Zero-shot learning approach to the problem of short text classification

N. Nikolaev, I. Bondarenko, T. Batura

Novosibirsk State University

18.05.2021

Outline

1. A brief description of zero-shot learning approach
2. Overview of the proposed data and solution
   - DBpedia: ontology classification dataset
   - VAE based approach for 0-shot text classification
   - Variational autoencoder loss
   - Cross-alignment loss
   - Distribution-Alignment Loss
   - Cross- and Distribution Alignment Loss
3. Results
4. ZeroShotEval: Unified Pipeline for ZSL Models Evaluation
5. Conclusion

A brief description of zero-shot learning approach

Classification problem

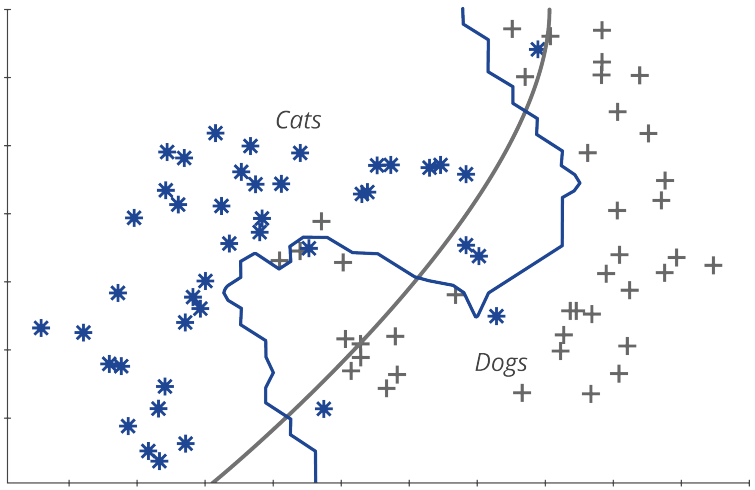

Figure 1: Problem Setting for Classification

A classification problem is when the output variable is a category, such as "cats" or "dogs" or "disease" and "no disease". A classification model attempts to draw some conclusion from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes.

What is few-shot learning?

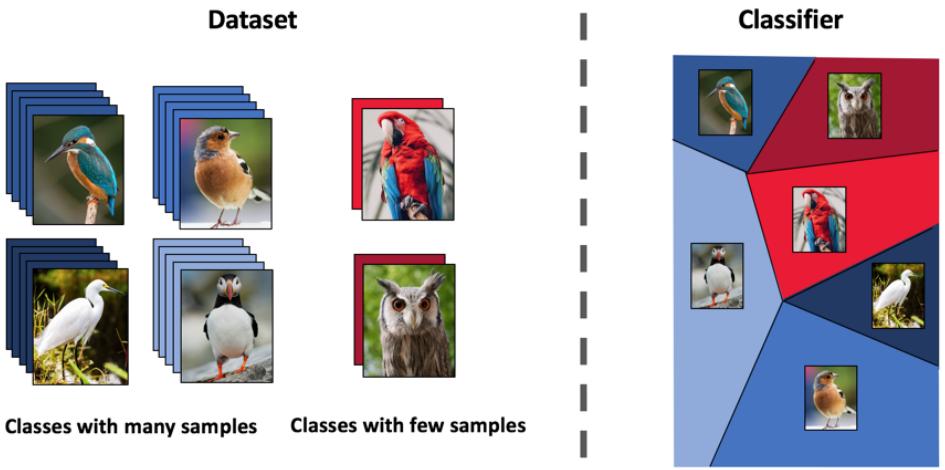

Figure 2: Problem Setting for Few-Shot Learning

As the name implies, few-shot learning refers to the practice of feeding a learning model with a very small amount of training data, contrary to the normal practice of using a large amount of data.

What is zero-shot learning?

