DL Thinking

Konrad Kording and Lyle Ungar

neuromatch academy

Section 2: Architectures and multimodal DL Thinking

Konrad Kording and Lyle Ungar

Last time in DL thinking

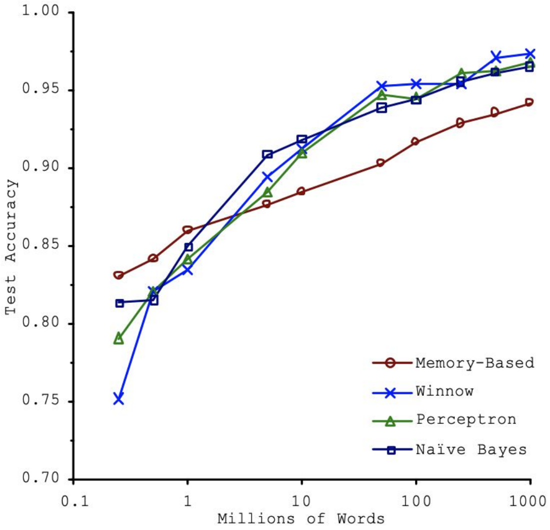
Cost functions - a way for us to build in what we want to achieve

Let us zoom out from last time a bit. What did we learn?

However, there are so many other things we can engineer.

Let us look at imagenet data

Reminder: data is great!



Lyle wears a hat

This is not a slide. We will make the video full screen at this time.

I give you dataset of cats and dogs. No you can’t have any other data. What can you tell me about dogs. Is something like a bigger dog still a dog? A smaller dog? What if the color was different? What if it was too different?

What should we do with the data

You design the strategy

TA guide

Hint 1: Look at a few photos of dogs (use an image search engine). How are they similar? How are they different? What makes them all be dogs?

Hint 2: Think about color, orientation, flipping, pixel noise, color noise, shearing, contrast, brightness, scaling

Hint 3: Discuss where each of these ideas will break down. Can too much of a good thing be good?

Solution: See next slide

Advanced: Ask how all of these strategies may vary by object class

Much of the progress of image recognition has been due to data augmentation!

The unsung hero of Machine Learning!

Here are typical augmentations

f_h	Horiz. flip	$1 - 2B(0.5)$
t_x	Horiz. translation	$\mathcal{U}(-0.1, 0.1)$
t_y	Vert. translation	$\mathcal{U}(-0.1, 0.1)$
z_x	Horiz. scale	$\mathcal{U}(0.85, 1.15)$
z_y	Vert. scale	$\mathcal{U}(0.85, 1.15)$
θ	Rotation angle	$\mathcal{U}(-22.5°, 22.5°)$
ϕ	Shear angle	$\mathcal{U}(-0.15, 0.15)$
γ	Contrast	$\mathcal{U}(0.5, 1.5)$
δ	Brightness	$\mathcal{U}(-0.25, 0.25)$

