Data Augmentation for Data-Centric AI Through the Lens of Semantic Technologies: A Position Paper

Daniele Bertillo[2], Luca Cabibbo[2], Gianluca Cima[1], Valter Crescenzi[2], Marco Console[1], Roberto Maria Delfino[1], Stefano Iannucci[2], Domenico Lembo[1], Maurizio Lenzerini[1], Lorenzo Marconi[1], Paolo Merialdo[2], Marco Napoleone[2], Laura Papi[1], Antonella Poggi[1], Federico Maria Scafoglieri[1] and Riccardo Torlone[2]

[1]*Sapienza University of Rome, Italy*
[2]*Roma Tre University, Italy*

Abstract

Data augmentation is a fundamental technique in machine learning to enhance model generalization by artificially expanding training datasets. However, conventional augmentation approaches often rely on heuristic transformations that may not fully capture domain-specific knowledge. This position paper advocates a data-centric AI perspective on data augmentation, emphasizing the integration of semantic technologies, particularly domain ontologies, to guide augmentation strategies. The use of techniques from Symbolic AI for data augmentation has been dealt with only in a few recent papers. Our goal is to explore further this idea, based on the consideration that an explicit representation of the domain may be helpful in two key tasks: optimizing the generation of new data, and validating the generated data, both fundamental steps for all data augmentation strategies. We aim at developing novel approaches that combine ontologies and data augmentation techniques to address these two tasks, in particular by relying on automated reasoning. We argue that leveraging knowledge representation and symbolic reasoning enables more principled and context-aware data augmentation, leading to improved model robustness and fairness.

Keywords

Data Augmentation, Ontology Based Data Access, Semantic Technologies, Data centric Artificial Intelligence

1. Introduction

In recent years, machine learning (ML) has emerged as a transformative technology across a wide range of domains, from healthcare and finance to autonomous systems and natural language processing. A critical factor in the success of ML models is their ability to generalize well to unseen data, which is heavily influenced by the quality and diversity of the training datasets. In this framework, data augmentation (DA) has become a fundamental technique to enhance model

SEBD 2025: 33nd Symposium on Advanced Database Systems, June XX-XX, 2025, Ischia, Italy
daniele.bertillo@uniroma3.it (D. Bertillo); luca.cabibbo@uniroma3.it (L. Cabibbo); cima@diag.uniroma1.it (G. Cima); valter.crescenzi@uniroma3.it (V. Crescenzi); console@diag.uniroma1.it (M. Console); delfino@diag.uniroma1.it (R. M. Delfino); stefano.ianucci@uniroma3.it (S. Iannucci); lembo@diag.uniroma1.it (D. Lembo); lenzerini@diag.uniroma1.it (M. Lenzerini); marconi@diag.uniroma1.it (L. Marconi); paolo.merialdo@uniroma3.it (P. Merialdo); mar.napoleone3@stud.uniroma3.it (M. Napoleone); papi@diag.uniroma1.it (L. Papi); poggi@diag.uniroma1.it (A. Poggi); scafoglieri@diag.uniroma1.it (F. M. Scafoglieri); riccardo.torlone@uniroma3.it (R. Torlone)

generalization by artificially expanding training datasets through data transformations [1, 2, 3]. These transformations aim to introduce variability into the data, thereby enabling models to learn more robust and invariant representations. Numerous studies have explored individual methods as well as combinations of multiple DA operations. Also, tools and libraries exist that implement diverse DA methods. However, current approaches frequently lack systematic methodologies for evaluating and optimizing DA strategies and crafting an effective DA strategy remains a challenging and time-consuming process, often requiring domain-specific expertise [2, 3]. Moreover, conventional DA approaches often rely on heuristic transformations, such as random rotations, translations, or noise injection, which may not fully capture the underlying domain-specific knowledge.

This position paper advocates for a novel data-centric AI perspective on data augmentation, emphasizing the integration of semantic technologies, particularly domain ontologies, to guide augmentation strategies. While research on data augmentation is far from new, the coupling with semantic technologies remains largely unexplored. Although some studies employ ontologies in the data augmentation process [4, 5], they predominantly utilize them as vocabulary resources [6]. In contrast, our goal is to investigate the idea that an explicit representation of the domain can play the first citizen role in two key tasks that are fundamental to the success of any data augmentation strategy: optimizing the generation of new data and validating the generated data.

