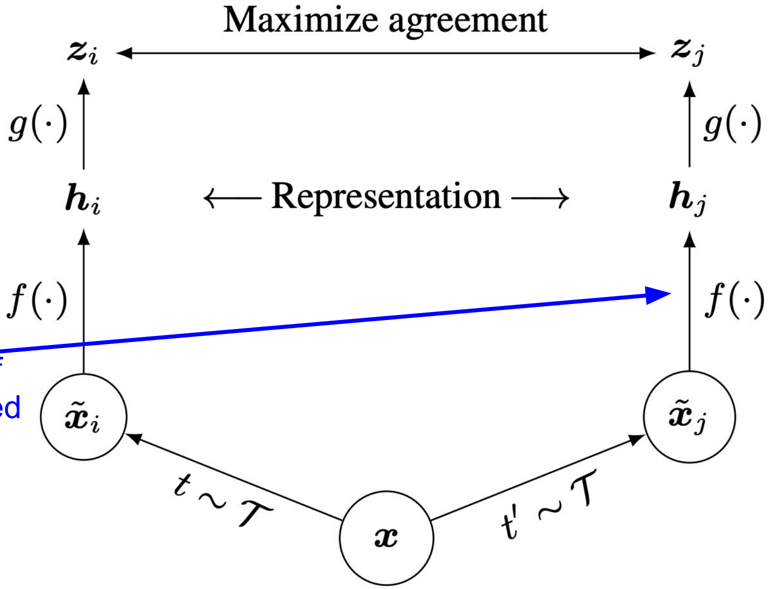
SimCLR: a common approach for contrastive self-supervised learning



Contrastive objective
Different views of the same input should have more similar representation to each other than with a different input

SimCLR formulation

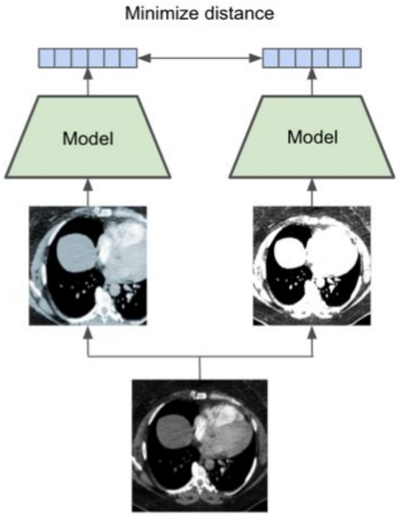
Encoder network f (same model applied to both image transformers)



Input

Chen et al. 2020

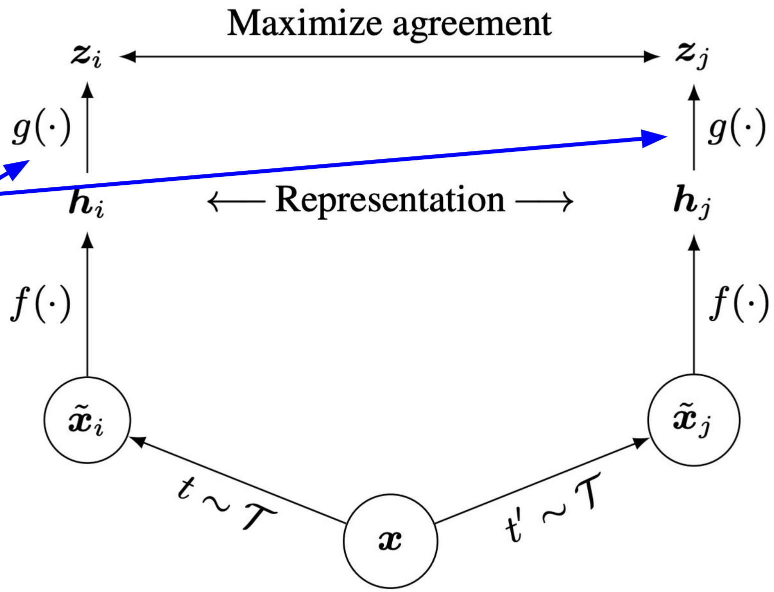
SimCLR: a common approach for contrastive self-supervised learning



Contrastive objective
Different views of the same input should have more similar representation to each other than with a different input

SimCLR formulation

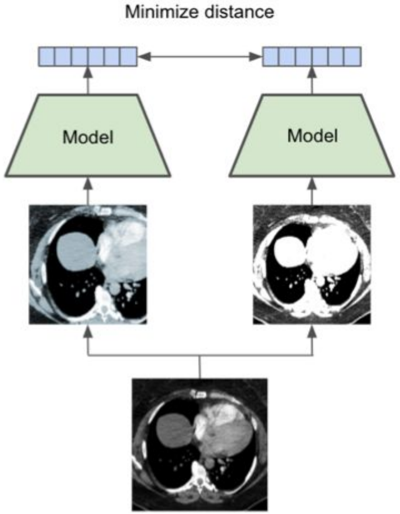
Projection head (MLP w/ one hidden layer), same network applied to both representations h_i, h_j



Input

Chen et al. 2020

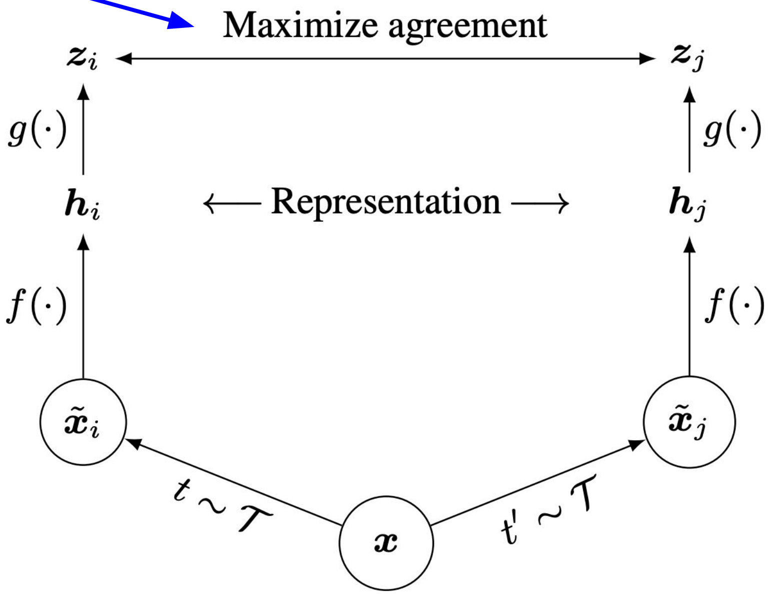
SimCLR: a common approach for contrastive self-supervised learning



Contrastive objective
Different views of the same input should have more similar representation to each other than with a different input

Contrastive loss

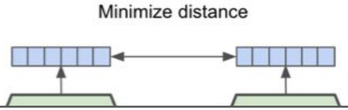
SimCLR formulation



Input

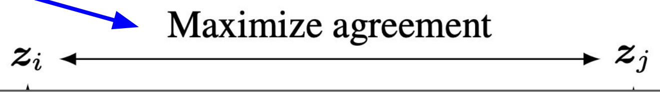
Chen et al. 2020

SimCLR: a common approach for contrastive self-supervised learning



Contrastive loss

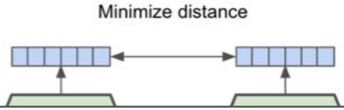
SimCLR formulation



Contrastive loss can take the form of a familiar cross-entropy loss!

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

SimCLR: a common approach for contrastive self-supervised learning



Contrastive loss

SimCLR formulation

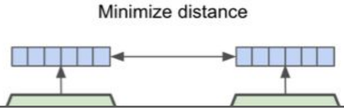
$$z_i \xleftrightarrow{\text{Maximize agreement}} z_j$$

Contrastive loss can take the form of a familiar cross-entropy loss!

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

Compute loss over a minibatch of N examples. Generate two augmented views of each example, resulting in 2N data points total. Now in the contrastive loss, we wish for a pair of data points (i,j) corresponding to augmentations of the same example to have closer representation similarity than with other data points generated from different examples. Use a cross-entropy formulation: given data point i, similarity with data point j should have higher score than with all other points such that it is “correctly classified”!

SimCLR: a common approach for contrastive self-supervised learning



Contrastive loss

SimCLR formulation

$$z_i \xleftrightarrow{\text{Maximize agreement}} z_j$$

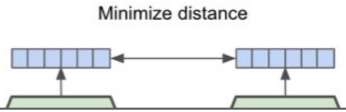
Contrastive loss can take the form of a familiar cross-entropy loss!

Loss for a pair of data points (i,j)

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

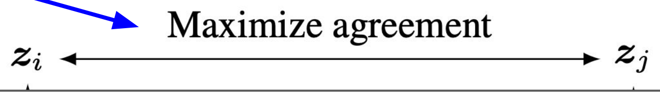
Compute loss over a minibatch of N examples. Generate two augmented views of each example, resulting in 2N data points total. Now in the contrastive loss, we wish for a pair of data points (i,j) corresponding to augmentations of the same example to have closer representation similarity than with other data points generated from different examples. Use a cross-entropy formulation: given data point i, similarity with data point j should have higher score than with all other points such that it is “correctly classified”!

