Regularizing Text Classification with Topics

Konstantinos Skianis, Michalis Vazirgiannis
LIX, École Polytechnique, France

Motivation

Text categorization is hard:
- high dimensionality
- prone to overfitting
- state-of-the-art structured regularization is slow due to overlapping clusters

Regularization is necessary:
- Critical for language modeling, structured prediction, and classification
- Prior on the feature weights
- Find the optimal weights: $\theta^* = \text{argmin}_\theta \underbrace{\underbrace{\sum_{i=1}^{N} \mathcal{L}(\mathcal{X}^i, \theta, y^i)}_{\text{empirical risk}} + \underbrace{\lambda\Omega(\theta)}_{\text{penalty term}}}_{\text{expected risk}}$

I. Structured Regularization

Group lasso: $\Omega(\theta) = \lambda \sum_g \lambda_g \|\theta_g\|_2$

Objective: $\Omega_{las}(\theta) + \Omega_{glas}(\mathrm{v}) + \mathcal{L}(\theta) + \mathrm{u}^\top(\mathrm{v} - M\theta) + \frac{\rho}{2}\|\mathrm{v} - M\theta\|_2^2$

Iterative update of θ, v and u:

$$\min_\theta \Omega_{las}(\theta) + \mathcal{L}(\theta) + \mathrm{u}^\top M\theta + \frac{\rho}{2}\|\mathrm{v} - M\theta\|_2^2$$

$$\min_\mathrm{v} \Omega_{glas}(\mathrm{v}) + \mathrm{u}^\top \mathrm{v} + \frac{\rho}{2}\|\mathrm{v} - M\theta\|_2^2$$

$$\mathrm{u} = \mathrm{u} + \rho(\mathrm{v} - M\theta)$$

Algorithm ADMM

Input: augmented Lagrangian variable ρ, λ_{glas} and λ_{las}

```
while update in weights not small do
  θ = argmin_θ Ω_las(θ) + L(θ) + ρ/2 Σ_{i=1}^{V} N_i(θ_i − μ_i)^2
  for g = 1 to G do
    v_g = prox_{Ω_glas, λ_g/ρ}(z_g)
  end for
  u = u + ρ(v − Mθ)
end while
```

II. Structured Regularization in NLP

STATISTICAL REGULARIZERS:
- Network of features
  - $\Omega_{net}(\theta) = \lambda_{net} \sum \theta_k^\top M \theta_k$, where $M = \alpha(I - P)^\top(I - P) + \beta I$.
- Sentence Regularizer
  - $\Omega_{sen}(\theta) = \sum_{d=1}^{D} \sum_{s=1}^{S_d} \lambda_{d,s} \|\theta_{d,s}\|_2$

SEMANTIC REGULARIZERS:
- LDA regularizer
- **LSI regularizer**
  - $\Omega_{LDA,LSI}(\theta) = \sum_{k=1}^{K} \lambda \|\theta_k\|_2$

GRAPHICAL REGULARIZERS:
- **Graph-of-words regularizer**
  - Community detection on document collection graph
  - $\Omega_{gow}(\theta) = \sum_{c=1}^{C} \lambda \|\theta_c\|_2$
  - c ranges over the C communities.
- **Word2vec regularizer**
  - Kmeans clustering on word2vec
  - $\Omega_{word2vec}(\theta) = \sum_{k=1}^{K} \lambda \|\theta_k\|_2$
  - K is the number of clusters

$\hookrightarrow$ **Why?** Clusters of words will capture same concepts & topics

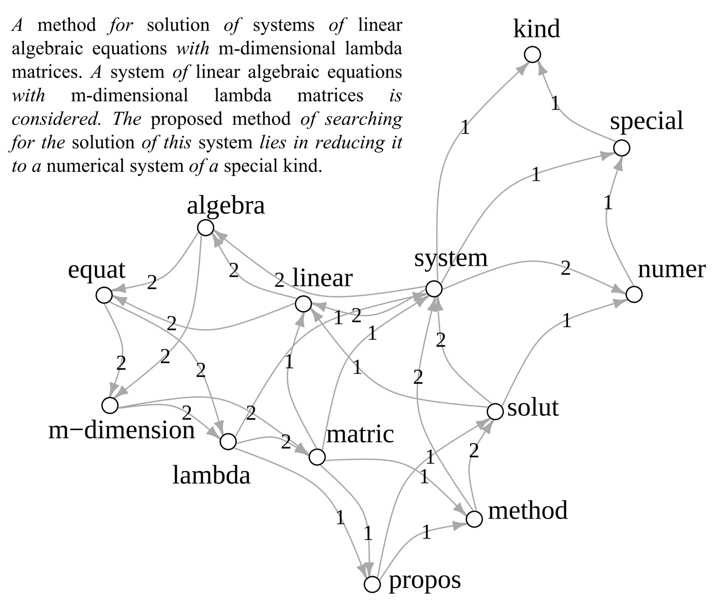


Figure: A Graph-of-words example.