The integration of domain ontologies into data augmentation processes offers several advantages. First, ontologies provide a structured and formal representation of domain knowledge, enabling the generation of data that is semantically consistent with the underlying domain. Second, ontologies can facilitate the validation of augmented data by providing a framework for automated reasoning, ensuring that the generated data adheres to domain-specific constraints and rules. Moreover, the use of symbolic reasoning in conjunction with data augmentation opens up new possibilities for addressing challenges related to model interpretability and fairness [7]. By explicitly encoding domain knowledge and constraints, we can ensure that the augmented data reflects the desired properties and avoids biases that may be inadvertently introduced through heuristic transformations. This is particularly relevant in applications where fairness and transparency are critical, such as in healthcare or criminal justice.

This work is part of a wider project, called ENDURANCE, which is a builds upon seminal work in data provenance [8, 9] and explainable AI (xAI) [10, 11, 12], aiming to provide a more holistic and interpretable approach to DA engineering.

The rest of the paper is organized as follows. In Section 2 we review the main DA methods proposed in the literature. In Section 3, we introduce ontologies and describe the role they play in data management. Then, in Section 4, we depict the approaches we aim at exploring in order to exploit ontologies for DA, while in Section 5 we conclude the paper by discussing future work.

2. Data Augmentation

DA methods can be broadly categorized into three main approaches, though it is important to acknowledge that many techniques draw inspiration from multiple domains, often blending

ideas across these categories [1, 2, 3].

Heuristic or rule-based augmentation methods rely on manually defined transformations, making them relatively simple to implement and computationally inexpensive. However, their effectiveness is often dependent on the expertise of domain specialists who tailor the transformations to the specific dataset. While this approach can yield useful variations of data, it may struggle to introduce the level of diversity needed to significantly enhance model performance.

A more dynamic class of techniques is **mixing-based augmentation**, which generates new training samples by interpolating between two or more existing examples. Originally designed for image processing, these methods have since been adapted for NLP, time series, and tabular data. However, the discrete nature of language tokens and structured data poses challenges, as direct linear interpolation is often not feasible. To circumvent this issue, mixing-based approaches typically operate within a continuous embedding space [2, 3, 13].

Recent methods involve **deep models** explicitly trained to generate new data. Techniques such as variational autoencoders (VAEs) [14], generative adversarial networks (GANs) [15], diffusion models, and large language models fall into this category. These approaches have the potential to create highly diverse and even novel examples that extend beyond the original dataset's distribution. However, if not carefully constrained, these models can generate unrealistic or misleading samples that degrade rather than enhance learning. To mitigate this issue, recent advances have explored leveraging large pre-trained models to generate data in a more context-aware manner, ensuring that the augmented data remains relevant and meaningful [2, 3].

Each of these approaches offers distinct advantages and trade-offs. While rule-based methods provide control and interpretability, they may lack the richness of data needed for more complex learning tasks. Mixing-based methods offer an efficient way to increase sample diversity but require careful adaptation for discrete data types. Deep generative models, on the other hand, represent the cutting edge of DA, capable of synthesizing high-quality data, though they require significant computational resources and robust safeguards to maintain reliability. As research in this area continues to evolve, the interplay between these approaches is likely to shape the next generation of DA strategies.

2.1. Image Data Augmentation

DA in computer vision (CV) aims to enrich image datasets with diverse, label-consistent, variations enabling exposure to previously unseen visual patterns.

Rule-based Transformations. Simple geometric or photometric operations (e.g., random flipping, cropping, rotation, color jittering) preserve semantic labels yet diversify the input distribution. These lightweight edits are ubiquitous in CNN training pipelines, offering consistent improvements with minimal overhead [1].

Noise and Erasing. Injecting noise into the image matrix (e.g., Gaussian noise injection [16]) or masking random regions, similar to dropout regularization (e.g., Cutout [17]) encourages models to develop more robust feature representations. Variants of these techniques incorporate content-aware erasing, allowing for more controlled manipulation of hidden information.