SimCLR: a common approach for contrastive self-supervised learning



Contrastive loss

SimCLR formulation

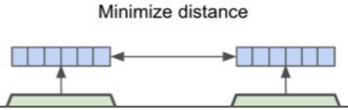


Contrastive loss can take the form of a familiar cross-entropy loss!

Similarity score between final-layer representations of i and j

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

SimCLR: a common approach for contrastive self-supervised learning



Contrastive loss

SimCLR formulation

Maximize agreement

$\boldsymbol{z}_i \longleftrightarrow \boldsymbol{z}_j$

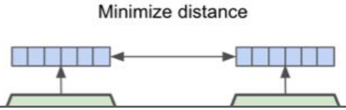
Contrastive loss can take the form of a familiar cross-entropy loss!

Similarity score between final-layer representations of i and j

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

Use cosine similarity $\mathrm{sim}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v} / \|\boldsymbol{u}\| \|\boldsymbol{v}\|$

SimCLR: a common approach for contrastive self-supervised learning



Contrastive loss

SimCLR formulation

Maximize agreement

$z_i \longleftrightarrow z_j$

Contrastive loss can take the form of a familiar cross-entropy loss!

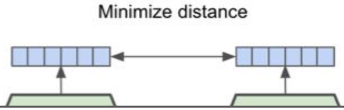
Exponentiate

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

From here, looks very similar to softmax loss (generalized cross entropy to multiple classes)

$$L_{Softmax} = \frac{1}{M} \sum_i -\log\left(\frac{e^{s_{y_i}}}{\sum_j e^{s_j}}\right)$$

SimCLR: a common approach for contrastive self-supervised learning



Contrastive loss

SimCLR formulation

Maximize agreement

$z_i \longleftrightarrow z_j$

Contrastive loss can take the form of a familiar cross-entropy loss!

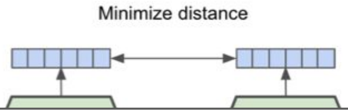
$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

Detail: Loss uses a temperature hyperparameter, controls peakiness of final probability distribution for better learning dynamics

From here, looks very similar to softmax loss (generalized cross entropy to multiple classes)

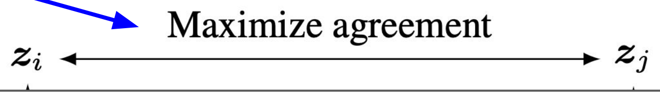
$$L_{Softmax} = \frac{1}{M} \sum_i -\log\left(\frac{e^{s_{y_i}}}{\sum_j e^{s_j}}\right)$$

SimCLR: a common approach for contrastive self-supervised learning



Contrastive loss

SimCLR formulation



Contrastive loss can take the form of a familiar cross-entropy loss!

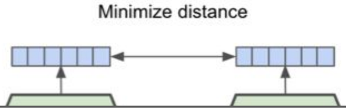
$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

Normalize over scores of similarity between i and all other data points in the minibatch (2N total)

From here, looks very similar to softmax loss (generalized cross entropy to multiple classes)

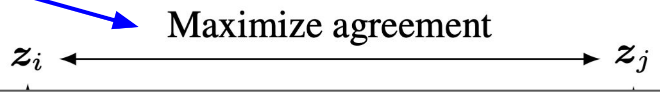
$$L_{Softmax} = \frac{1}{M} \sum_i -\log\left(\frac{e^{s_{y_i}}}{\sum_j e^{s_j}}\right)$$

SimCLR: a common approach for contrastive self-supervised learning



Contrastive loss

SimCLR formulation



Contrastive loss can take the form of a familiar cross-entropy loss!

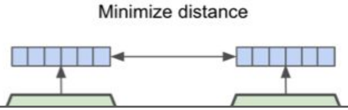
Negative log likelihood, as in softmax / cross-entropy

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

From here, looks very similar to softmax loss (generalized cross entropy to multiple classes)

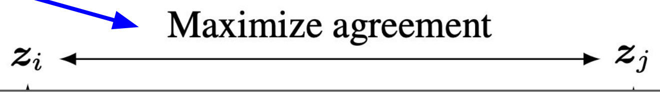
$$L_{Softmax} = \frac{1}{M} \sum_i -\log\left(\frac{e^{s_{y_i}}}{\sum_j e^{s_j}}\right)$$

SimCLR: a common approach for contrastive self-supervised learning



Contrastive loss

SimCLR formulation



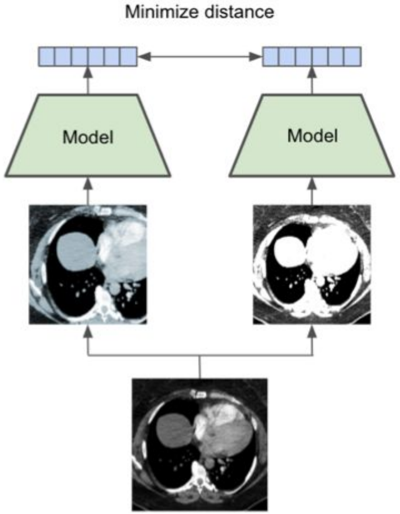
Contrastive loss can take the form of a familiar cross-entropy loss!

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

For a minibatch of N examples (2N augmented data points), compute this loss over all corresponding pairs (i,j), as well as (j,i) for symmetry of the loss, and then average these individual loss terms (2N terms total)

$$\mathcal{L} = \frac{1}{2N} \sum_{k=1}^{N} [\ell(2k-1, 2k) + \ell(2k, 2k-1)]$$

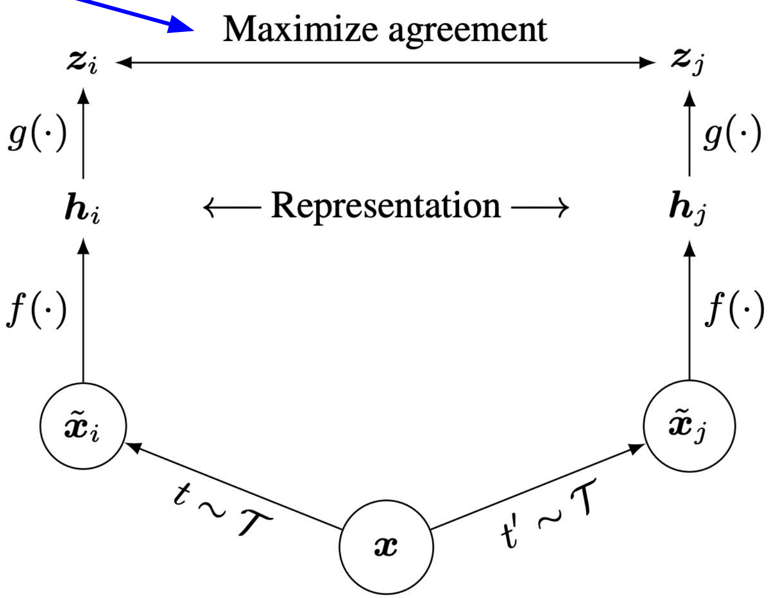
SimCLR: a common approach for contrastive self-supervised learning



Contrastive objective
Different views of the same input should have more similar representation to each other than with a different input

