Knowledge Base Representation Learning – Baselines and Challenges

Timothée Lacroix, Nicolas Usunier, Guillaume Obozinski – Facebook AI Research / Ecole des Ponts

1. Knowledge Base Representation Learning

- A **Knowledge-Base** is a set of triples defining relations between entities.
- Examples :
  - `FB15K` : (Wall-E, genre, Fantasy), (Lil Wayne, born in, New Orleans)
  - `Wordnet`: (score, hypernym, evaluation)
  - `SVO`: (cat, eat, food)
- Given an incomplete triple, the goal is to provide a good ranking with a score $s(\mathbf{u}, \mathbf{r}, \mathbf{v})$, function of the *embeddings* $(\mathbf{u}, \mathbf{r}, \mathbf{v})$.

2. Previous Models

Model	Score Function	Regularization
RESCAL [1]	$e_s^T W_r e_o$	$\lambda_W \|W\|_F^2 + \lambda_E \|E\|_F^2$
TransE [2]	$-\|e_s + w_r - e_o\|_2^2$	$\|e_i\|_2 \leq 1$
DistMult [3]	$\langle w_r, e_s, e_o \rangle$	$\lambda \|W\|_2^2$ and $\|e_i\|_2 \leq 1$
ComplEx [4]	$Re\left(\langle w_r, e_s, \overline{e_o} \rangle\right)$	$\lambda\left(\|W\|_F^2 + \|E\|_F^2\right)$

Previous works focused on the structure of the model (same embeddings for `lhs` and `rhs` embeddings, use of complex numbers, translations, ...) and used the rank and early-stopping to regularize. Different losses have been proposed, in this work, we used :

- Ranking loss :

$$\ell(s, \mathbf{u}, \mathbf{r}, \mathbf{v}, \overline{\mathbf{v}}) = \ln(1 + \exp(s(\mathbf{u}, \mathbf{r}, \mathbf{v}) - s(\mathbf{u}, \mathbf{r}, \overline{\mathbf{v}})))$$

- Binary classification loss (number of negatives is a hyperparameter):

$$\ell(s, Y, \mathbf{u}, \mathbf{r}, \mathbf{v}) = \ln(1 + \exp(-Y \cdot s(\mathbf{u}, \mathbf{r}, \mathbf{v})))$$

3. Matrix regularizers and tensors

- Trace-norm is easy to control thanks to :

$$\|X\|_\Sigma = \sum_i |\lambda_i| = \min_{X = UV'} \|U\|_2 \|V\|_2 = \min_{X = UV'} \frac{1}{2}(\|U\|_2^2 + \|V\|_2^2)$$

- Tensor trace-norm defined for the Tucker decomposition, hard to control.
- Max-norm :

$$\|X\|_{max} = \min_{X = UV'} \left(\max_i \|U_i\|_2\right)\left(\max_i \|V_i\|_2\right)$$

- Tensor extension to max-norm is not a norm, we try to be in the hull that controls the complexity instead.

4. Atomic Norm Regularization

- We use the canonical tensor decomposition (Hitchcock, 1927) with explicit factor weights σ to ease capacity control:

$$X = \sum_{k=1}^{K} \sigma_k \mathbf{e}^\ell_{:,k} \otimes \mathbf{w}_{:,k} \otimes \mathbf{e}^\mathbf{r}_{:,k} \qquad \text{score} : \langle \sigma, e_o^\ell, w_r, e_s^r \rangle = X_{o,r,s}$$

- We bound the complexity of our function class by setting :

$$(\tilde{\sigma})_k = \sigma_k \|\mathbf{e}^\ell_{:,k} \otimes \mathbf{w}_{:,k} \otimes \mathbf{e}^\mathbf{r}_{:,k}\|_\infty \in \mathcal{B} \quad \|\mathbf{e}^\ell\|_\infty, \|\mathbf{w}\|_\infty, \|\mathbf{e}^\mathbf{r}\|_\infty \leq 1$$

- $\mathcal{B} = \mathcal{B}_1 = \{x \mid \|x\|_1 \leq c_1\}$, controls the Atomic Norm with atoms :

$$\{u \otimes v \otimes w \mid u, v, w \in \{-1, 1\}^{n \times m \times p}\}$$

  Similar to Max-norm in complexity [5]. Doesn't yield good reslts on `SVO`.
- $\mathcal{B} = \mathcal{B}_\infty = \{x \mid \|x\|_\infty \leq c_\infty\}$, the generalization bound depends on the rank, yet we observe good behaviour on `SVO`.
- Finally, $\mathcal{B} = \mathcal{B}_1 \cap \mathcal{B}_\infty$ leads to sparsity and good behaviour on `SVO`.
- We use mini-batch projected `Adagrad` on σ, `Hogwild` and projected `Adagrad` for the factors. [6]