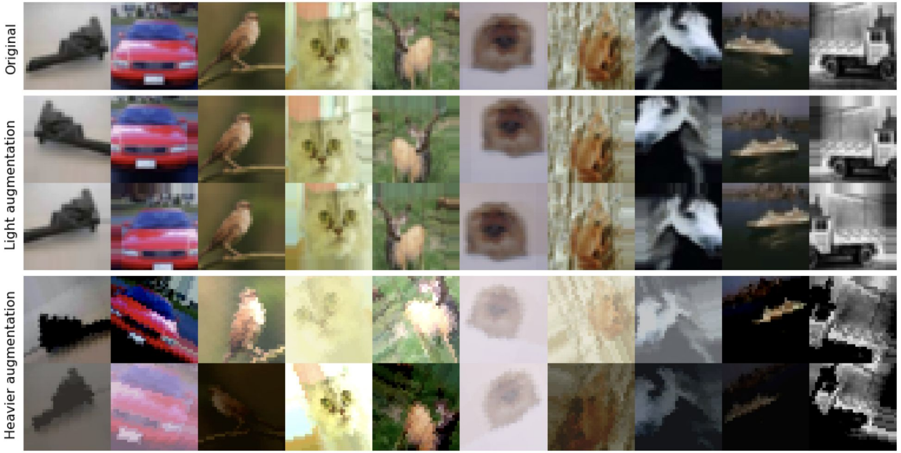
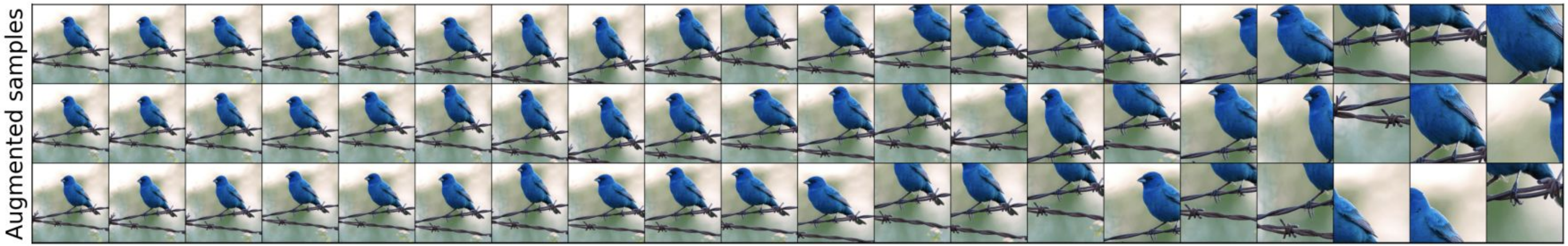


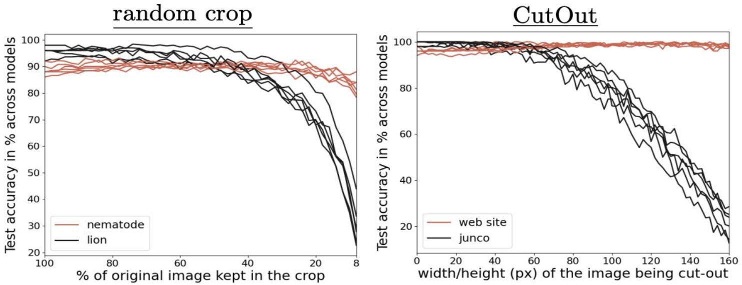
Figure 4.1: Illustration of the most extreme transformations performed by the data augmentation schemes on ten images—one per class—from CIFAR-10.

Class dependence of augmentation



The Effects of Regularization and Data Augmentation are Class Dependent

Randall Balestriero[1], Léon Bottou[1], and Yann LeCun[1,2]



Konrad wears a hat

This is not a slide. We will make the video full screen at this time.

So, I work for med school. We have lots of photos of brains. Some have cancers. I want to automatically detect them. But I only have 10,000 of them (shock). Lyle: Tell me how they look. I talk about tumors being like objects. Rotation symmetry. Left right. But also squishing. I want to use that things are similar to other objects.

What do we want?

Good performance

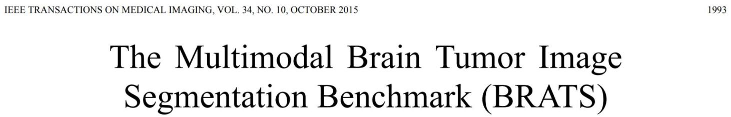
Despite small dataset

Lets talk:

Strategy for data

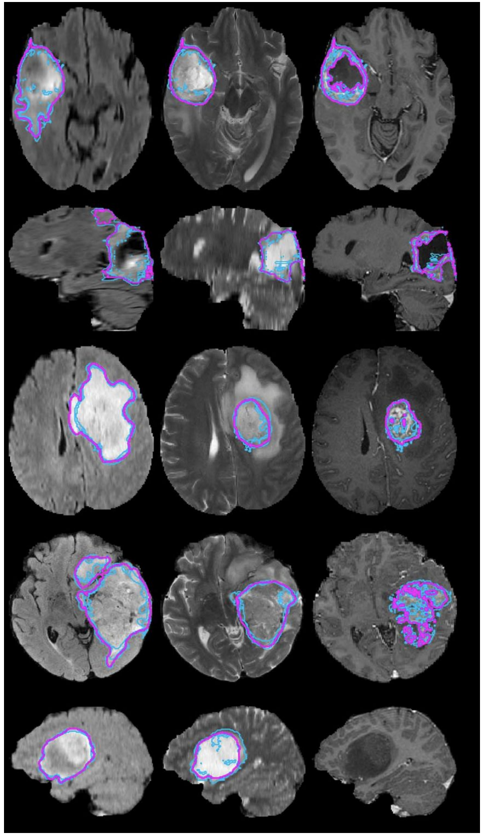
Image courtesy Wikipedia / spring of hope

How do tumors look like?



IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. 34, NO. 10, OCTOBER 2015 1993

The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)



What do we want to use

Images of things do similar things to images of tumors

Lets exploit that!

But also exploit whatever else we know about data

So which data architecture/ ideas should we use?

You design the data strategy

Hints

Hint 1: Data augmentation is always something to consider.

Hin 2: Could you train on another dataset first? What properties should such a dataset have?

Solution: Always do data augmentation. First pre-train on something like imagenet. Imagenet is kinda similar to tumors. Then retrain whole thing (or top layers only) on the tumor dataset. Even better if there exist other datasets that are big and are similar to tumors (say xray image datasets).

Advanced: Think about tradeoffs in the strategy for pre-training. Consider the use of multiple datasets. Consider hyperparameter optimization.

First idea: Data augmentation

Consider symmetries (flip left right).

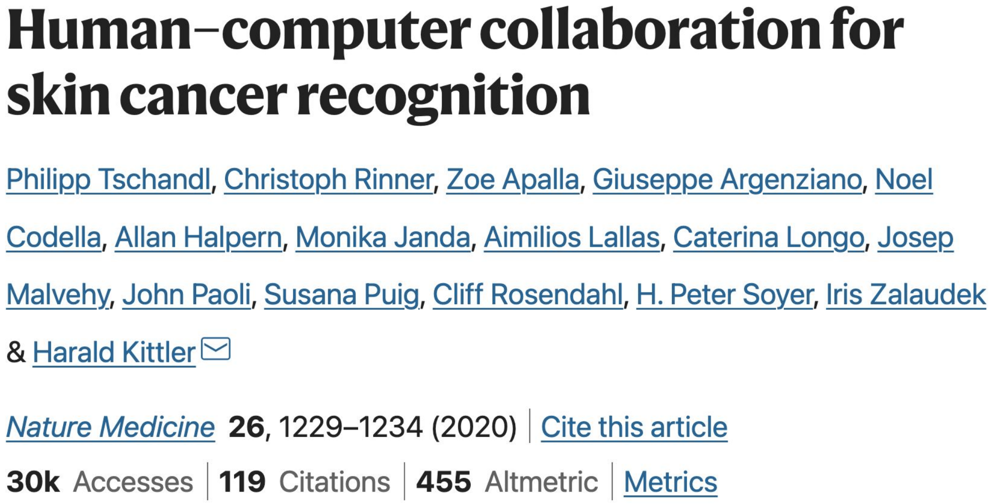
Consider small rotations (or any rotation).

Consider illumination noise (but no color)

Second Idea: Pretrain on imagenet (or other image data)

Because cancers are a bit like other visual objects (coherent, local transformations, soft)