Contrastive loss

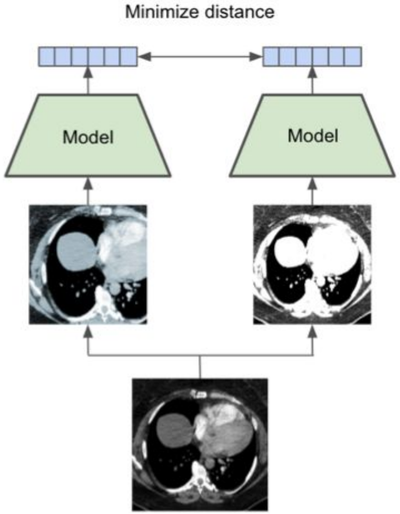
SimCLR formulation



Input

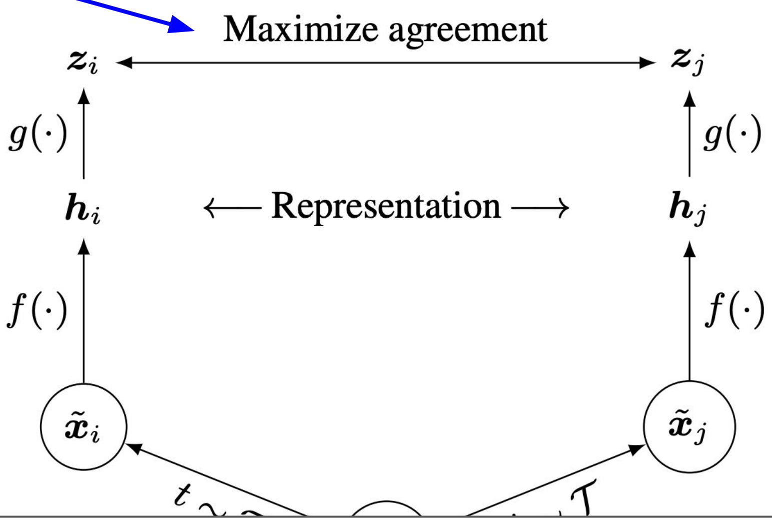
Chen et al. 2020

SimCLR: a common approach for contrastive self-supervised learning



Contrastive loss

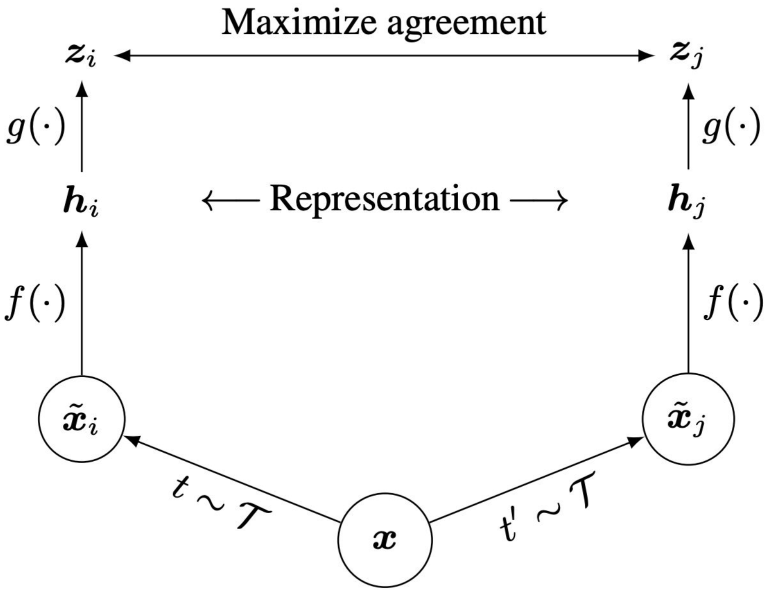
SimCLR formulation



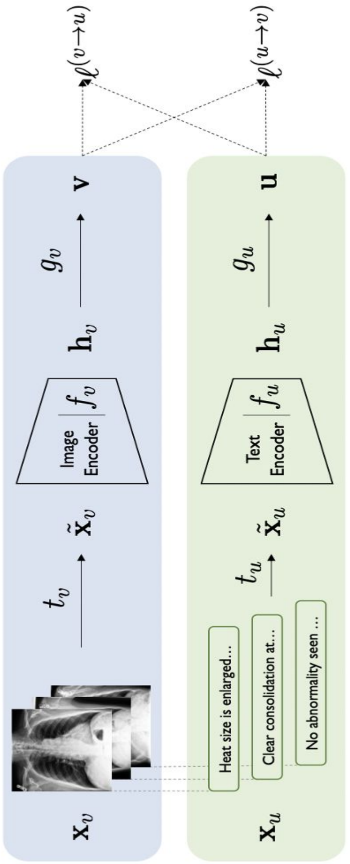
After self-supervised training, can fine-tune the encoder *f* on smaller labeled datasets. Can also directly extract learned representations h for downstream tasks.

Multimodal contrastive learning

SimCLR (single-modality)

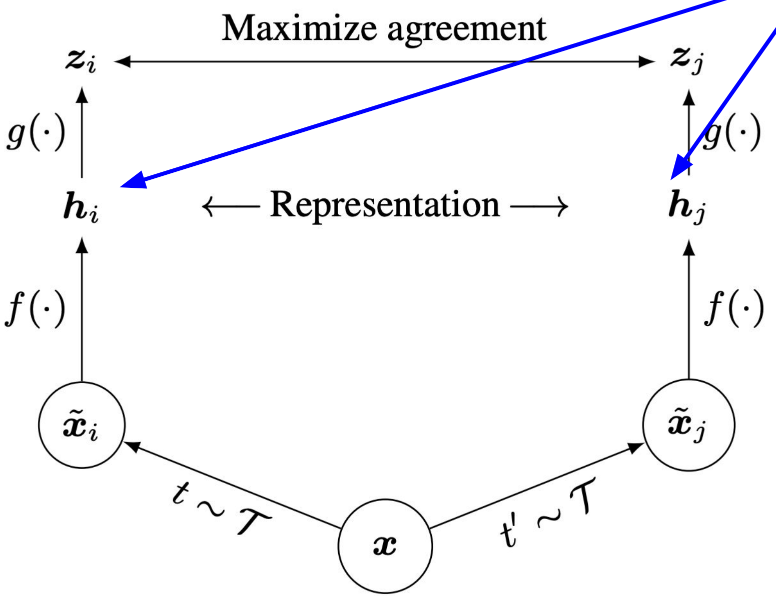


ConVIRT (multi-modality)



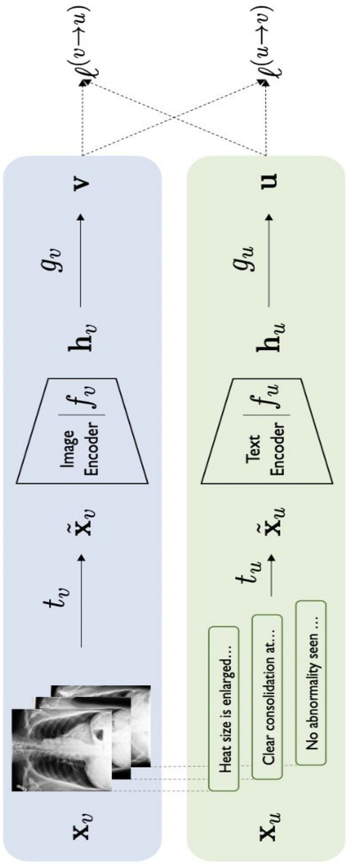
Multimodal contrastive learning

SimCLR (single-modality)



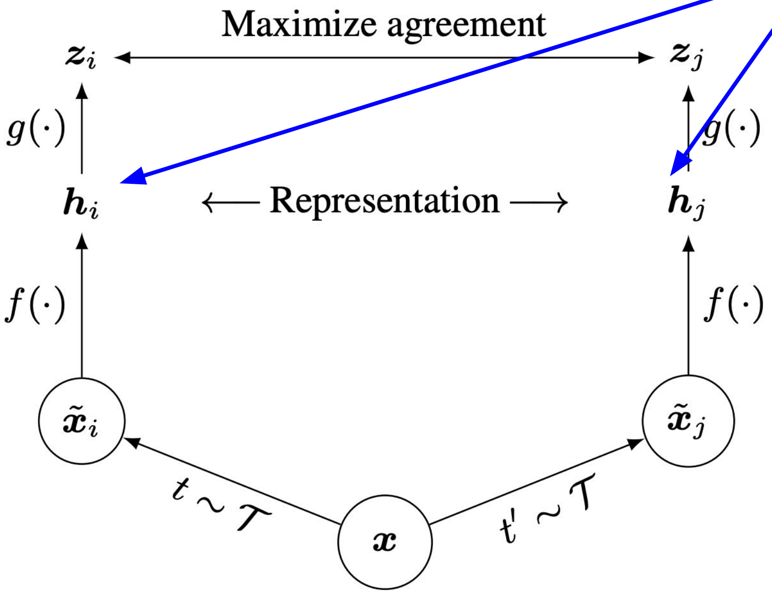
In **single-modality contrastive learning**, representations h are **shared-encoder** outputs of two different augmentations of the same input. Want augmentations corresponding to the same input to be more similar to each other than to those corresponding to different inputs

ConVIRT (multi-modality)



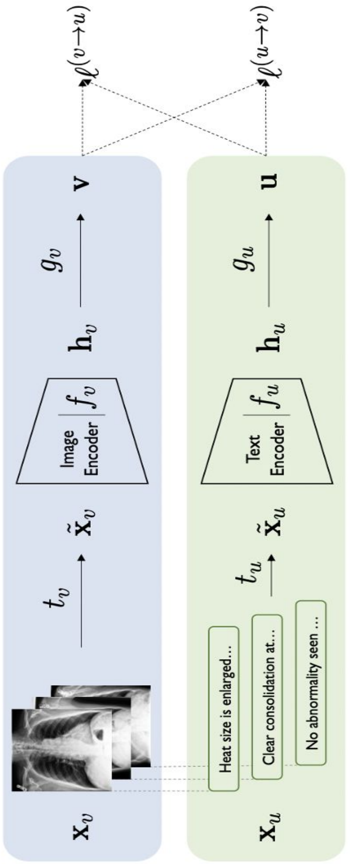
Multimodal contrastive learning

SimCLR (single-modality)