Zero-shot learning approach aims to solve a task without receiving any example of that task at training phase, e.g. the task of recognizing an object from a given image where there were not any example images of that object during training phase can be considered as a zero-shot learning task. Briefly, it simply allows us to recognize objects we have not seen before.

$$S = \{(x, y, c(y))|\ x \in X,\ y \in Y^S,\ c(y) \in C\}$$
$x -$ image-features, $y -$ labels

$$U = \{(u, c(u))|\ u \in Y^u,\ c(u) \in C\}$$
$C(U) = \{c(u_1), \ldots, c(u_L)\} -$ class-embeddings of unseen classes

$$f_{ZSL} : X \rightarrow Y^U$$
$$f_{GZSL} : X \rightarrow Y^U \cup Y^S$$

Application of additional data modalities

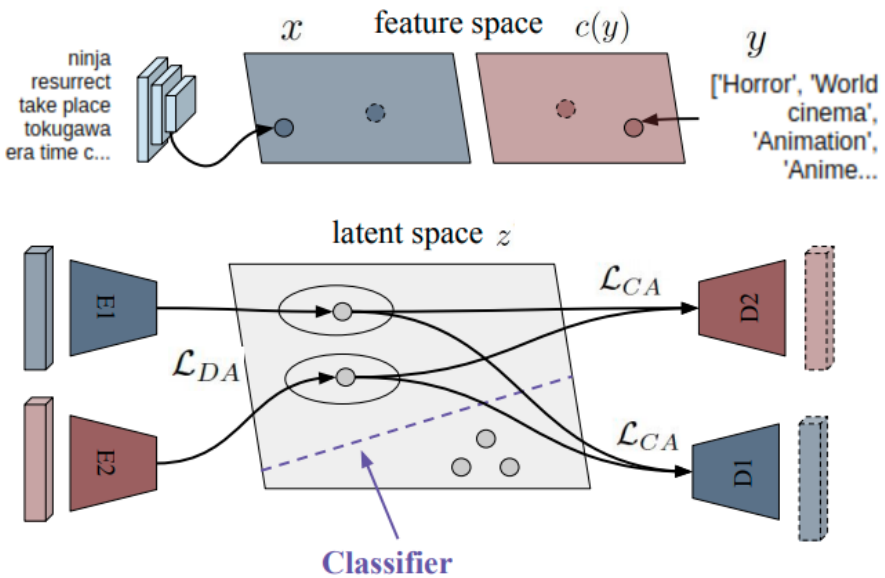

Figure 3: Diagram of the approach based on VAE.

The model learns a latent embedding (z) of different modalities via aligned VAEs optimized with cross-alignment ($\mathcal{L}_{CA}$) and distribution alignment ($\mathcal{L}_{DA}$) objectives, and subsequently trains a classifier on sampled latent features of seen and unseen classes.

Overview of the proposed data and solution

DBpedia: ontology classification dataset


```
{
  "label" : [
    0 : "Company"
    1 : "EducationalInstitution"
    2 : "Artist"
    3 : "Athlete"
    4 : "OfficeHolder"
    5 : "MeanOfTransportation"
    6 : "Building"
    7 : "NaturalPlace"
    8 : "Village"
    9 : "Animal"
    10 : "Plant"
    11 : "Album"
    12 : "Film"
    13 : "WrittenWork"
  ]
  "title" : "string"
  "content" : "string"
}
```


Figure 4: 14 non-overlapping classes from DBpedia 2014

	content	label
0	Abbott of Farnham E D Abbott Limited was a British coachbuilding business based in Farnham Surrey trading under that name from 1929. A major part of their output was under sub-contract to motor vehicle manufacturers. Their business closed in 1972.	0
1	Schwan-STABILO is a German maker of pens for writing colouring and cosmetics as well as markers and highlighters for office use. It is the world's largest manufacturer of highlighter pens Stabilo Boss.	0
2	Q-workshop is a Polish company located in Poznań that specializes in designand production of polyhedral dice and dice accessories for use in various games (role-playing gamesboard games and tabletop wargames). They also run an online retail store and maintainan active forum community.Q-workshop was established in 2001 by Patryk Strzelewicz – a student from Poznań. Initiallythe company sold its products via online auction services but in 2005 a website and online store wereestablished.	0
3	Marvell Software Solutions Israel known as RADLAN Computer Communications Limited before 2007 is a wholly owned subsidiary of Marvell Technology Group that specializes in local area network (LAN) technologies.	0

Figure 5: Total size of the training dataset is 560,000 and the testing dataset is 70,000.

