DATALOGISK INSTITUT
KØBENHAVNS UNIVERSITET

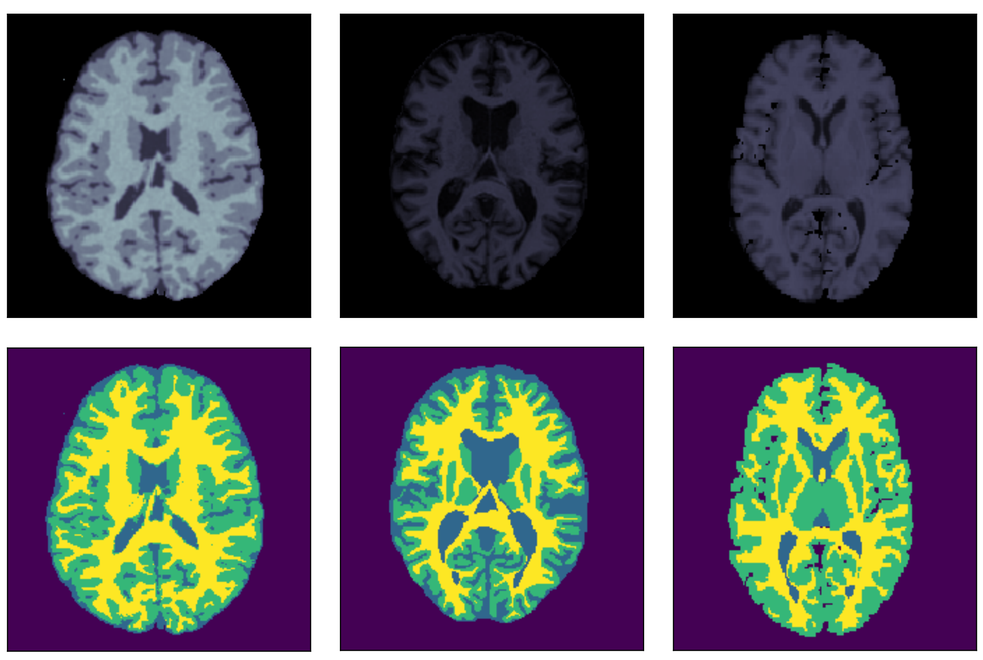
Sequential domain-adaptive machine learning

Wouter M Kouw

Machine learning algorithms have limited generalization capacity. One important concern is sampling bias; local sampling in space, such as at single universities or medical centers, will produce data that is not representative of larger populations. Local sampling in time, such as collecting data in one month of the year, will produce data that is not representative of larger time spans. *Domain adaptation* is concerned with generalizing from one biased sample to another and is used to make machine learning algorithms more robust to sampling bias. In this Fellowship, I studied domain adaptation from a sequential perspective.

Spatial adaptation of biomedical images

In biomedical imaging, machine learning algorithms are often used to perform tasks such as tissue segmentation and pathology detection. Their strength lies in the fact that they are very precise and pick up minute differences in image intensities. However, that same strength is a weakness in the presence of sampling bias. Minute differences in image intensities due to shifts in imaging protocols, can cause an algorithm to mistake an up-shifted gray matter tissue voxel with a white matter tissue voxel, for instance. Unfortunately, MRI scans vary across medical centers due to differences in acquisition and experimental protocols. This makes it difficult to train an algorithm on image data from one medical center and apply it to data from another. Below are examples of brain scans with tissue labels 'white matter' (yellow), 'gray matter' (green), 'cerebro-spinal fluid' (blue) and background (purple).



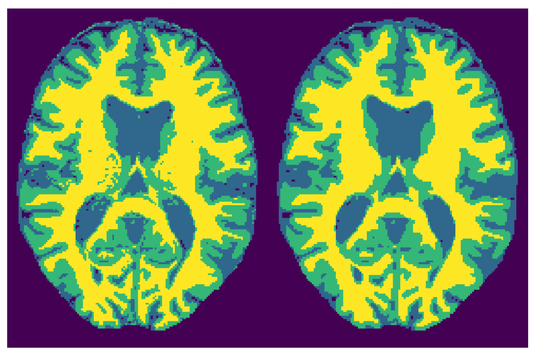
Utilizing prior knowledge from other medical centers

Even if imaging data from another medical center cannot directly be used to train an algorithm for deployment in another center, there is still information that can indirectly be used. For example, spatial smoothness or relative position of tissues can be learned from annotations produced at another center. In Bayesian statistical models, one forms 'prior' distributions to describe how probable particular values of variables are, before they observed. For example, it could be learned that gray matter tends to consist of "wiggly" structures. A tissue segmentation algorithm with such a prior belief will not produce outcomes with isolated gray matter voxels, because it knows that such a solution was improbable *a priori*.

Cross-center brain tissue segmentation

We fit a statistical model capturing spatial smoothness to tissue segmentations of brain scans in medical center A. It essentially counts how often each voxel has another voxel of the same tissue as its neighbour in the image. This statistical model is fed as a prior belief to another model that actually performs tissue segmentation in medical center B. The segmentation model assumes that each tissue generates a distinct mean intensity value in a voxel in the scan, and that the intensity spreads depending on the proportion of that tissue in the voxel.