Example paper



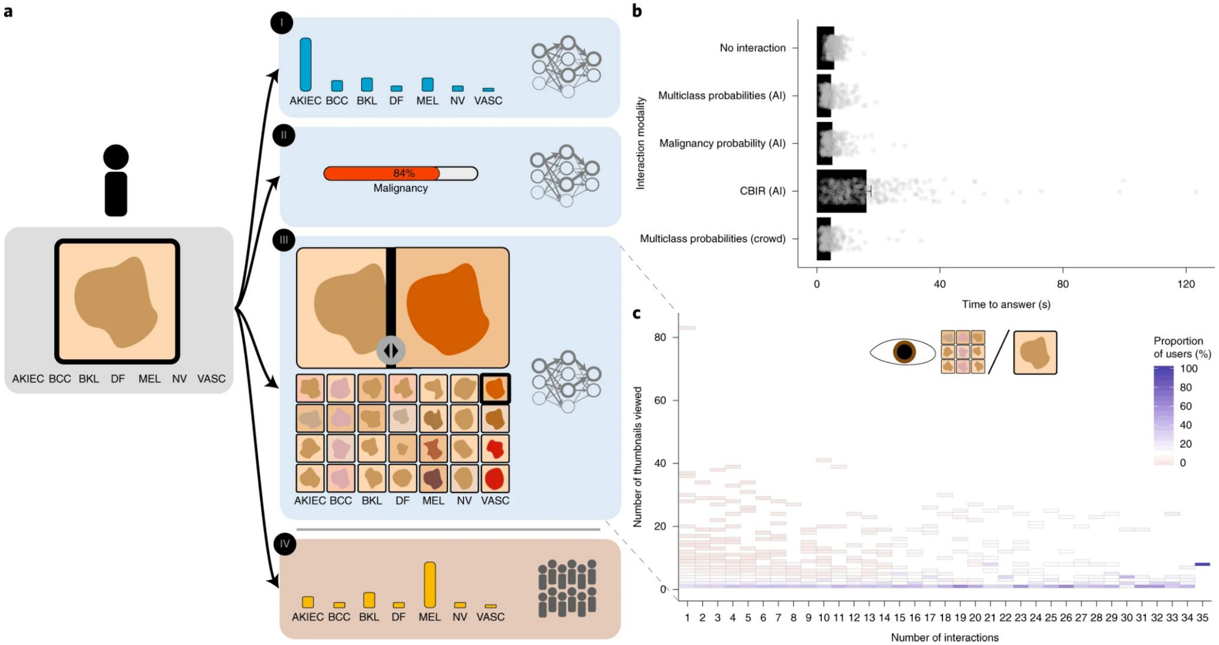
Human–computer collaboration for skin cancer recognition

Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, John Paoli, Susana Puig, Cliff Rosendahl, H. Peter Soyer, Iris Zalaudek & Harald Kittler ✉

Nature Medicine **26**, 1229–1234 (2020) | Cite this article

30k Accesses | **119** Citations | **455** Altmetric | Metrics

Collaborate with humans



Konrad wears a hat

This is not a slide. We will make the video full screen at this time.

So, I have two paired datasets - a long video with all kinds of labels and a dataset of simultaneously recorded brain data. I want to figure out what these two datasets have in common. Data high-d. Brain data and video data change roughly every second. We believe that coding in both is nonlinear.

What do we want?

Pull the shared information

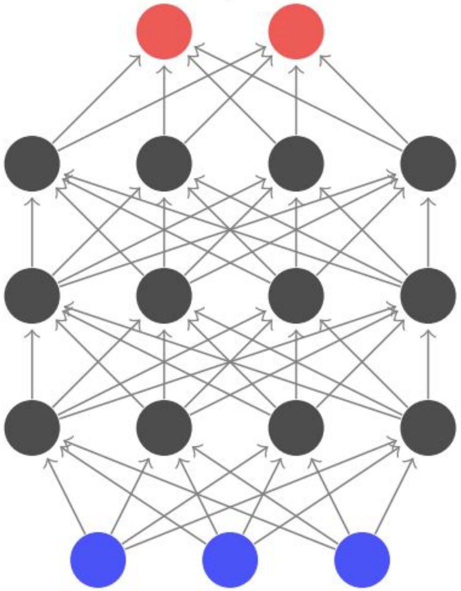
From two data modalities

Lets talk:

Strategy for data

Image courtesy Wikipedia / spring of hope

I believe that I could use an ANN to extract relevant information, from each



Cost

They should be as related

Or as correlated as possible

So which data architecture/ cost function should we use?

You design both!