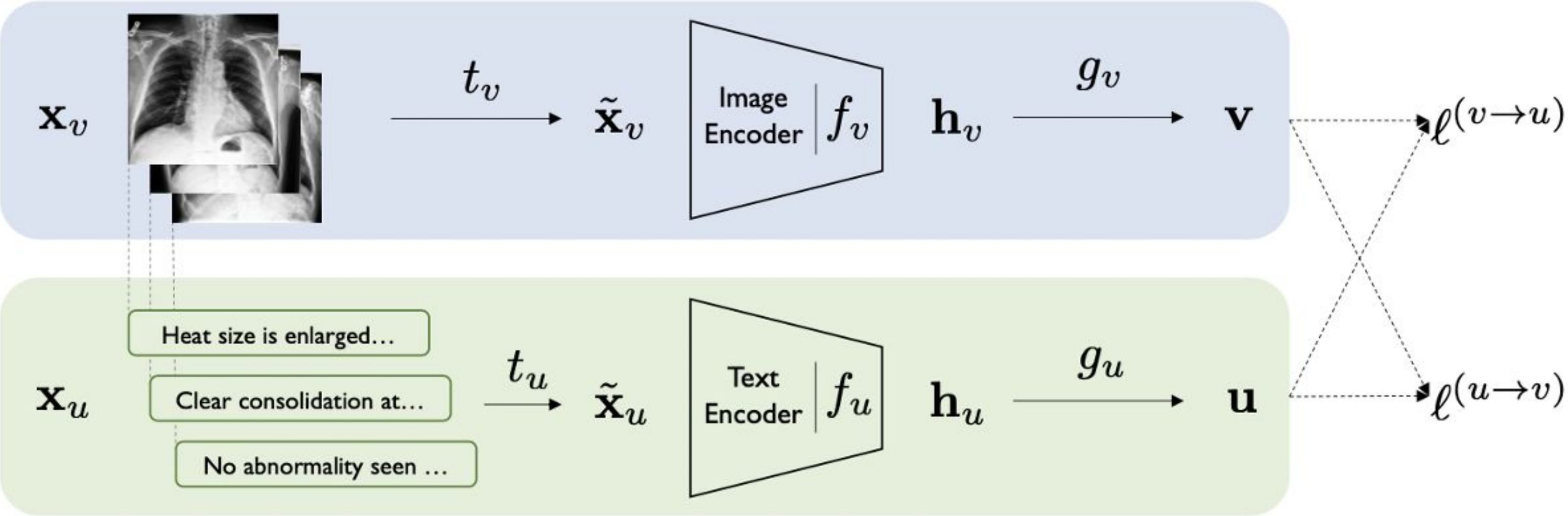
In **multi-modality contrastive learning**, representations h are encoder outputs of the same concept (e.g. radiology image and corresponding report), from **two different modality-specific encoders**. Want these to be more similar to each other than with non-corresponding images / reports.

ConVIRT (multi-modality)



ConVIRT

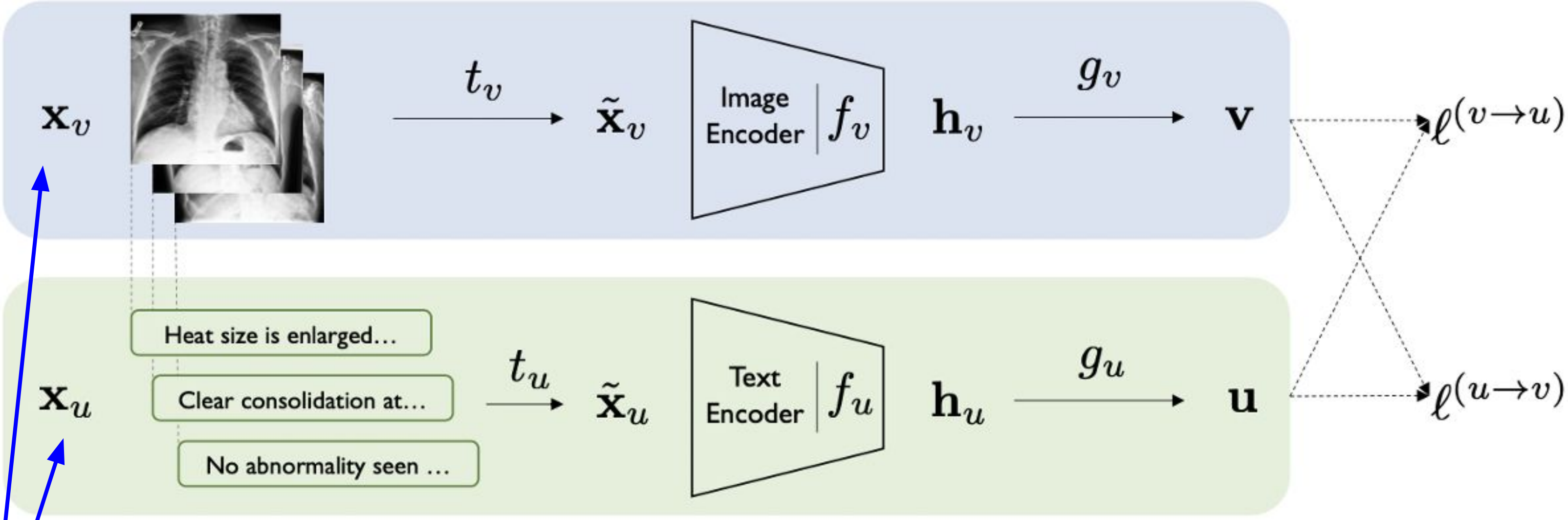
Multimodal contrastive pre-training on 217k image-text from the MIMIC-CXR dataset



Zhang et al. 2020.

ConVIRT

Multimodal contrastive pre-training on 217k image-text from the MIMIC-CXR dataset

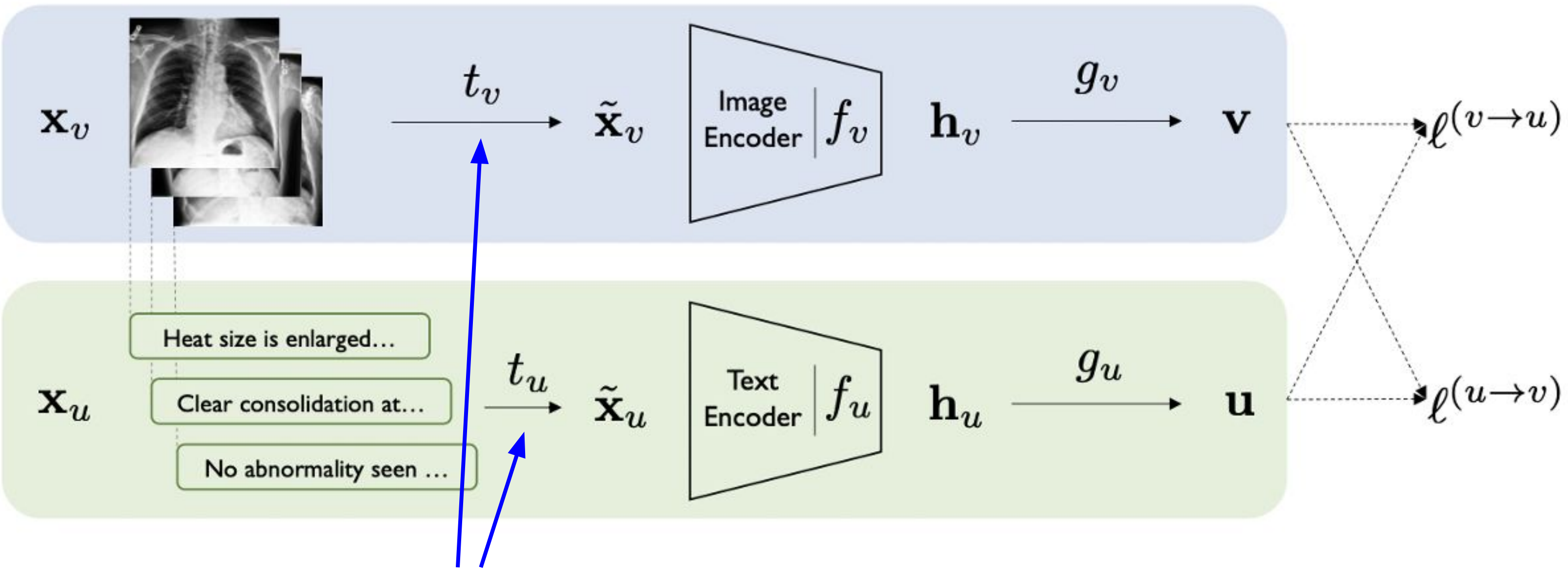


Input image/text data

Zhang et al. 2020.