DBpedia: additional data modality

As an additional modality we can use class descriptions that represent a general knowledge about class instances.

0	A company is a legal entity representing an as...
1	An educational institution is a place where pe...
2	An artist is a person engaged in an activity r...
3	An athlete (also sportsman or sportswoman) is ...
4	A person who's been appointed to a position by...
5	Any of the different kinds of transport facili...
6	Building is a structure with a roof and walls ...
7	Natural place means an area of the rural or no...
8	Village is a group of houses and associated bu...
9	Animal is a living organism that feeds on orga...
10	Plant is a living organism of the kind exempli...
11	An album is a collection of audio recordings i...
12	A film, also called a movie, motion picture or...
13	Literature broadly is any collection of writte...

Figure 6: Classes definitions taken from Wikipedia as an additional modality for training a zero-shot net.

VAE based approach for 0-shot text classification

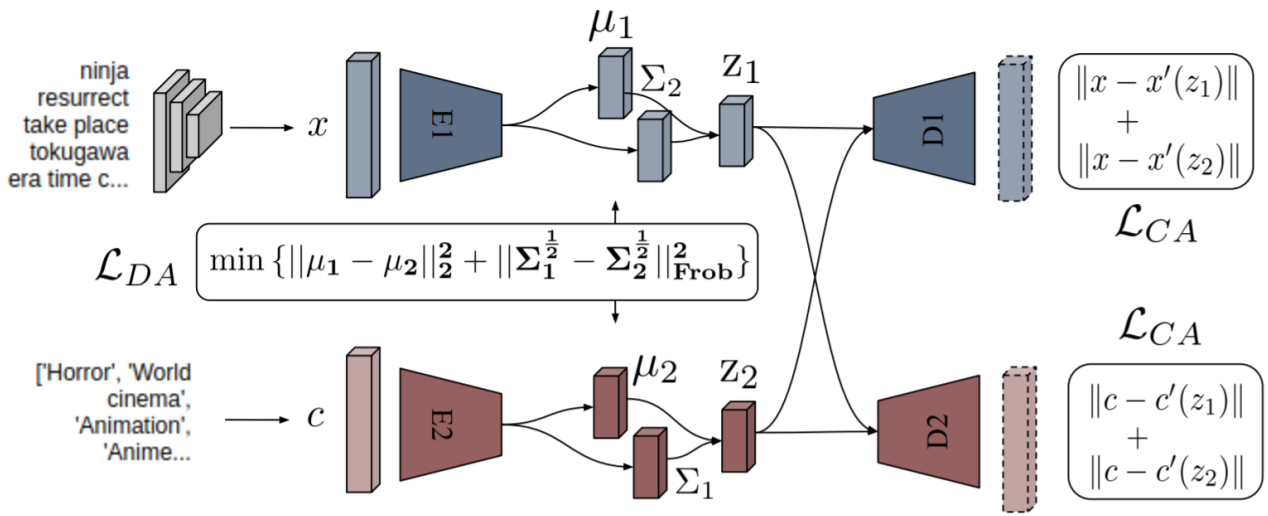

Figure 7: Architecture based on VAE with 2 data modalities. The main part of the architecture is 2 autoencoders for each modality. At the same time, each autoencoder receives contribution from the cross-alignment (CA) loss function and from the distribution-alignment (DA) loss function.

Variational autoencoder loss

The basic VAE loss of our model is the sum of M VAE-losses:

$$\mathcal{L}_{VAE} = \sum_{i}^{M} \mathbb{E}_{q_\phi(x|z)}[\log p_\theta(x^{(i)}|z)] - \beta D_{KL}(q_\phi(z|x^{(i)})||p_\theta(z)) \quad (1)$$

where β weights the KL-Divergence.