TA guide

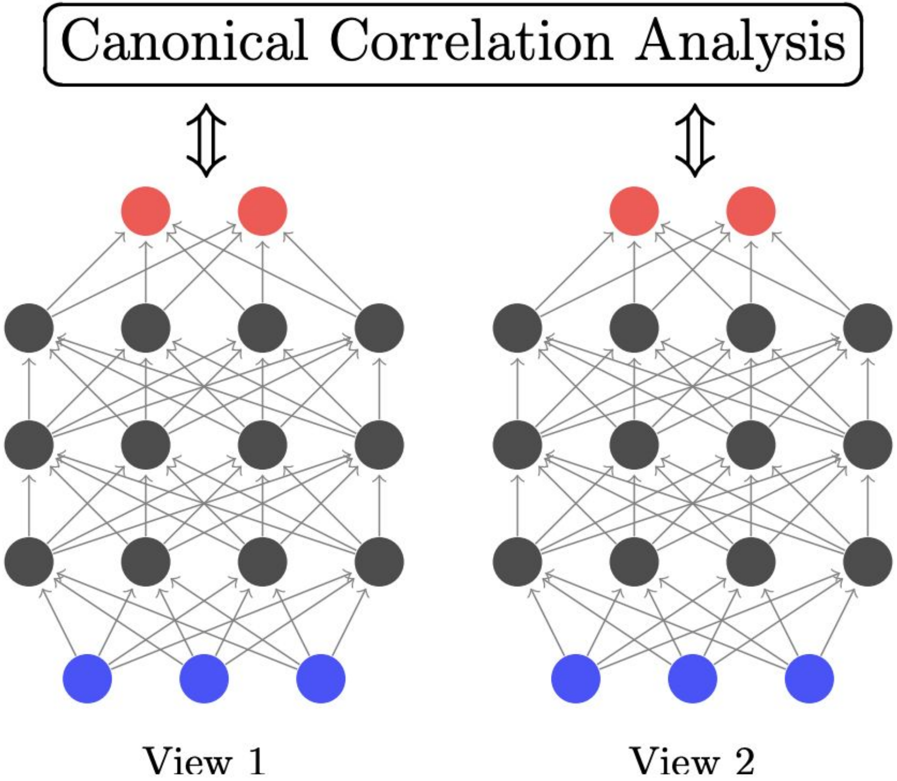
Hint 1: We want the two datasets to share something. What does that mean?

Hint 2: We want the two embeddings to have correlations. What kind of correlations? Well, a linear one. That carries a lot of variance. How to measure that?

Hint 3: What happens if we multiply all activities by 2? Need a scale invariant solution.

Solution: Equations from next slide

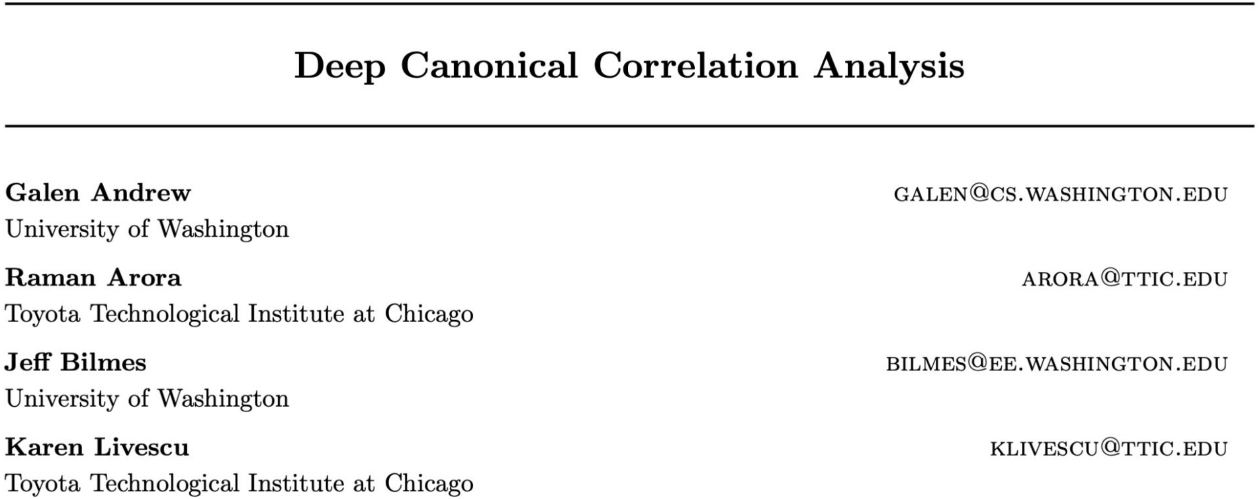
Advanced: How can you combine these ideas with other DL approaches?