ConVIRT

Multimodal contrastive pre-training on 217k image-text from the MIMIC-CXR dataset

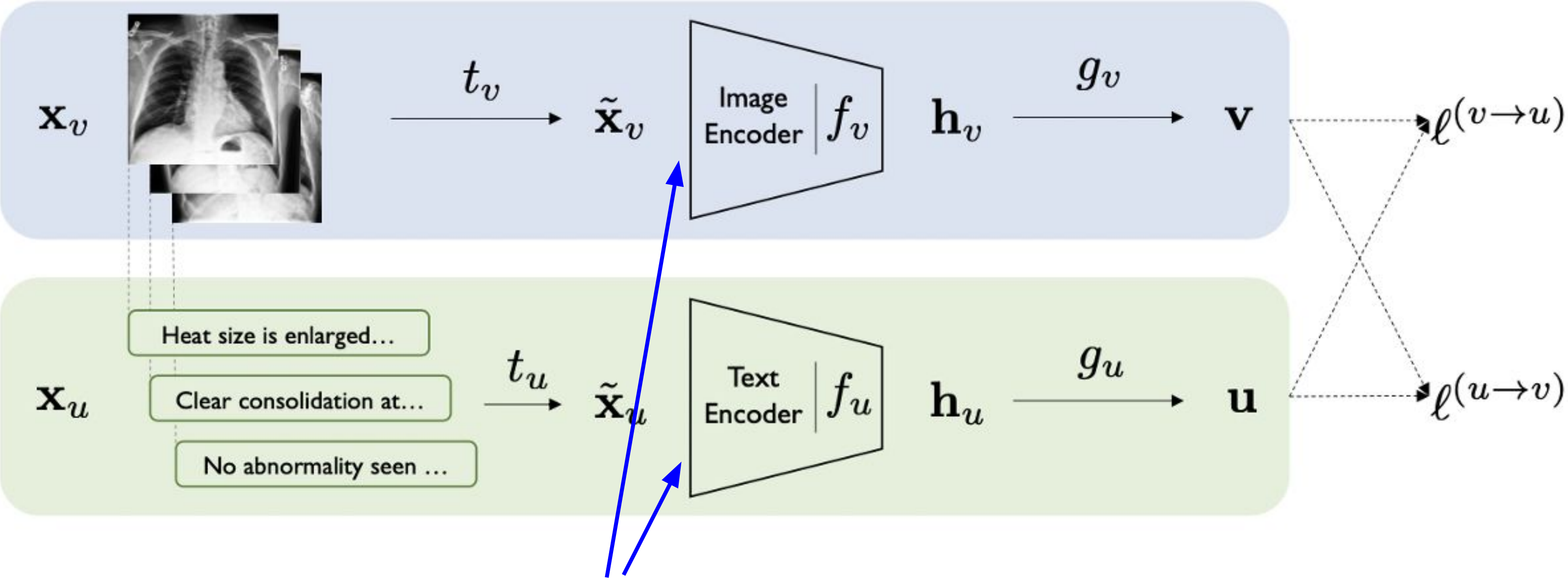


Modality-specific sampling
and transformation

Zhang et al. 2020.

ConVIRT

Multimodal contrastive pre-training on 217k image-text from the MIMIC-CXR dataset

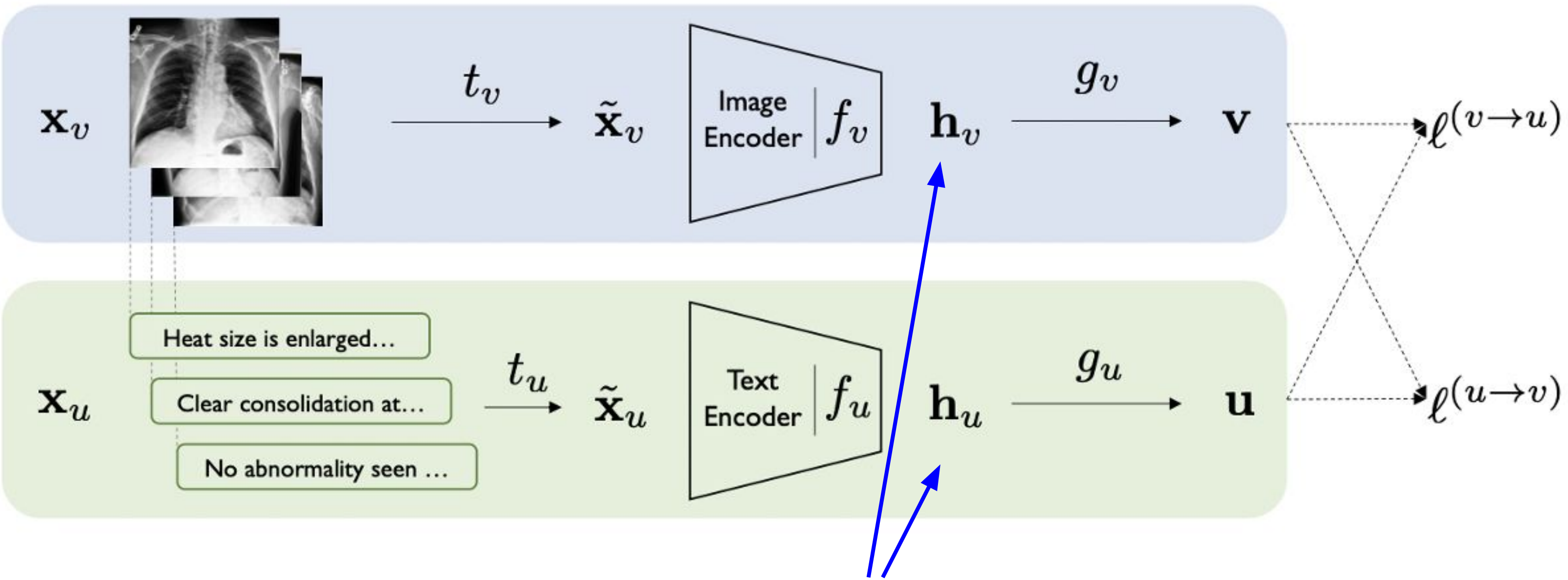


Modality-specific encoders: ResNet-50 for images and BERT (initialized with ClinicalBERT) for text. Only fine-tune last 6 layers of BERT encoder during pre-training.

Zhang et al. 2020.

ConVIRT

Multimodal contrastive pre-training on 217k image-text from the MIMIC-CXR dataset

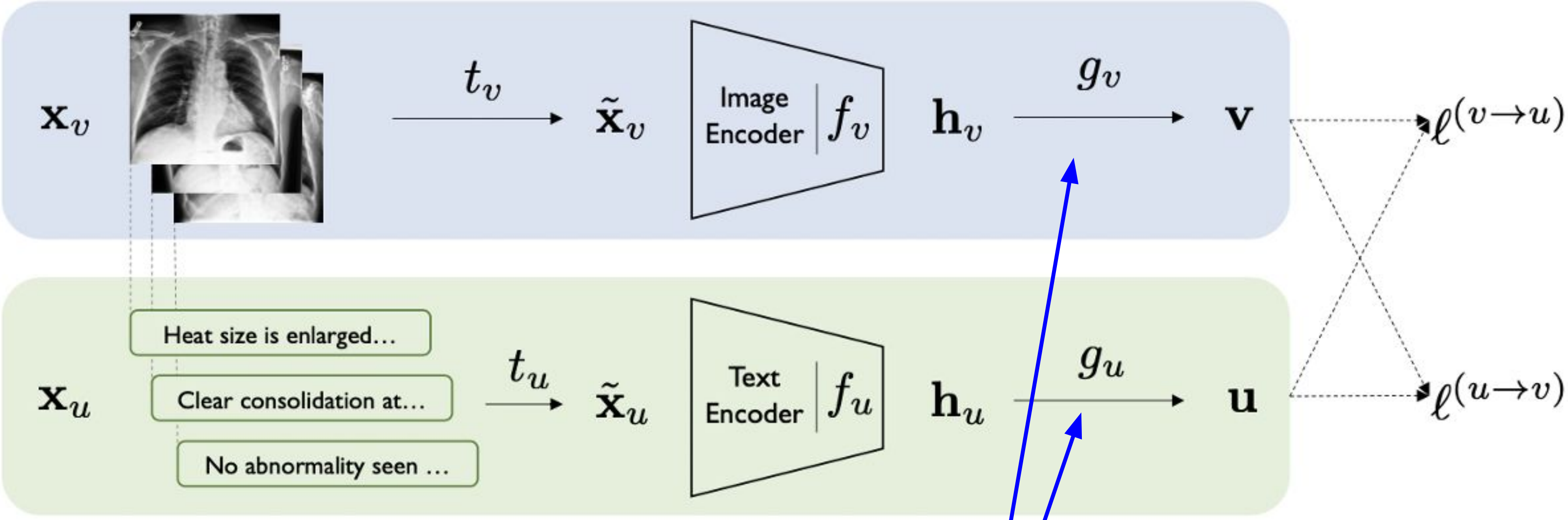


Encoded representations of image and text

Zhang et al. 2020.