Cross-alignment loss

Each modality specific decoder is trained on the latent vectors derived from the other modalities. This cross-reconstruction loss is:

$$\mathcal{L}_{CA} = \sum_{i}^{M} \sum_{j \neq i}^{M} |x^{(j)} - D_j(E_i(x^{(i)}))| \tag{2}$$

where E_i is the encoder of a feature of i^{th} modality and D_j is the decoder of a feature of the same class but the j^{th} modality.

Distribution-Alignment Loss

The 2-Wasserstein distance between two distributions i and j is given as:

$$W_{ij} = [\,\|\mu_i - \mu_j\|_2^2 + Tr(\Sigma_i) + Tr(\Sigma_j) - 2(\Sigma_i^{\frac{1}{2}}\Sigma_i\Sigma_j^{\frac{1}{2}})^{\frac{1}{2}}]^{\frac{1}{2}} \tag{3}$$

Since the encoder predicts diagonal covariance matrices, which are commutative, this distance simplifies to:

$$W_{ij} = (\|\mu_i - \mu_j\|_2^2 + \|\Sigma_i^{\frac{1}{2}} - \Sigma_j^{\frac{1}{2}}\|_{Frobenius}^2)^{\frac{1}{2}} \tag{4}$$

$$\mathcal{L}_{DA} = \sum_{i}^{M}\sum_{j\neq i}^{M} W_{ij} \tag{5}$$

Cross- and Distribution Alignment Loss

The cross- and distribution aligned VAE combines the basic VAE-loss with L_{CA} and L_{DA}:

$$\mathcal{L} = \mathcal{L}_{VAE} + \gamma \mathcal{L}_{CA} + \delta \mathcal{L}_{DA} \tag{6}$$

where γ and δ are the weighting factors of the cross alignment and the distribution alignment loss, respectively.

Results

Results

	XLNet	**BERT**	**ULMFiT**	**Ours**
Feature Size	1024	1024	400	1024
Seen	**0.62**	0.68	0.80	1.00
Unseen	-	-	-	**1.01**
Harmonic	-	-	-	**1.00**

Table 1: Comparison with state-of-the-art error rates on test sets of several text classification dataset. It is important to note, that only our approach allows to measure an error rate on unseen classes as the model were trained in zero-shot setting. All the experiments were performed on DBpedia dataset with predefined splits into train/test. We also report a harmonic mean of errors on seen and unseen classes.

ZeroShotEval: Unified Pipeline for ZSL Models Evaluation

ZeroShotEval: Unified Pipeline for ZSL Models Evaluation [1]

