We propose the Unified Visual-Semantic Embeddings (Unified VSE) for learning a joint space of visual representation and textual semantics. The model unifies