5. Results

Model	KIN (AUC)	WN (MRR)	FB15K (MRR)
RESCAL	0.948 ± 8.10^{-3}	0.89	0.35
DistMult	-	0.89	0.80
ComplEx	0.98 ± 3.10^{-3}	0.94	0.80
Reg. Candecomp	0.951 ± 3.10^{-3}	0.94	0.80

7. Conclusions and Future work

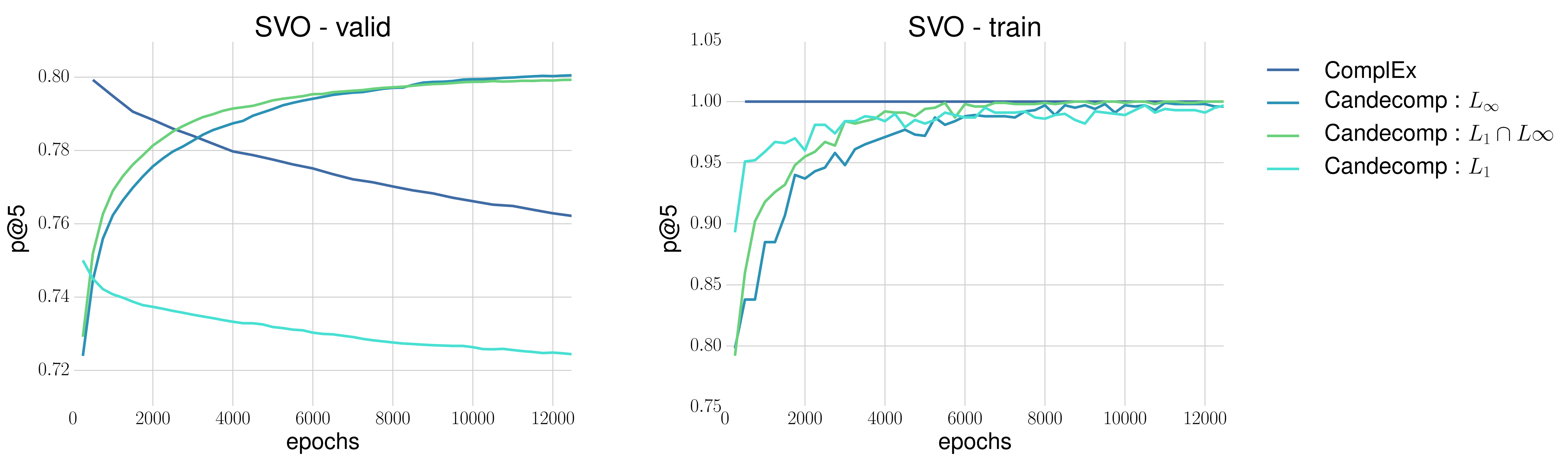
Conclusion

- No structural tricks needed to get state of the art results on current benchmarks.
- Currently no regularization needed on `FB15K` and `WN`.

Future work

- Understand what makes the L_∞ generalize better than L_1.
- Explain the slow convergence of regularized methods.
- Explore the differences between the possible losses.
- Try the same regularizers on `ComplEx`.

6. Figures



. References

[1] Nickel, M., Tresp, V., & Kriegel, H. P. (2011). A three-way model for collective learning on multi-relational data. In Proceedings of the 28th international conference on machine learning (ICML-11)

[2] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In Advances in neural information processing systems (pp. 2787-2795)

[3] Yang, B., Yih, W. T., He, X., Gao, J., & Deng, L. (2014). Embedding entities and relations for learning and inference in knowledge bases. In International Conference on Learning Representations (ICLR-15)

[4] Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., & Bouchard, G. (2016, June). Complex embeddings for simple link prediction. In International Conference on Machine Learning (pp. 2071-2080)

[5] Srebro, N., & Shraibman, A. (2005, June). Rank, Trace-Norm and Max-Norm. In COLT (Vol. 5, pp. 545-560)

[6] Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research, 12(Jul), 2121-2159.