III. Results

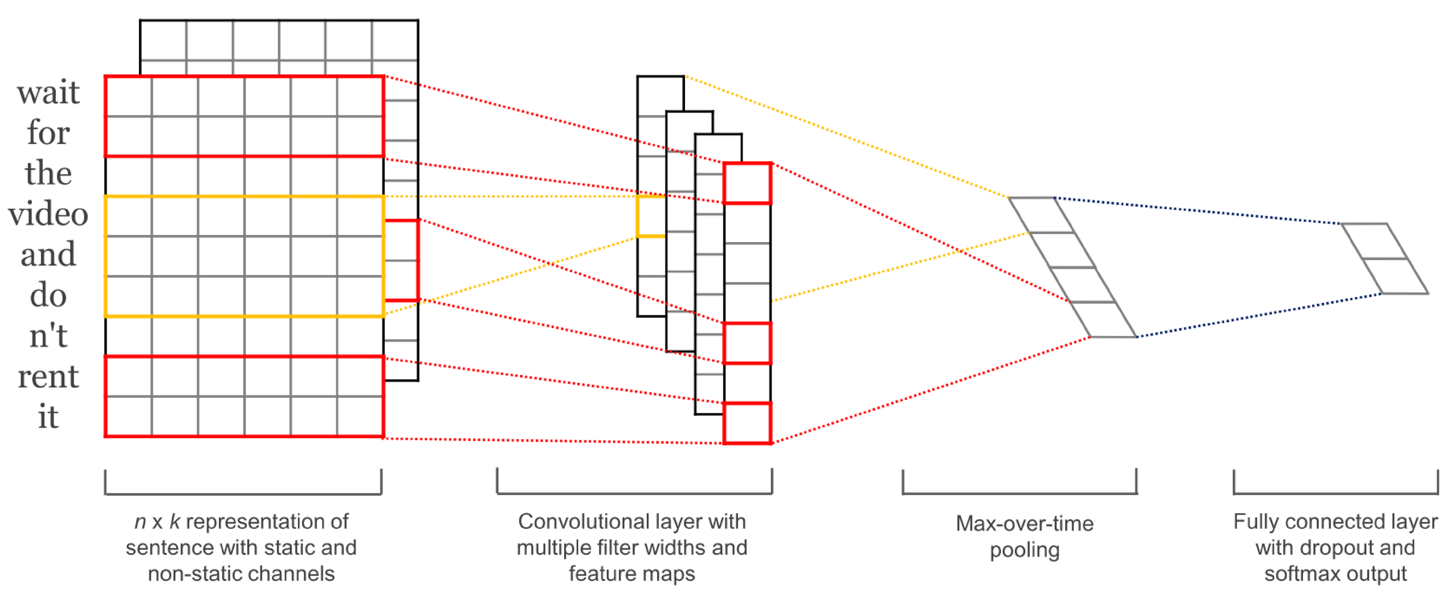
Table: Accuracy

	dataset	no reg.	lasso	ridge	elastic	group lasso LDA	LSI	sentence	GoW	word2vec
20NG	science	0.946	0.916	0.954	0.954	**0.968**	**0.968***	0.942	0.967*	**0.968***
	sports	0.908	0.907	0.925	0.920	0.959	0.964*	**0.966**	0.959*	0.946*
	religion	0.894	0.876	0.895	0.890	0.918	0.907*	**0.934**	0.911*	0.916*
	computer	0.846	0.843	0.869	0.856	0.891	0.885*	0.904	0.885*	**0.911***
Sentiment	vote	0.606	0.643	0.616	0.622	**0.658**	0.653	0.656	0.640	0.651
	movie	0.865	0.860	0.870	0.875	**0.900**	0.895	0.895	0.895	0.890
	books	0.750	0.770	0.760	0.780	0.790	0.795	0.785	0.790	**0.800**
	dvd	0.765	0.735	0.770	0.760	0.800	**0.805***	0.785	0.795*	0.795*
	electr.	0.790	0.800	0.800	**0.825**	0.800	0.815	0.805	0.820	0.815
	kitch.	0.760	0.800	0.775	0.800	0.845	**0.860***	0.855	0.840	0.855*

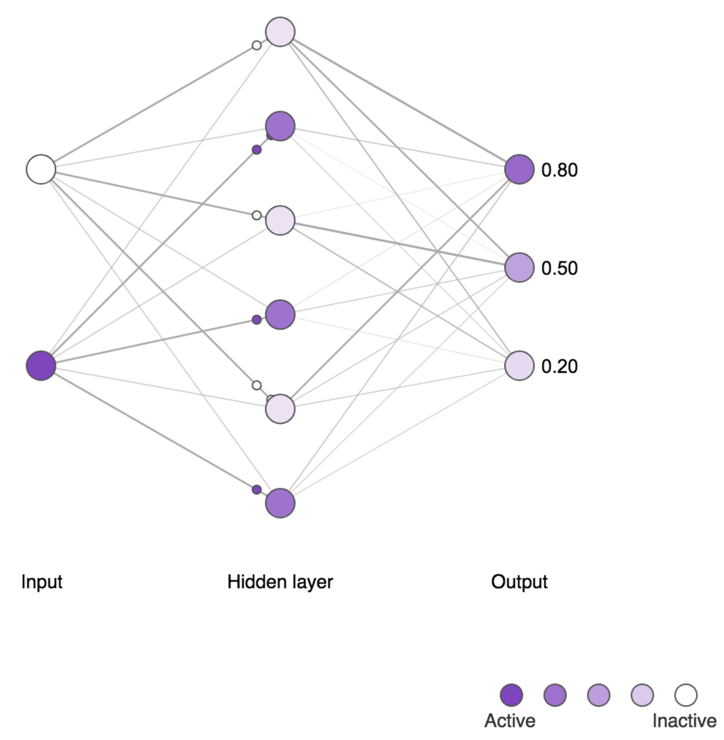
Table: Model size (∼sparsity)