Mixing Strategies. Inspired by Mixup [18], several methods overlay or interpolate pairs of images. This smooths the decision boundary by forcing the model to learn from "mixed" samples (e.g., partial dog, partial cat). Such mixing may reduce overfitting and improve adversarial robustness. Variants like CutMix [19] combine random patches from different images, effectively blending object context while encouraging the network to leverage global cues.

Deep Generative Models. These methods utilize powerful generative models, such as GANs and diffusion-based approaches, to create realistic synthetic images while preserving essential features like class labels [1]. Conditional GANs have been employed to generate images tailored to underrepresented classes. Similarly, diffusion models have gained traction for generating diverse yet coherent augmentations by sampling from learned image distributions. These models progressively refine images through a series of de-noising steps, ensuring realistic variations while maintaining consistency with the original dataset.

2.2. NLP Data Augmentation

DA for natural language focuses on modifying text corpora by altering syntax while maintaining semantic integrity.

Rule-based Transformations. Simple lexical edits like synonym substitution or random swaps expand a dataset with minimal overhead. For example, EDA [20] replaces or deletes words to yield new training samples while preserving semantics—often enough to boost text classification performance under low-resource conditions.

Embedding-based Mixup. Adapted from computer vision's Mixup [18], methods like MixText [13] generate "soft" data points by blending sentence embeddings. Unlike direct text-based augmentations, this approach operates in the embedding space, where interpolated vectors may not always correspond to syntactically valid sentences. Despite potential syntactic inconsistencies, these synthetic samples can preserve meaningful semantic properties, contributing to improved generalization in text classification tasks. This label-preserving technique smooths decision boundaries by exposing the model to diverse latent-space representations.

Backtranslation. A popular sequence-to-sequence approach, backtranslation [21] passes text through two or more translation steps (e.g., English→Italian→English). The resulting paraphrases have altered surface forms yet retain original meaning, yielding label-consistent diversity valuable in low-resource or domain-shift scenarios.

Language Models Generative architectures such as GPT [22] create synthetic text samples by fine-tuning on specific labels or topics, expanding the linguistic diversity of training data. Similarly, sequence-to-sequence models like T5 [23] facilitate augmentation through tasks such as paraphrasing and controlled text transformations. Meanwhile, masked language models (e.g., BERT [24]) contribute to DA by replacing tokens with contextually plausible alternatives, improving coverage of rare patterns and linguistic styles. More recently, large language models (LLMs) such as GPT-3.5 and LLaMA have demonstrated advanced capabilities in generating fluent and coherent text with minimal prompting, making them valuable tools for NLP DA. Their ability to generate highly context-aware transformations introduces new possibilities for enriching training corpora, though challenges such as computational costs and potential biases in generated content must be carefully managed.

2.3. Tabular Data Augmentation

Tabular data are widely used across various domains, typically found as web tables, databases, or spreadsheets, containing both numerical and categorical features. Due to the discrete and heterogeneous nature of tabular data, any augmentation procedure must adhere to domain-specific constraints, such as column types, positivity constraints, and permissible value ranges [3].

Some research has proposed end-to-end pipelines for tabular DA (TDA), incorporating multiple stages to ensure high-quality augmentation [3]. Broadly, TDA can be divided into two main steps: **pre-augmentation** and **augmentation**. The pre-augmentation stage includes essential table-processing tasks such as schema alignment, entity resolution, data fusion, and data cleaning to address missing or incorrect values [3]. The **augmentation** step can be further categorized into retrieval-based methods, which enrich table data by integrating relevant information from external but available sources, and **generative** methods, which synthesize new data by learning the statistical properties of the original dataset [3]. Here, we focus on **generative augmentation** approaches.

Row-level Augmentation (Record Generation): New rows can be generated using models that sample from learned data distributions, such as SMOTE [25], Gaussian Mixture Models, or GANs. These methods help mitigate class imbalance or data scarcity by generating additional training examples that preserve the statistical properties of the original data [3].

Column-level Augmentation (Feature Construction): When additional features are needed, transformation-based feature engineering can be applied to existing attributes, or generative models can be used to create entirely new columns. A key challenge in this process is ensuring that the added features contribute meaningful information rather than introducing noise [3].