Two example segmentations of the middle scan in the figure above are shown. The left image is made by the segmentation algorithm when it did not have a spatial smoothness prior. We then trained a prior on the segmentation image in the left part of the figure above. The right image is made by the segmentation algorithm when it did have a prior.



Article

WM Kouw, SN Ørting, J Petersen, KS Pedersen & M de Bruijne. A cross-center smoothness prior for variational Bayesian brain tissue segmentation. Information Processing in Medical Imaging, 2019.

Temporal adaptation of text representations

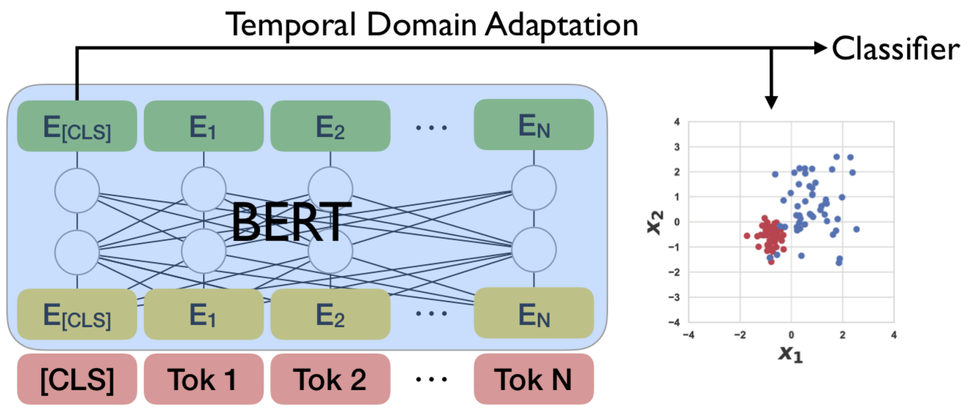
Language use drifts over time, in particular on online media. Say you're interested in automatically detecting the stance (support, deny, etc.) of a tweet author on a particular issue. You could collect a set of tweets and manually annotate them. If you train a machine learning algorithm on that set and then deploy it to monitor Twitter, it will probably perform well at first. But after a while, its performance will start to deteriorate; it looks for patterns in language that are no longer being used.

Representations of language

Modern computational models of language encode characters, words and sentences as vectors. Popular methods, such as artificial neural networks, move these vectors around in vector space such that semantic properties are maintained. For example, the words "Bert" and "Ernie" are mapped close together (two muppets on Sesame Street). But these networks have to be trained and are thus sensitive to sampling bias. If you base training on a snapshot data set, it will continue to operate in the same manner while language use evolves. In online media, language evolves so rapidly that re-collecting data and re-training is not a practical solution.

Sequential alignment of representations

We have looked at conceptually simple and computationally cheap ways of keeping up with language evolution. We now align word representations from one time-step to a next one, such that vectors with new semantic associations are mapped closer together. For example, "BERT" is also the name of a popular neural network architecture and will be moved to be closer to the word "neural-network".



Author stance recognition

Sequential alignment of word representations allows us to recognize the stance - support, deny, query or comment - of a Tweet author on online rumours (e.g. "fake news") through their use of language. Between rumours, the general tone and semantics change, but now we can align tweets to control for these subtle changes.

A tweet in January 2015 by a newspaper reporting on the "Charlie Hebdo" incident. It supports the veracity of the rumour by providing details and a reference.

> Support: France: 10 people dead after shooting at HQ of satirical weekly newspaper #CharlieHebdo, according to witnesses <URL>

If you look at tweets in March 2015, after the crash of Germanwings Flight 9525, then the neural network trained on the set of tweets from January 2015, will produce the following tweets as "neighbours" in embedding space:

> Comment: @USER Praying for the families and friends of those involved in crash. I'm so sorry for your loss.

> Query: @USER @USER if they had, its likely the descent rate wouldve been steeper and the speed not reduce, no ?

The above tweets do not support the veracity of the rumour. Now, we align the learned embedding of the neural network trained on tweets from Jan 2015 to the embeddings of tweets from Mar 2015. It produces the following two tweets as "neighbours" in embedding space:

> Support: Report: Co-Pilot Locked Out Of Cockpit Before Fatal Plane Crash <URL> #Germanwings <URL>

> Support: @USER: 148 passengers were on board #GermanWings Airbus A320 which has crashed in the southern French Alps <URL>

Article

J Bjerva, WM Kouw & I Augenstein. Back to the future - temporal adaptation of text representations. Empirical Methods in Natural Language Processing (submitted), 2019.

Discussion

Sequential domain adaptation is a solution to fitting statistical models in multi-site studies or fitting models based on snapshots of slowly-drifting processes. One advantage is that it becomes easier to combine data sets from different sources. Since many researchers benefit from utilizing other data sets, it increases the incentive to share data.

https://wmkouw.github.io | github.com/wmkouw