	dataset	no reg.	lasso	ridge	elastic	group lasso LDA	LSI	sentence	GoW	word2vec
20NG	science	100	1	100	63	19	20	86	19	21
	sports	100	1	100	5	60	11	6.4	55	44
	religion	100	1	100	3	94	31	99	10	85
	computer	100	2	100	7	40	35	77	38	18
Sentiment	vote	100	1	100	8	15	16	13	97	13
	movie	100	1	100	59	72	81	55	90	62
	books	100	3	100	14	41	74	72	90	99
	dvd	100	2	100	28	64	8	8	58	64
	electr.	100	4	100	6	10	8	43	8	9
	kitch.	100	5	100	79	73	44	27	75	46

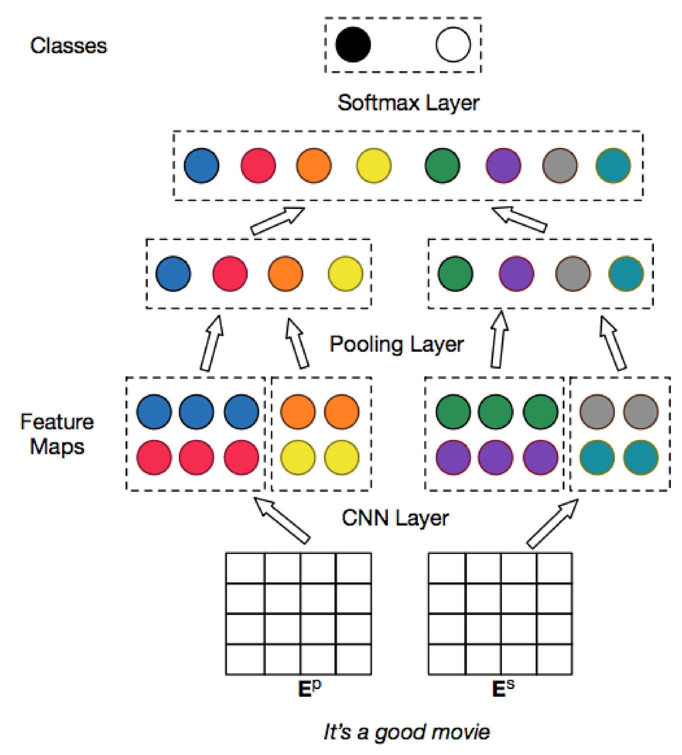
IV. Regularizing Deep Learning



Sentence Classification with CNNs (Kim et al., EMNLP 2014)



Neuron Deletion (Morcos et al., ICLR 2018)



Grouped Weight Sharing (Zhang et al., ACL 2017)

V. Discussion & Future Work

- Superior proposed regularizers: more effective, more efficient and sparser
- GoW-based regularization although very fast, did not outperform the other methods
  - Overlapping community detection algorithms failed to identify "good" groups
- Incorporating prior knowledge into neural models
- Interpretable neurons are no more important
- Networks which **generalise better are harder to break**

	dataset	GoW	word2vec
20NG	science	79	691
	sports	137	630
	religion	35	639
	computer	95	594

Table: Number of groups.

	dataset	lasso	ridge	elastic	group lasso LDA	LSI	sentence	GoW	word2vec
20NG	science	10	1.6	1.6	15	11	76	12	19
	sports	12	3	3	7	20	67	5	9
	religion	12	3	7	10	4	248	6	20
	computer	7	1.4	0.8	8	6	43	5	10

Table: Time (in seconds) for learning with best hyperparameters.

CONCLUSION
- Find and extract semantic and syntactic structures that lead to sparser feature spaces $\rightarrow$ faster learning times
- Linguistic prior knowledge in the data can be used to improve categorization performance for baseline bag-of-words models, by mining inherent structures
- No significant change in results with different loss functions as the proposed regularizers are not log loss specific

FUTURE WORK
- Find better clusters in word2vec (+overlapping with GMM)
- Explore alternative regularization algorithms diverging from group-lasso
- Interpret neurons in a deeper level
- Different tasks like qa, chatbots etc.
- Study sparsity along with groups
- Learning the topics simultaneously

kskianis@lix.polytechnique.fr