Cell-level Augmentation (Cell Imputation): Missing or incorrect cell entries can degrade model performance. Generative methods such as GANs, diffusion models, or large pre-trained models can be used to impute missing values by leveraging learned distributions. Classical imputation techniques, such as regression-based methods or MICE (Multiple Imputation by Chained Equations), also remain effective for filling incomplete data [3].

3. Ontologies

In Computer Science, "ontology" is a specific term denoting an artifact that is designed for enabling the modeling of knowledge about some domain. In Artificial Intelligence, an *ontology* is a symbolic representation of a domain interest, a formal tool used to capture and reason over an explicit specification of the domain knowledge [26].

There is actually a long history of this notion. The early "semantic networks" [27] and "frame systems" [28] proposed and studied in the ’70s are the first examples of ontology formalisms developed with the rise of symbolic artificial intelligence (AI) and structured knowledge representation languages.

In the ’80s, the research on object-oriented languages and conceptual data modeling borrowed several principles from AI and adapt them towards the goal of devising new methodologies for software development and database design. The common aspect of these formalisms are

constituted by representational primitives with which to model a domain, namely classes (or concepts), attributes (or properties), and relationships (or relations among class members).

In the '90s, the research on Description Logics [29] provided solid logical foundations of ontologies, continuing the work started by Brachman and Levesque in 1986 [30], and carrying out a large body of research with the goal of understanding how the expressiveness of the class definition mechanisms used in a given Description Logic influences the feasibility and the effectiveness of automated reasoning algorithms. After more than three decades, we have now a global picture of the expressiveness/complexity trade-off in Description Logics.

In recent years ontologies are advocated as tools to provide a semantic layer for data integration, knowledge graphs and machine learning, with the goal of improving effectiveness, quality and explainability of several AI tasks.

Ontology-Based Data Management (OBDM) [31] is a well-founded paradigm aiming at enriching data with semantics, so as to rely on ontology reasoning for data management. Specifically, an OBDM system is a three-layered architecture comprising an ontology, a set of autonomous data sources and a so-called *mapping* establishing the relationship between the two. Given that within this setting data are typically very large, literature on OBDM has mainly focused on the study of forms of reasoning over the data that are tractable with respect to the data.

While traditional reasoning over ontologies focused on inferring schema-level properties, OBDM introduces the issue of using the conceptual level to access and manipulate data. For this reason, the most studied form of reasoning in this setting is query answering, i.e. the problem of logically deriving answers to queries, based on the knowledge expressed in the ontology and the mappings. Notably, query answering is the basis for other relevant tasks in OBDM, such as data quality checking [32, 33] and privacy preserving query processing.

In general, OBDM resorts to symbolic techniques to infer (new) insights from data, as opposed to ML, that induces insights from data through sub-symbolic techniques.

4. Proposed Approach

As discussed in section 2, current DA techniques can be categorized into two broad classes: rule-based or sub-symbolic, i.e., based on sub-symbolic AI approaches. As each approach shows strengths that could help mitigating the other's weaknesses, we argue that a more robust DA framework should mix together techniques from both approaches. Specifically, taking into account the semantics of the application domain defined by *ontological axioms* may improve two fundamental steps performed via purely sub-symbolic DA techniques: *optimizing* the generation, and *strengthening* the validation of synthetic data.

In what follows, we propose several novel approaches to combine ontologies and sub-symbolic DA techniques to address these two tasks. A crucial assumption for all these techniques is that *ontological axioms and training data are compatible*, i.e., we can perform standard reasoning tasks over the specification defined by the data and the axioms. For this purpose, we envision the use of specifically crafted mappings, able to reconcile ML models and domain knowledge in the ontology.

Additionally, we assume that ontological axioms are always able to define *correct and meaningful* information on the domain of interest. Due to the specific use cases we are considering,

this may require the use of expressive ontological languages that can handle probabilistic rules.

Generation The first task we aim to tackle is the optimization of the data generation process. In this context, we discuss two different approaches in which an ontology is used to influence the data generation process. For this purpose, let the generative process be denoted by a function $G(\cdot)$, called *generator*, which outputs new unseen data samples.

