

Streszczenie

Uczenie głębokie stanowi podstawę percepcji w robotyce, inspekcji przemysłowej i inteligentnej infrastrukturze. Zastosowania te wymagają niskich opóźnień, działania przy niestabilnej łączności oraz lokalnego przetwarzania danych, dlatego wnioskowanie przenosi się na urządzenia brzegowe umieszczone bezpośrednio przy czujnikach. Urządzenia te dysponują jednak ograniczoną pamięcią, mocą obliczeniową i energią, a wąski zestaw obsługiwanych operacji obliczeniowych zawęża zakres modeli, które można uruchomić. Nawet jeśli uda się pokonać te bariery, po etapie wdrożenia pojawiają się kolejne wyzwania. Od systemów brzegowej sztucznej inteligencji (ang. *edge AI*) oczekuje się niezawodnego działania przez lata, a zmienność środowiska sprawia, że operują one w warunkach odbiegających od tych z etapu uczenia. Zjawisko dryfu danych wymusza na modelach adaptację, która z kolei jest znacząco utrudniona ze względu na wspomniane bariery sprzętowe oraz słaby sygnał nadzorujący.

W rozprawie zbadano, czy algorytmy percepcji dostosowane do zadania, implementacja uwzględniająca specyfikę sprzętową, selektywne aktualizacje modelu oraz kompaktowa wymiana wiedzy pozwalają uzyskać korzystniejszy kompromis między jakością predykcji a kosztami wnioskowania, adaptacji i współpracy. Układ pracy odpowiada cyklowi operacji związanych z systemem percepcji na urządzeniu brzegowym, obejmując wdrożenie, adaptację i współdziałanie między urządzeniami. Dla każdego zadania wskaźniki jakościowe zestawiono z parametrami wydajnościowymi zmierzonymi na platformach referencyjnych, od mikrokontrolerów po procesory graficzne dla systemów wbudowanych. Do tego celu wykorzystano trzy zadania wizyjne: szacowanie liczby osób na podstawie obrazów termowizyjnych niskiej rozdzielczości, diagnostykę uszkodzeń silników indukcyjnych z wykorzystaniem obrazów termowizyjnych oraz prognozowanie natężenia promieniowania słonecznego w krótkim horyzoncie czasowym w oparciu o obrazy całego nieba. Ostatnie z nich posłużyło do oceny wszystkich trzech etapów. Wkładem rozprawy jest weryfikacja możliwości realizacji niezawodnej percepcji na urządzeniach o ograniczonych zasobach, obejmująca pełny cykl operacyjny i oparta na rzeczywistych pomiarach kosztów. Uzyskane wyniki, zestawione według schematu raportowania opartego na dziewięciu kryteriach niezawodności, potwierdzają postawioną hipotezę, której cztery składowe zweryfikowano w odrębnych studiach przypadku. Rozprawa dostarcza również praktycznych wytycznych wskazujących, kiedy optymalizacja po stronie modelu przynosi korzyści oraz kiedy koszty adaptacji i uczenia zespołowego są uzasadnione.

Abstract

Deep learning underpins perception in robotics, industrial inspection, and smart infrastructure. These applications require low latency, operation under unstable connectivity, and local data processing, so inference is moving to edge devices placed close to the sensors. Such devices, however, have limited memory, compute, and energy, and the limited set of supported operations narrows the range of models that can execute at all. Even if these barriers are overcome, further challenges appear after deployment. Edge AI systems are expected to operate for years, and environmental variability means that they run under conditions that differ from those encountered during training. Data drift necessitates adaptation, which is, in turn, severely hindered by the aforementioned hardware constraints and weak supervisory signals.

This dissertation examines whether task-specific perception algorithms, hardware-aware execution, selective model updates, and compact knowledge exchange yield a better trade-off between predictive performance and the costs of inference, adaptation, and collaboration. The work follows the lifecycle of a perception system on an edge device, covering deployment, adaptation, and interaction across devices. For every task, quality metrics are set against performance figures measured on reference platforms ranging from microcontrollers to embedded GPUs. Three vision tasks serve this purpose, namely occupancy estimation in low-resolution thermal images, induction-motor fault diagnosis from thermal images, and short-horizon solar irradiance forecasting from all-sky imagery, the last of which is used to evaluate all three stages. The dissertation contributes a measured, lifecycle-wide validation that reliable edge perception can be sustained on devices with limited hardware resources. The evidence, gathered with a reporting framework of nine reliability criteria, confirms the hypothesis through its four components, verified in separate case studies. The dissertation also provides practical guidelines on when model-side optimization is beneficial, and when adaptation and collaborative learning justify their costs.