Canonical Correlation Analysis
View 1
View 2

Regular CCA

$$
\begin{aligned}
(w_1^*, w_2^*) &= \underset{w_1, w_2}{\operatorname{argmax}} \; \operatorname{corr}(w_1' X_1, w_2' X_2) \\
&= \underset{w_1, w_2}{\operatorname{argmax}} \; \frac{w_1' \Sigma_{12} w_2}{\sqrt{w_1' \Sigma_{11} w_1 w_2' \Sigma_{22} w_2}}.
\end{aligned}
$$

Example paper



Deep Canonical Correlation Analysis

Galen Andrew GALEN@CS.WASHINGTON.EDU
University of Washington

Raman Arora ARORA@TTIC.EDU
Toyota Technological Institute at Chicago

Jeff Bilmes BILMES@EE.WASHINGTON.EDU
University of Washington

Karen Livescu KLIVESCU@TTIC.EDU
Toyota Technological Institute at Chicago

How do you think ideas in deep learning are made? Which ones are you most likely to miss?

Discuss with your pod

TA guide

Hint 1: Problem knowledge: what do we want to achieve?

Hint 2: Domain knowledge: what can we build in

Hint 3: Algorithm agility (needed to implement that)

Hint 4: Whom can we ask for such knowledge

Data thinking

Almost every paper you read contains interesting data thinking

It is mixed in with a lot of standard things.

All courses teach the standard things

Go out and do the domain specific things.

Lyle wears a hat

This is not a slide. We will make the video full screen at this time.

You have a really big dataset of text. Say all of wikipedia. You want to tell Konrad which papers to read. What does it mean to understand text. Have models that maximize predictions within the same text. Goal is to have a good metric between paper embeddings.

Background: talk recommenders

Use embedding to describe each talk by a vector

Estimate interest vector for each user (e.g. by averaging the vectors associated with talks they liked)

Then use algorithm to choose what to show

Try out Konrad's groups current UPenn talk recommender:

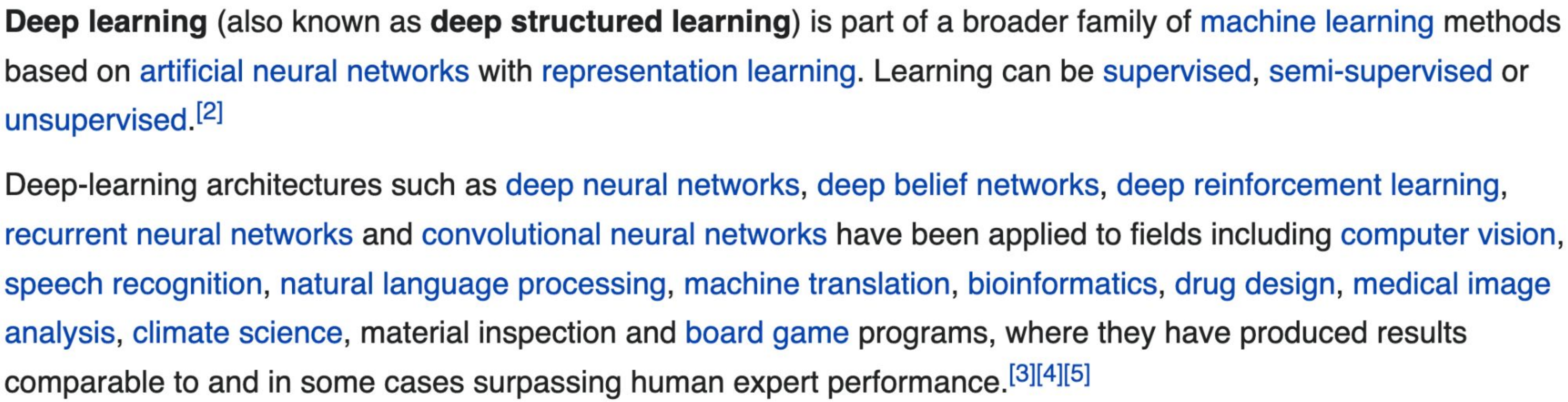
https://events.faculty.upenn.edu/

What did we hear?