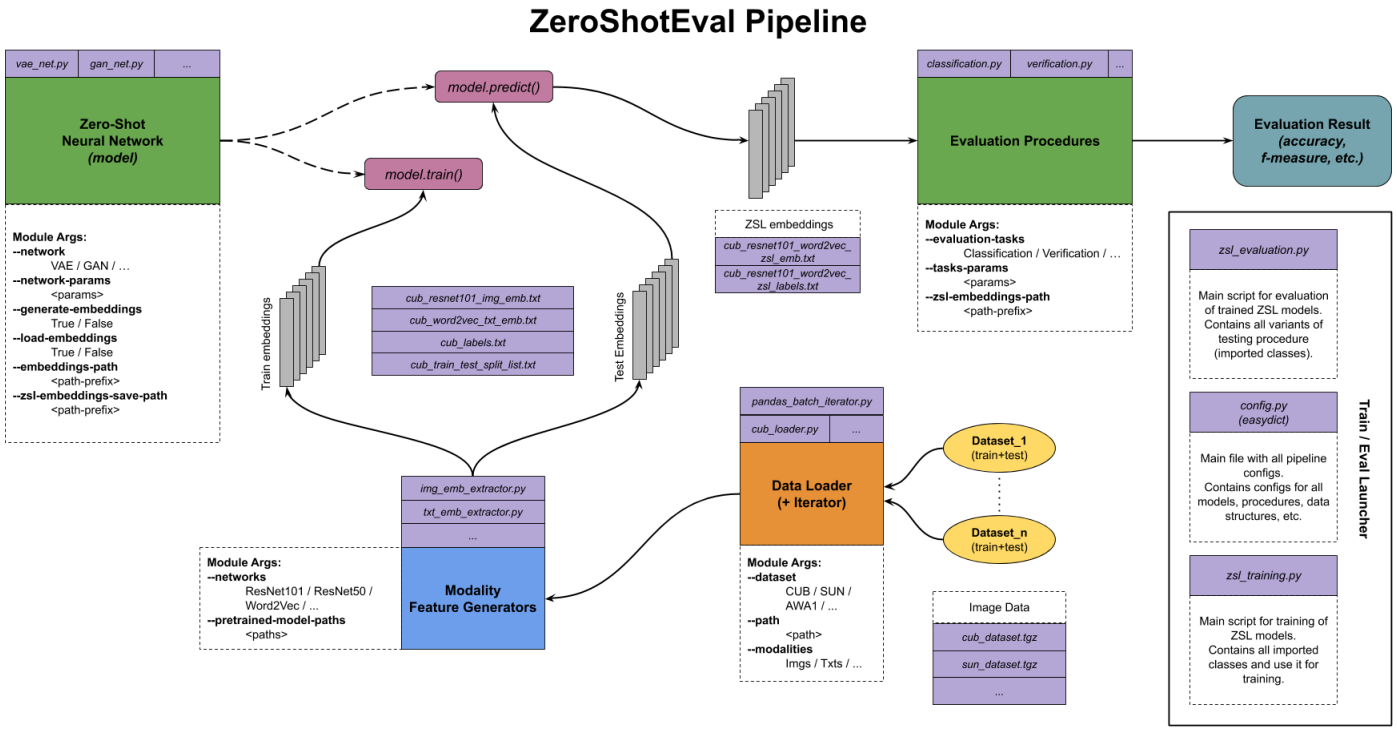

Figure 8: The overview of the proposed framework with four main parts-phases: 1) data loading and preprocessing, 2) modality feature generators, 3) zero-shot neural network, 4) evaluation procedures.

[1] GitHub Repository: https://github.com/ZSLresearch-team/ZeroShotEval

Conclusion

Conclusion

- 0-Shot TC is a scantily explored field and task which has a great potential for further improvements.
- All existing solutions are based on extremely different approaches and datasets, so they require a single form of quality assessment - a way to ZeroShotEval.
- It is necessary to adapt existing datasets or develop a new one to solve the problem. The lack of data is one of the main difficulties in solving this problem.
- Even with usual classification datasets zero-shot approach allows to achieve comparable performance with usual classification models, but with a greater prospective of application.

References

- Y. Xian, B. Schiele and Z. Akata. Zero-Shot Learning - The Good, the Bad and the Ugly. CVPR 2017
- W. Yin, J. Hay and D. Roth. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. EMNLP 2019.
- Y. Xian, T. Lorenz, B. Schiele, and Z. Akata. Feature generating networks for zero-shot learning. CVPR, 2018.
- E. Schonfeld, S. Ebrahimi, S. Sinha, T. Darrell and Z. Akata. Generalized Zero- and Few-Shot Learning via Aligned Variational Autoencoders. CVPR, 2019.
- P. K. Pushp, M. M. Srivastava. Train Once, Test Anywhere: Zero-Shot Learning for Text Classification. ICLR 2018.
- J. Zhang, P. Lertvittayakumjorn, Y. Guo. Integrating Semantic Knowledge to Tackle Zero-shot Text Classification. NAACL-HLT 2019.
- P. Qin, X. Wang, W. Chen, C. Zhang, W. Xu, W. Y. Wang. Generative Adversarial Zero-Shot Relational Learning for Knowledge Graphs
- D. Bamman, B. O'Connor, and N. Smith. CMU Movie Summary Corpus. Language Technologies Institute and Machine Learning Department at Carnegie Mellon University

Publications

- Nikolaev N.A. Zero-shot learning approach to the problem of short text classification. International Scientific Student Conference. Novosibirsk. April, 2021. (In English, RSCI)

Thank You for Your Attention!