- *Ontology-guided generation*: The ontology may be used to process the output of $G(\cdot)$ and improve it according to its axioms. In particular, one could use an ontology to verify the consistency of the generated samples against the domain knowledge possessed by experts and fix possible inconsistencies. A similar approach is proposed by [34] where logical constraints are incorporated in the output layer of a neural network. However, such constraints are expressed purely in terms of formulae over the attributes of the data samples. In the approach we propose, on the contrary, we want to formalize these constraints using a higher level of abstraction, i.e. using the symbols defined by an ontology. This will allow to perform reasoning over the domain, possibly inferring new knowledge concerning the generated sample, and therefore further augmenting the data.
- *Ontology as the generator*: Another possible approach is to use the ontology as a set of rules that define the behavior of the generator $G(\cdot)$ itself. In this context, the ontology could be used to generate data samples that are either already consistent with the domain knowledge or even partially consistent. In the latter case, we could also exploit the ontology to measure the inconsistencies and eventually address them.

Validation Performing a validation step on the output of a generative technique is a common practice that has been shown to improve the quality of the output. Usually, given a generator $G(\cdot)$, its validation involves training a binary classifier called discriminative model, or discriminator, $D(\cdot)$. This discriminator is trained to distinguish between *real* samples, i.e., those close enough to the original dataset, and *synthetic* samples, i.e., those that that lack of some defining features. Combining this technique with ontologies could lead to improved validation results.

- *Ontology-enhanced validation*: this approach employs the ontology to improve the performance of a given discriminator $D(\cdot)$, i.e., providing a more accurate classification. To this end, we propose two alternative strategies: *ontology-first* and *discriminator-first*. The *ontology-first* strategy consists in applying the ontology rules to the *input* of $D(\cdot)$, before the classification is computed. The intuition behind this approach is that ontological axioms could be used to infer additional knowledge on the data samples. This knowledge could enhance the input of $D(\cdot)$ thus making the classification more precise. Adding this information to the data samples, however, may require the modification of the the architecture of $D(\cdot)$ to ensure compatibility. In contrast, the *discriminator-first* approach applies the ontological axioms to the *output* of $D(\cdot)$, verifying its consistency with the domain knowledge, without interfering with the model architecture. For example, a sample misclassified as *real* by $D(\cdot)$, could get the correct label *synthetic* after verifying the violation of some domain rules.
- *Ontology as the discriminator*: another novel approach we propose is the use of an ontology as the discriminator $D(\cdot)$ itself. Instead of training a model to recognize synthetic data, the ontology could be used to directly validate the output of a generator $G(\cdot)$ checking consistency directly against domain knowledge.

Combining generation and validation in an online approach In the previous paragraphs we described techniques to employ ontologies in an offline fashion, i.e., given G and D. A natural, and particularly interesting, extension of our study is to investigate online approaches where ontologies are integrated in the training phase of generative and discriminative machine learning models. In particular, we propose to use this approach in conjunction with GAN models, in which ontologies could be used to guide the learning process of both G and D.

Some of the most notable approaches in the literature are based on compiling the semantic constraints into the loss function of the training algorithm [35, 36], or directly into the model structure [37, 38]. However, none of these approach resort to an ontology to model the domain knowledge, which is at the core of our proposal. We argue that leveraging domain ontologies in the training phase of both the generative model $G(\cdot)$ and the discriminative model $D(\cdot)$ of a GAN leads to a context-aware generation of data, and therefore to improve DA.

5. Conclusions and future work

This paper advocates for integrating domain ontologies into DA to address the limitations of conventional heuristic approaches in knowledge-intensive ML applications. By formalizing domain knowledge through semantic technologies, ontology-driven DA enhances both data generation and validation processes.

Three key findings emerge: (1) ontologies enable semantically consistent training data expansion, overcoming the domain-agnostic nature of traditional transformations like rotations or noise injection; (2) automated reasoning via ontological constraints ensures augmented data validity, preventing semantic inconsistencies that could degrade model performance, and (3) explicit knowledge representation supports fairness and interpretability in high-risks domains like healthcare by mitigating biases inherent in purely statistical DA methods.