ConVIRT

Multimodal contrastive pre-training on 217k image-text from the MIMIC-CXR dataset

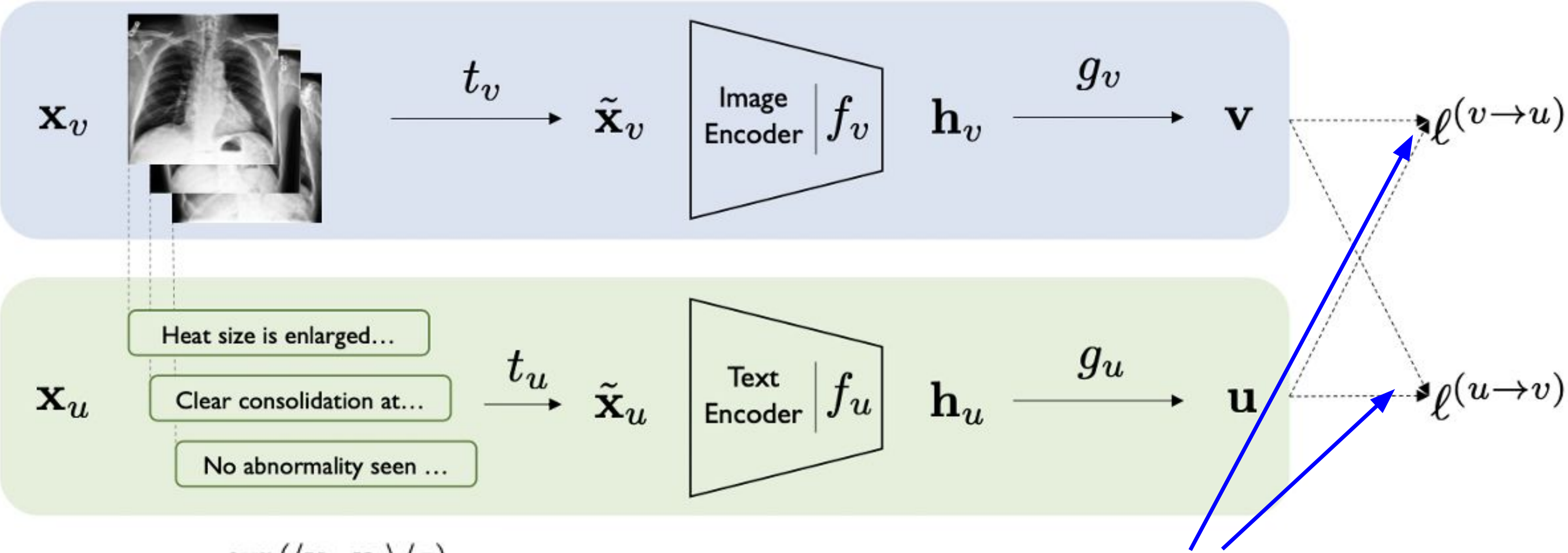


Small projection function (MLP) used only during contrastive learning, not downstream task fine-tuning, as with SimCLR

Zhang et al. 2020.

ConVIRT

Multimodal contrastive pre-training on 217k image-text from the MIMIC-CXR dataset



$$\ell_i^{(v \to u)} = -\log \frac{\exp(\langle \mathbf{v}_i, \mathbf{u}_i \rangle / \tau)}{\sum_{k=1}^{N} \exp(\langle \mathbf{v}_i, \mathbf{u}_k \rangle / \tau)}$$

Same contrastive loss on projection function outputs, as in SimCLR. “Correct” matched pairs are now those from the same patient image/text case, different from the two augmented views of the same input in SimCLR.

Zhang et al. 2020.

GLORIA

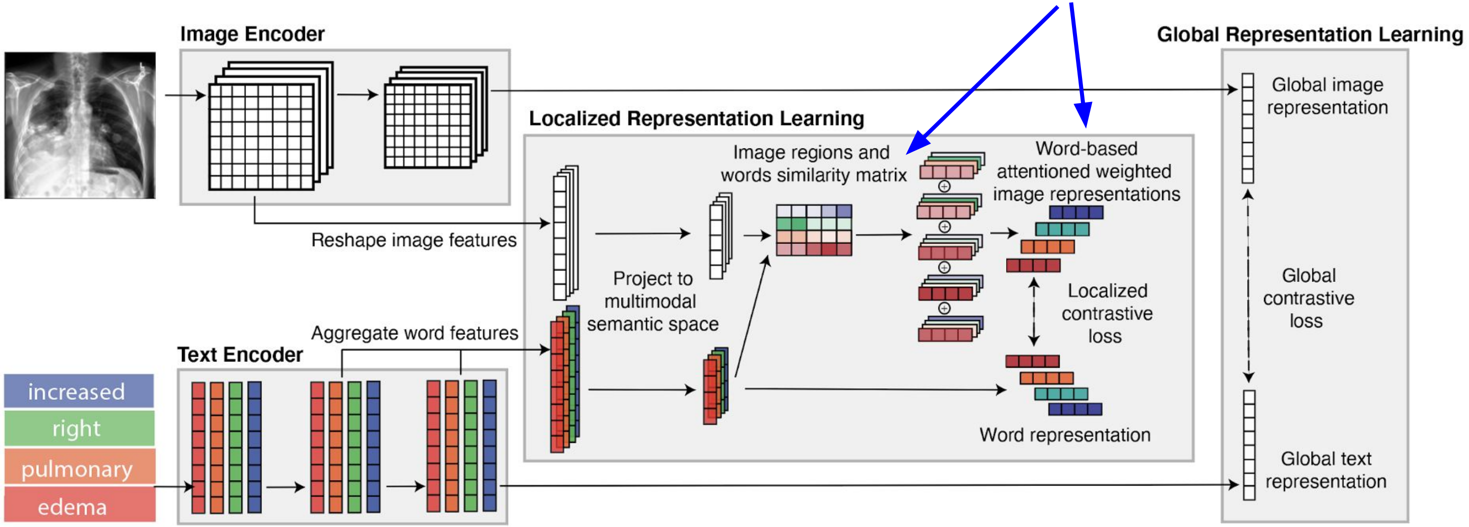
Many radiology reports are long – associating all parts of a report equally with all regions of an image may be too coarse

Huang et al. 2021.

GLORIA

Many radiology reports are long – associating all parts of a report equally with all regions of an image may be too coarse

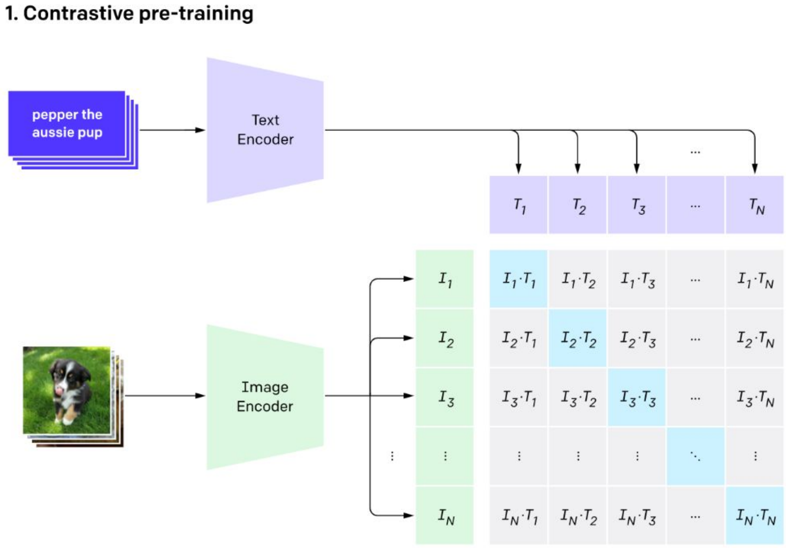
Extension to ConVIRT: beyond global contrastive loss, jointly train with a localized contrastive loss between words and attention-weighted regions of images (learn the attention weighting, as in previous lectures)



Huang et al. 2021.

CLIP

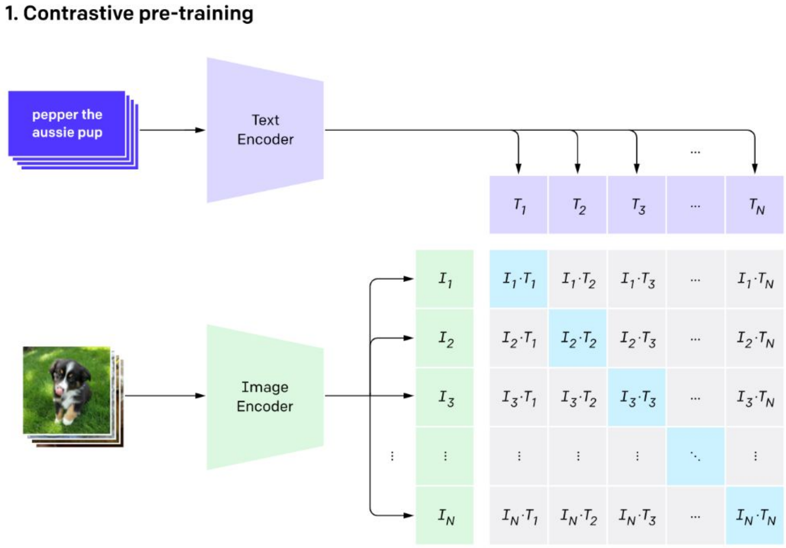
Multimodal contrastive learning similar to ConVIRT, but now on very large dataset of 400 million image-text pairs



Radford et al. 2021.

