Coherent and Consistent Relational Transfer Learning with Auto-encoders

Harald Strömfelt, Imperial College London
Luke Dickens, University College London
Artur Garcez, City, University of London
Alessandra Russo, Imperial College London

Imperial College London

Motivation

Human defined concepts are inherently transferable. After learning them in one domain/context, they can be applied in related scenarios [1].

Under what conditions can we say the same about neural-symbolic learners?

Partial Relation Transfer (PRT)

Relation Learning Phase (RLP): Learn set of relations in source (s) domain and task.
Transfer Phase (TP): Apply (subset) of the relations to different, but related, target (t) domain and task, with fixed parameters (no relation learning). Included relations act to guide a domain-specific auto-encoder. Held-out relations rely upon zero-shot transfer.
Relations: R = {isGreater, isLess, isEqual, isSuccessor, isPredecessor}.

Architecture

- Combines domain-specific β-VAE [2] with cross-domain φ_r, $r \in R$, relation-decoders [3].

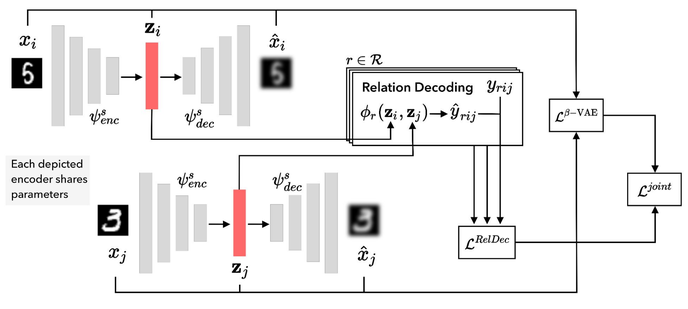

Figure 1. Model depiction. We apply a domain-specific auto-encoder for each domain, where $\varphi^s_{enc/dec}$ is the auto-encoder for the source domain, X_s. Encodings, $(\mathbf{z}_i, \mathbf{z}_j)$, of different domain images, $(\mathbf{x}_i, \mathbf{x}_j)$, are used as input for relation-decoders, φ_r, $r \in R$. In source domain gradients flow from L_{KGE}, but in the target domain they do not.

Joint objective:

$$\mathscr{L}^{joint} = \mathscr{L}^{ELBO}_{\beta\text{-VAE}} - \lambda \underbrace{\mathbb{E}_{r,y_{rij},\mathbf{z}_i,\mathbf{z}_j}[y_{rij}\log(\hat{y}_{rij}) + (1-y_{rij})\log(1-\hat{y}_{rij})]}_{\mathscr{L}^{RelDec}},$$

Dynamic Comparator

Proposed binary relation-decoder with two modes:

$$\phi^{DC}_r(\boldsymbol{z}_i, \boldsymbol{z}_j) = a_0 \cdot \underbrace{\sigma_0\big(\eta_0(\|\boldsymbol{u} \odot (\boldsymbol{z}_i - \boldsymbol{z}_j + \boldsymbol{b}_\dagger)\|^2_2)\big)}_{\phi^\dagger_r} + a_1 \cdot \underbrace{\sigma_1\big((\eta_1 \cdot \boldsymbol{u}^\top(\boldsymbol{z}_i - \boldsymbol{z}_j + \boldsymbol{b}_\ddagger))\big)}_{\phi^\ddagger_r}$$

- $\varphi_r^\dagger$, similarity by distance function; $\varphi_r^\ddagger$, step-like function.
- a_0 and a_1 perform attention-style weighting to select mode, $\mathbf{b}_{\dagger/\ddagger}$ gives offset capability and $\mathbf{u}$ acts as a directional mask. σ_0 and σ_1 are an exponential and a sigmoid function. η_0 and η_1 are any-value and positive-value scalars, respectively.