The proposed framework bridges symbolic AI and data-centric paradigms, offering systematic strategies to align synthetic data with domain-specific rules. To validate the effectiveness of the proposed approach, the project will conduct experimental analysis across multidisciplinary scenarios encompassing three distinct use cases:

Entity Resolution: The first use case will involve entity resolution tasks on structured and semi-structured data [11, 39, 40]. In this scenario, our objective is to evaluate how the proposed *validation approaches* can improve known techniques for matching and linking records across different data sources.

NLP Scenario: In collaboration with the Senate of the Italian Republic, the second use case will focus on clustering textual data (law amendments) [41], exploring how ontologies can support *generative approaches* for text generation and NLP tasks.

Hand Writer Identification: The third use case will address the *writer identification* problem for medieval manuscripts [42]. This use case in the digital humanities will test the applicability of DA techniques to image-based tasks. Tasks involving images may benefit from an *online approach* to augmentation as we discussed in the previous sections.

Future work should explore: (1) neuro-symbolic integration combining ontological reasoning with neural data generation, (2) constraint-based augmentation using ontological rules, and (3) automated validation pipelines verifying semantic consistency across augmentation cycles.

Acknowledgements

This work has been supported by PNRR MUR project PE0000013-FAIR.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] C. Shorten, T. M. Khoshgoftaar, A survey on image data augmentation for deep learning, Journal of big data 6 (2019) 1–48.
[2] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, E. Hovy, A survey of data augmentation approaches for nlp, arXiv preprint arXiv:2105.03075 (2021).
[3] L. Cui, H. Li, K. Chen, L. Shou, G. Chen, Tabular data augmentation for machine learning: Progress and prospects of embracing generative ai, arXiv preprint arXiv:2407.21523 (2024).
[4] M. Skreta, A. Arbabi, J. Wang, E. Drysdale, J. Kelly, D. Singh, M. Brudno, Automatically disambiguating medical acronyms with ontology-aware deep learning, Nature communications 12 (2021) 5319.
[5] C. Sun, M. Dumontier, Generating unseen diseases patient data using ontology enhanced generative adversarial networks, npj Digital Medicine 8 (2025) 4.
[6] Y. Qiu, Y. Jin, A method for synthesizing ontology-based textual design datasets: Evaluating the potential of llm in domain-specific dataset generation, Journal of Mechanical Design (2024) 1–41.
[7] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Computing Surveys (CSUR) 54 (2021) 1–35.
[8] A. Chapman, P. Missier, G. Simonelli, R. Torlone, Capturing and querying fine-grained provenance of preprocessing pipelines in data science, Proceedings of the VLDB Endowment 14 (2020) 507–520.
[9] A. Chapman, P. Missier, L. Lauro, R. Torlone, Supporting better insights of data science pipelines with fine-grained provenance, ACM Transactions on Database Systems 49 (2024) 1–42.
[10] A. Rossi, D. Firmani, P. Merialdo, T. Teofili, Explaining link prediction systems based on knowledge graph embeddings, in: Proceedings of the 2022 international conference on management of data, 2022, pp. 2062–2075.
[11] T. Teofili, D. Firmani, N. Koudas, V. Martello, P. Merialdo, D. Srivastava, Effective explanations for entity resolution models, in: 2022 IEEE 38th International Conference on Data Engineering (ICDE), IEEE, 2022, pp. 2709–2721.
[12] L. Lastilla, S. Ammirati, P. Merialdo, How explainable is automatic hand identification? a case study, in: UWA Conference 2023, 2023.
[13] J. Chen, Z. Yang, D. Yang, Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification, arXiv preprint arXiv:2004.12239 (2020).
[14] D. P. Kingma, M. Welling, et al., Auto-encoding variational bayes, 2013.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Communications of the ACM 63 (2020) 139–144.
[16] F. J. Moreno-Barea, F. Strazzera, J. M. Jerez, D. Urda, L. Franco, Forward noise adjustment scheme for data augmentation, in: 2018 IEEE symposium series on computational intelligence (SSCI), IEEE, 2018, pp. 728–734.
[17] T. DeVries, G. W. Taylor, Improved regularization of convolutional neural networks with cutout, arXiv preprint arXiv:1708.04552 (2017).
[18] H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, arXiv preprint arXiv:1710.09412 (2017).
[19] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, Y. Yoo, Cutmix: Regularization strategy to train strong classifiers with localizable features, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 6023–6032.
[20] J. Wei, K. Zou, Eda: Easy data augmentation techniques for boosting performance on text classification tasks, arXiv preprint arXiv:1901.11196 (2019).
[21] R. Sennrich, B. Haddow, A. Birch, Improving neural machine translation models with monolingual data, arXiv preprint arXiv:1511.06709 (2015).
[22] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.
[23] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of machine learning research 21 (2020) 1–67.
[24] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.
[25] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of artificial intelligence research 16 (2002) 321–357.
[26] N. Guarino, D. Oberle, S. Staab, What is an ontology?, Handbook on ontologies (2009) 1–17.
[27] W. A. Woods, What's in a link: Foundations for semantic networks, in: Representation and understanding, Elsevier, 1975, pp. 35–82.
[28] M. Minsky, A framework for representing knowledge, 1974.
[29] F. Baader, W. Nutt, Basic description logics, in: The description logic handbook: theory, implementation, and applications, 2003, pp. 43–95.
[30] R. J. Brachman, H. J. Levesque, Readings in knowledge representation, Technical Report, AT and T Bell Labs., 1985.
[31] A. Poggi, D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, R. Rosati, Linking data to ontologies, in: Journal on data semantics X, Springer, 2008, pp. 133–173.
[32] C. Daraio, M. Lenzerini, C. Leporelli, P. Naggar, A. Bonaccorsi, A. Bartolucci, The advantages of an ontology-based data management approach: openness, interoperability and data quality, Scientometrics 108 (2016) 441–455.
[33] M. Console, M. Lenzerini, Data quality in ontology-based data access: The case of consistency, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 28,