CLIP

Multimodal contrastive learning similar to ConVIRT, but now on very large dataset of 400 million image-text pairs



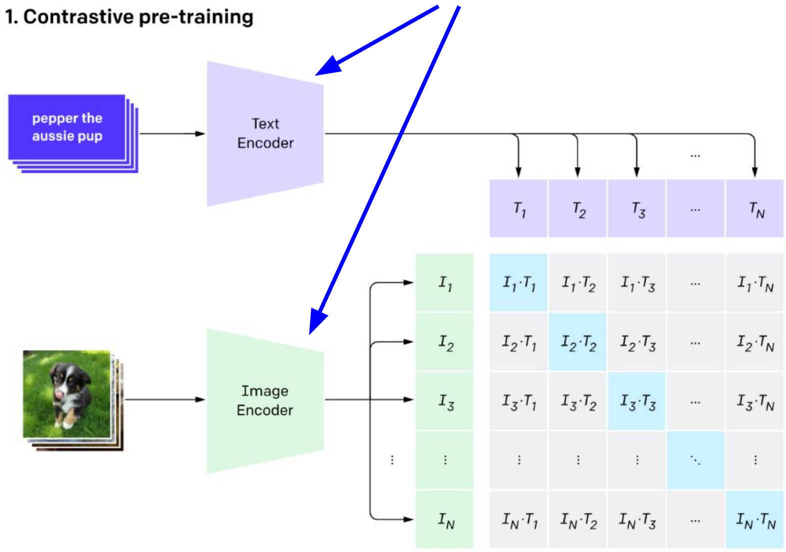
Dataset generated by searching for image-text pairs on the web, where text comes from a base query list of 500,000 queries comprising all words occurring at least 100 times in the English version of Wikipedia. This is augmented and processed in various ways, see paper for details.

Radford et al. 2021.

CLIP

Multimodal contrastive learning similar to ConVIRT, but now on very large dataset of 400 million image-text pairs

Transformer-based, trained from scratch

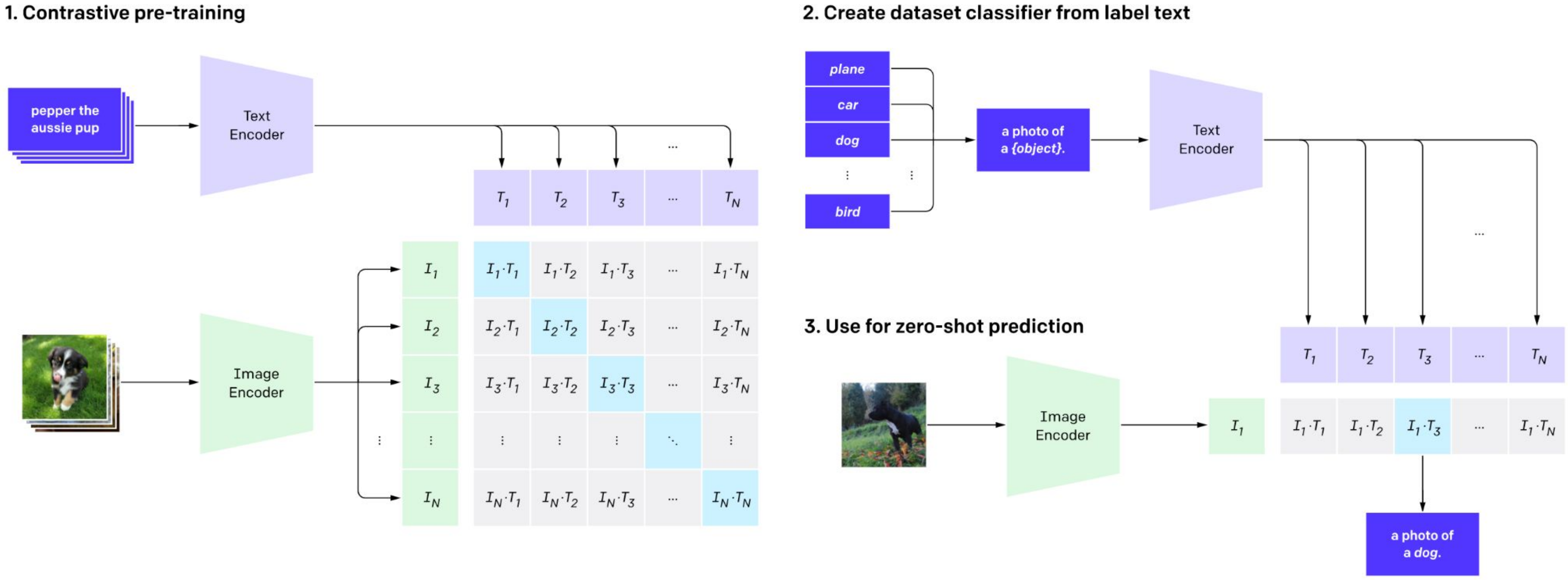


Radford et al. 2021.

CLIP

Multimodal contrastive learning similar to ConVIRT, but now on very large dataset of 400 million image-text pairs

Can be used for **zero-shot** prediction tasks



Radford et al. 2021.

Complementary to self-supervision: **weak supervision** is another class of methods to improve learning in limited label scenarios

- Machine learning paradigm where labels for supervised training are obtained from noisy or imprecise (but more easily accessible) sources
- One possibility is through corresponding data available in a different modality! (e.g., radiology reports as a source of weak supervision for radiology images)

Weak supervision from radiology reports

Can use rule-based approaches for obtaining labels from free-text radiology reports

```
Indication: Chest pain. Findings:
Mediastinal contours are within
normal limits. Heart size is
within normal limits. No focal
consolidation, pneumothorax or
pleural effusion. Impression: No
acute cardiopulmonary
abnormality.
```

Normal Report

```
def LF_pneumothorax(c):
  if re.search(r'pneumo.*', c.report.text):
    return "ABNORMAL"

def LF_pleural_effusion(c):
  if "pleural effusion" in c.report.text:
    return "ABNORMAL"

def LF_normal_report(c, thresh=2):
    if len(NORMAL_TERMS.intersection(c.
    report.words)) > thresh:
      return "NORMAL"
```

LFs

Figure credit: Nishith Khandwala et al., 2017.
Dunmon et al. Cross-Modal Data Programming Enables Rapid Medical Machine Learning, 2020.

How can we produce good labels from noisy sources?

One approach: Aggregate multiple rules (labeling functions) with majority voting

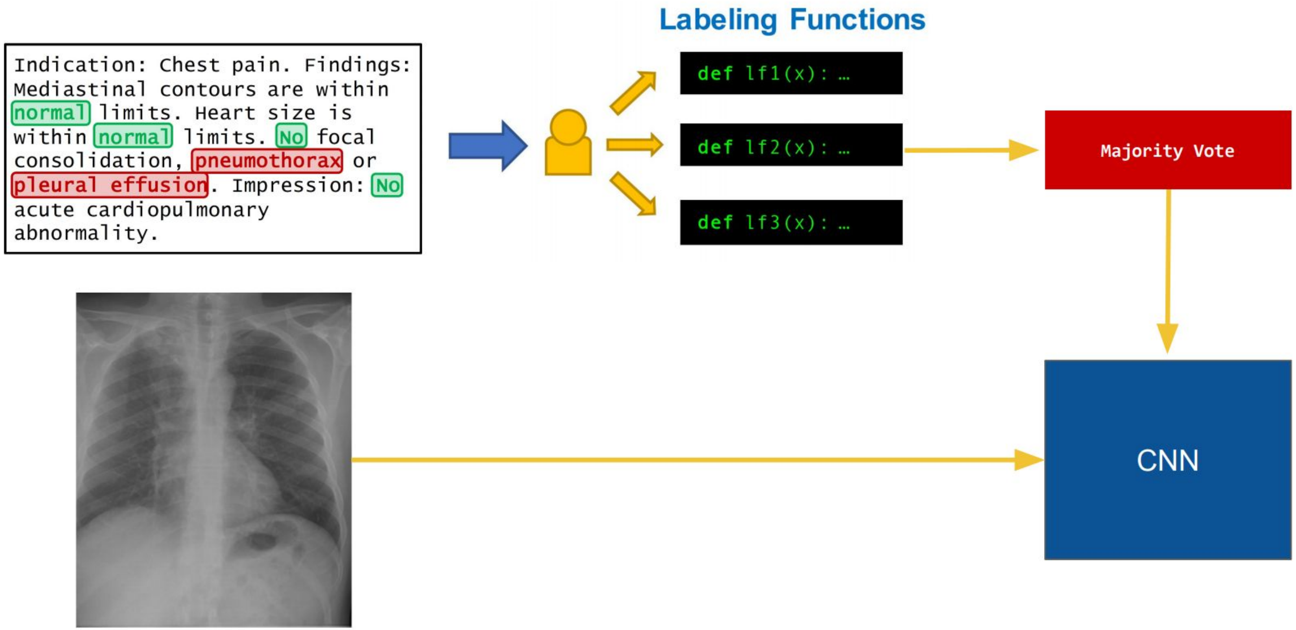


Figure credit: Nishith Khandwala et al., 2017.
Dunmon et al. Cross-Modal Data Programming Enables Rapid Medical Machine Learning, 2020.