Results

Accuracy profiles

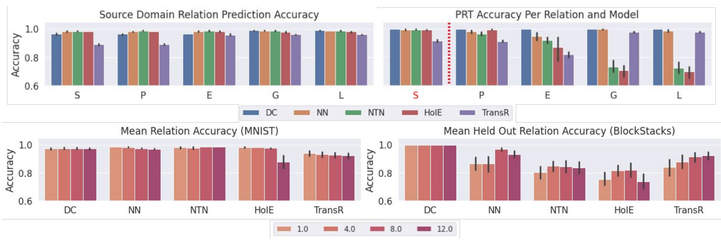

Figure 2. Top: accuracy of each relation-decoder for each model choice on source (**left**) and target (**right**) domains. Red relation labels indicate which relations were included as a guide in the TP phase. **Bottom:** β effect on mean accuracy of all relation-decoders (**left**) and only held-out relation-decoders (**right**).

Compared test relation-prediction accuracy (**Fig. 2**) between a variety of relation-decoder models: NTN, HolE, TransR from literature, DC proposed, NN simple 3-layer neural-network reference model.
Varied β ∈ {1, 4, 8, 12} during RLP but fixed in TP. λ fixed in both. Only isSuccessor (S) is included in TP, all others are held-out.

Key observations:

- DC retains accuracy to held-out relations. Other models see degradation, though NN does surprisingly well.
- β increase has positive effect on accuracy at moderate level.

Consistency and Gradient Conformity

Next, (**Fig. 3-top**) we looked at logical consistency across (**Con-A**) collective truth-assignments of each relation-decoder for different input pairs, sampled from three latent subspaces: **data-embeddings**, encodings of domain's test data; **interpolation**, samples from a corresponding Gaussian distribution using empirical mean/variance of data-embeddings; and **extrapolation**, samples from regions strictly outside the data-embeddings region.

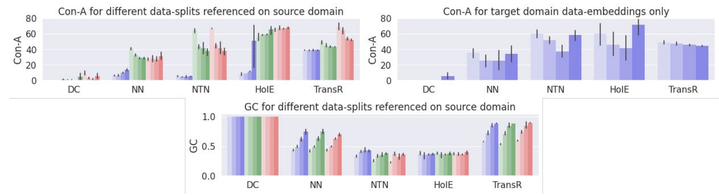

Figure 3. Top-Left: Con-A comparison on different source domain data splits. **Top-Right:** Con-A on target domain data-embeddings (lower Con-A values are better). **Bottom:** GC values for different source domain data splits, where larger values indicate greater gradient-conformity between relation-decoders.

Additionally, to gain insight into the mechanism by which β affects PRT performance, we looked at Gradient-Conformity (GC) (**Fig. 3-bottom**), which evaluates the 'parallel-ness' of relation-decoder gradient fields:

$$GC = \left| \frac{\boldsymbol{d}_i^T \boldsymbol{d}_j}{\|\boldsymbol{d}_i\|_2 \, \|\boldsymbol{d}_j\|_2} \right| \quad \text{where } \boldsymbol{d}_i = \left.\frac{d\phi_{r_i}}{d\boldsymbol{z}^c}\right|_{\boldsymbol{z}^c = \boldsymbol{z}^c_s} \text{ and } \boldsymbol{d}_j = \left.\frac{d\phi_{r_j}}{d\boldsymbol{z}^c}\right|_{\boldsymbol{z}^c = \boldsymbol{z}^c_s}, \quad \forall i \neq j$$

Key observations:

- DC obtains best consistency across data splits. Other models struggle outside data-embeddings.
- Target domain data-embedding consistency profile similar to β accuracy profile, where intermediate values are best. Also reflective of source domain extrapolation/interpolation.
- GC improves with β for most models. DC has maximum GC and is independent of β.
- Models that obtain higher overall GC tend to do better at PRT.

Conclusions and possible future directions

Our work indicates that a relation-decoder's ability to remain consistent over regions outside the train/test distribution enhance its ability to learn a coherent concept across domains/contexts, and that β regularisation seems to benefit models lacking a similar inductive bias. Future work can continue to investigate applying these insights as regularisation to a NN, to see if it can then obtain DC performance.

References

[1] Jean Piaget.The Psychology of Intelligence. Routledge and Kegan Paul, 2005. ISBN 0521781604

[2] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, A. Lerchner, β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, 5th International Conference on Learning Representations, Toulon, France, 2017

[3] Théo Trouillon, Éric Gaussier, Christopher R. Dance, and Guillaume Bouchard. On inductive abilities of latent factor models for relational learning. Journal of Artificial Intelligence Research, 64:21–53, 2019. ISSN 10769757. doi: 10.1613/jair.1.11305