2014.

[34] N. Hoernle, R. M. Karampatsis, V. Belle, K. Gal, Multiplexnet: Towards fully satisfied logical constraints in neural networks, 2021. URL: https://arxiv.org/abs/2111.01564. arXiv:2111.01564.

[35] J. Xu, Z. Zhang, T. Friedman, Y. Liang, G. V. den Broeck, A semantic loss function for deep learning with symbolic knowledge, 2018. URL: https://arxiv.org/abs/1711.11157. arXiv:1711.11157.

[36] M. Fischer, M. Balunovic, D. Drachsler-Cohen, T. Gehr, C. Zhang, M. Vechev, DL2: Training and querying neural networks with logic, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 1931–1941. URL: https://proceedings.mlr.press/v97/fischer19a.html.

[37] G. G. Towell, J. W. Shavlik, M. O. Noordewier, Refinement of approximate domain theories by knowledge-based neural networks, in: Proceedings of the Eighth National Conference on Artificial Intelligence - Volume 2, AAAI'90, AAAI Press, 1990, p. 861–866.

[38] A. Daniele, L. Serafini, Neural networks enhancement with logical knowledge, 2021. URL: https://arxiv.org/abs/2009.06087. arXiv:2009.06087.

[39] V. Di Cicco, D. Firmani, N. Koudas, P. Merialdo, D. Srivastava, Interpreting deep learning models for entity resolution: an experience report using lime, in: Proceedings of the second international workshop on exploiting artificial intelligence techniques for data management, 2019, pp. 1–4.

[40] R. Fagin, P. G. Kolaitis, D. Lembo, L. Popa, F. Scafoglieri, A Framework for Combining Entity Resolution and Query Answering in Knowledge Bases, in: Proceedings of the 20th International Conference on Principles of Knowledge Representation and Reasoning, 2023, pp. 229–239.

[41] A. Sajeva, S. Iannucci, C. Marchetti, P. Merialdo, R. Torlone, Clustering amendments with semantic embeddings, SEBD 2024 (2024).

[42] L. Lastilla, S. Ammirati, D. Firmani, N. Komodakis, P. Merialdo, S. Scardapane, Self-supervised learning for medieval handwriting identification: A case study from the vatican apostolic library, Information Processing & Management 59 (2022) 102875.