A good understanding of text allows us to predict text

A good model for text will probably allow us good embeddings

Example text



Deep learning (also known as **deep structured learning**) is part of a broader family of machine learning methods based on artificial neural networks with representation learning. Learning can be supervised, semi-supervised or unsupervised.[2]

Deep-learning architectures such as deep neural networks, deep belief networks, deep reinforcement learning, recurrent neural networks and convolutional neural networks have been applied to fields including computer vision, speech recognition, natural language processing, machine translation, bioinformatics, drug design, medical image analysis, climate science, material inspection and board game programs, where they have produced results comparable to and in some cases surpassing human expert performance.[3][4][5]

What did we hear

Transformers:

After all we want to train on very large datasets

Of text

Using self-supervised learning (predict some words on the others)

Self supervised learning

Predict one part of dataset (withheld words)

From other parts (shown words)

Fine tuning

Train on one data set

Improve or fine-tune on another

Datasets in that space

EBM-NLP: Clinical trials

SciERC: Computer science

SciScite: Citations

Microsoft Academic Graph: a lot of information on countless papers, defunct

Open Alex: Open replacement for MAG

What should we do with the data

You design the strategy

TA guide

Hint 1: A good model allows us to predict text. But what does good mean?

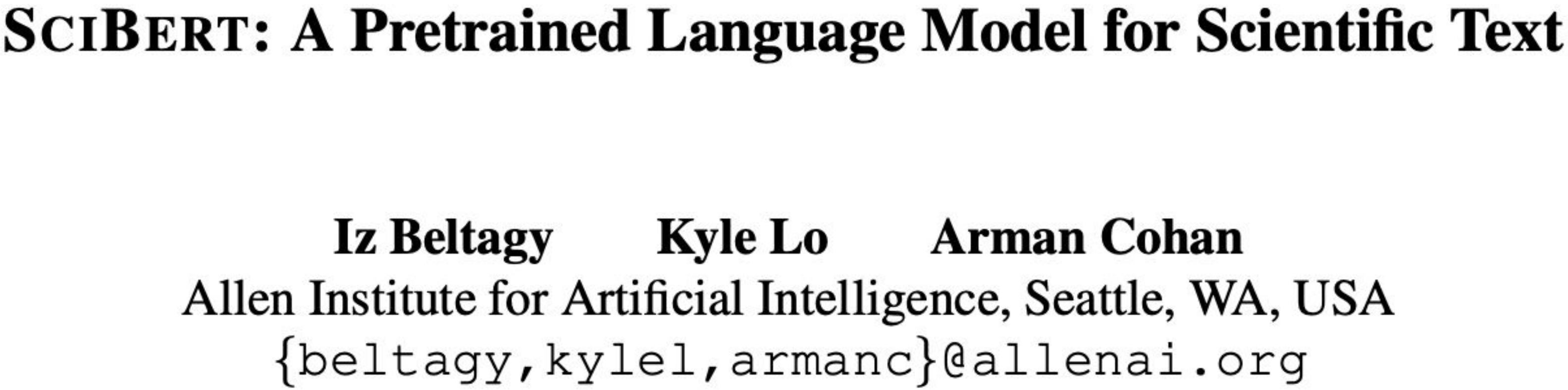
Hint 2: How can we quantify what a good model is?

Hint 3: If we had a good model, how would we then make paper recommendations?

Solution: A model that predicts the next word is, effectively, doing something quite similar to CCA - we want to maximize predictive value, or mutual information, with the next word. Once we do that, e.g. with a transformer we will have an embedding which we used for making the predictions. This embedding space may be quite good for making similar papers be meaningfully similar.

Advanced: How would you transfer the data from a general text model (say trained on wikipedia) to a more specialized model, say of pubmed papers?

A cool paper in that area



SciBert: A Pretrained Language Model for Scientific Text

Iz Beltagy **Kyle Lo** **Arman Cohan**
Allen Institute for Artificial Intelligence, Seattle, WA, USA
{beltagy,kylel,armanc}@allenai.org

Their strategy

Start with Pre-trained Bert

Then fine-tune with scientific data

Wrap up

Wrapping up DL thinking

This is not a slide. This is just Konrad and Lyle.

Why it matters? Good approaches = those that use the information. Mention No free lunch.

What can we do? Ask questions of experts and ourselves

What should we listen for? Everything