How can we produce good labels from noisy sources?

More sophisticated approach: learn models for how to best aggregate noisy labeling functions!

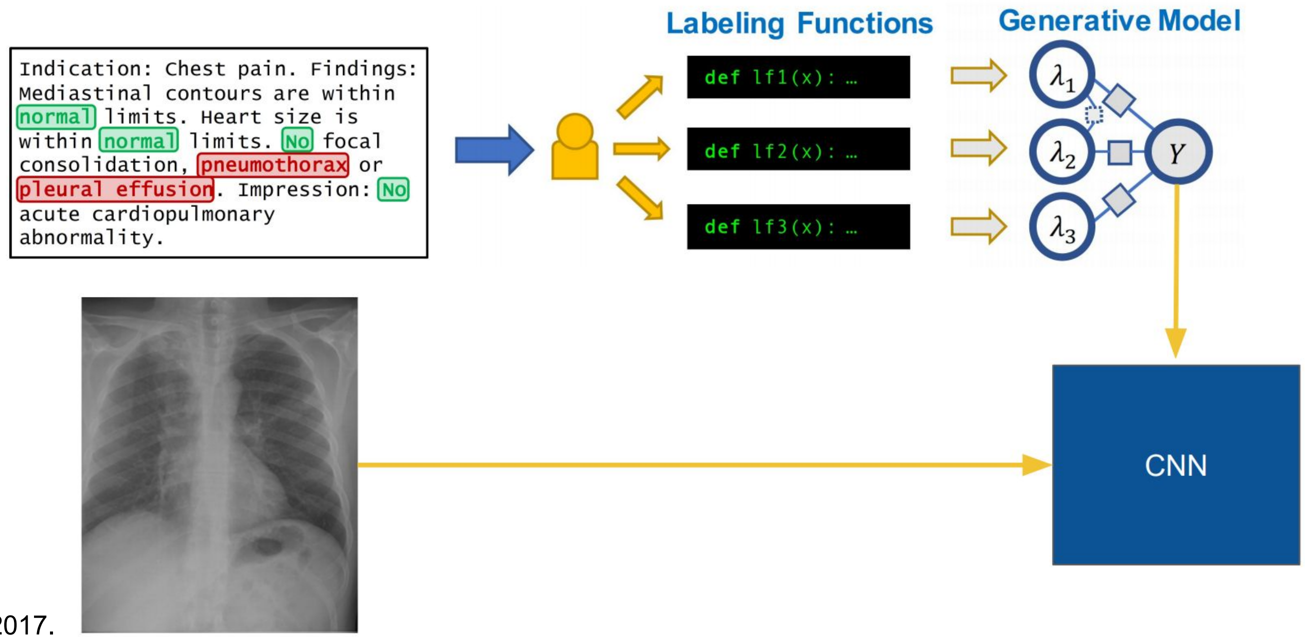


Figure credit: Nishith Khandwala et al., 2017.
Dunmon et al. Cross-Modal Data Programming Enables Rapid Medical Machine Learning, 2020.

How can we produce good labels from noisy sources?

More sophisticated approach: learn models for how to best aggregate noisy labeling functions!

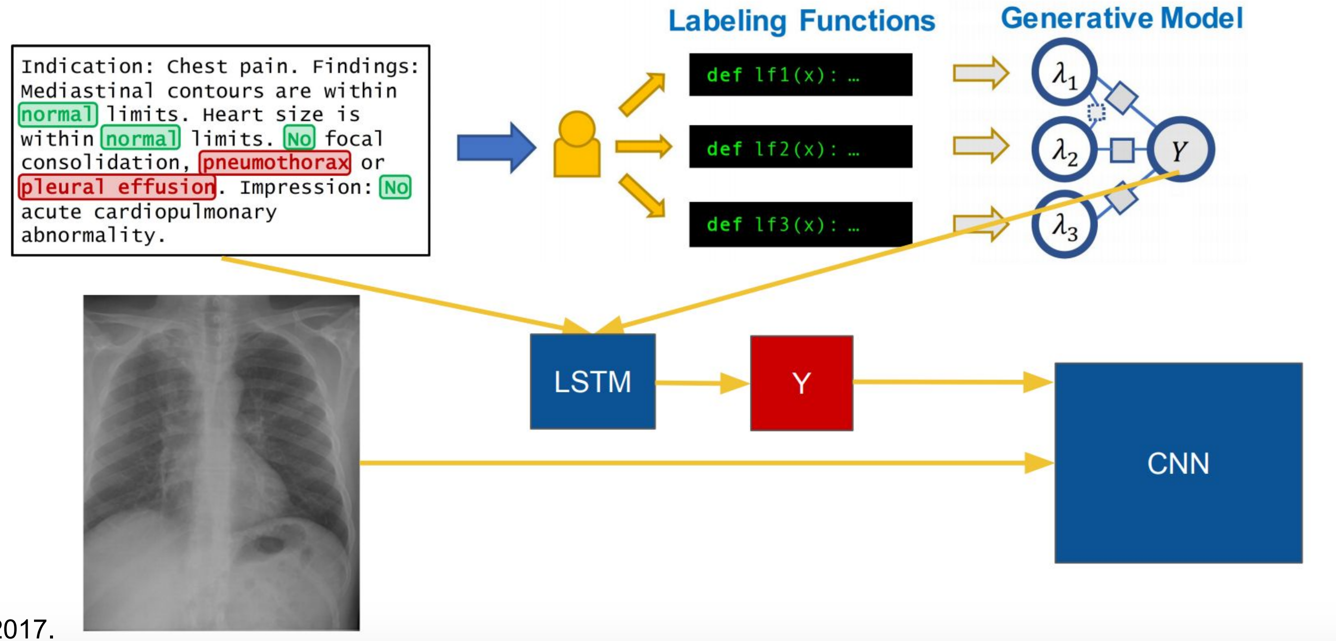
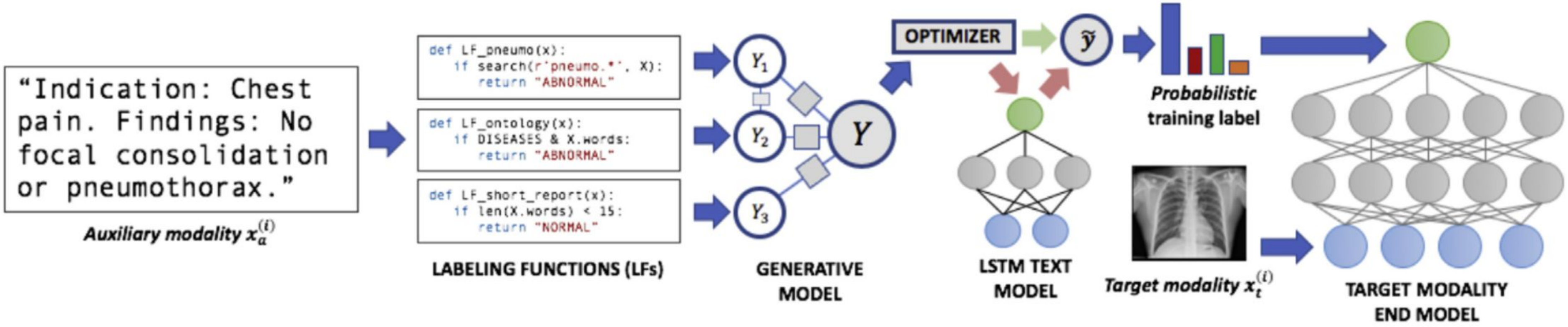


Figure credit: Nishith Khandwala et al., 2017.
Dunmon et al. Cross-Modal Data Programming Enables Rapid Medical Machine Learning, 2020.

"Data programming" paradigm for weak supervision



Dunmon et al. Cross-Modal Data Programming Enables Rapid Medical Machine Learning, 2020.

Summary

Today we covered:

- Multimodal data and models
- Self-supervised learning (including contrastive learning)
  - Both single-modality and multi-modality
- Weakly supervised learning

Next time:

- More on Transformers and Multimodal Models