cases	doc_1		doc_2		decision	id
	authors	• K. Leidal • David F. Harwath • James R. Glass	**authors**	• Kenneth Leidal • David Harwath • James Glass	DUPLICATES	370
	title	Learning modality-invariant representations for speech and images	**title**	Learning Modality-Invariant Representations for Speech and Images		
	publication_date	2017-12-01 00:00:00	**publication_date**	2017-12-11 17:18:34+00:00		
	source	SupportedSources.SEMANTIC_SCHOLAR	**source**	SupportedSources.ARXIV		
	journal		**journal**	None		
	volume		**volume**			
	doi	10.1109/ASRU.2017.8268967	**doi**			
	urls	• https://www.semanticscholar.org/paper/f69b80515b553bab0564b58555dd92780e606792	**urls**	• http://arxiv.org/pdf/1712.03897v1 • http://arxiv.org/abs/1712.03897v1 • http://arxiv.org/pdf/1712.03897v1		
	id	id-4630528516066062966	**id**	id-7531702360856458584		
	abstract	In this paper, we explore the unsupervised learning of a semantic embedding space for co-occurring sensory inputs. Specifically, we focus on the task of learning a semantic vector space for both spoken and handwritten digits using the TIDIGITs and MNIST datasets. Current techniques encode image and audio/textual inputs directly to semantic embeddings. In contrast, our technique maps an input to the mean and log variance vectors of a diagonal Gaussian from which sample semantic embeddings are drawn. In addition to encouraging semantic similarity between co-occurring inputs, our loss function includes a regularization term borrowed from variational autoencoders (VAEs) which drives the posterior distributions over embeddings to be unit Gaussian. We can use this regularization term to filter out modality information while preserving semantic information. We speculate this technique may be more broadly applicable to other areas of cross-modality/domain information retrieval and transfer learning.	**abstract**	In this paper, we explore the unsupervised learning of a semantic embedding space for co-occurring sensory inputs. Specifically, we focus on the task of learning a semantic vector space for both spoken and handwritten digits using the TIDIGITs and MNIST datasets. Current techniques encode image and audio/textual inputs directly to semantic embeddings. In contrast, our technique maps an input to the mean and log variance vectors of a diagonal Gaussian from which sample semantic embeddings are drawn. In addition to encouraging semantic similarity between co-occurring inputs,our loss function includes a regularization term borrowed from variational autoencoders (VAEs) which drives the posterior distributions over embeddings to be unit Gaussian. We can use this regularization term to filter out modality information while preserving semantic information. We speculate this technique may be more broadly applicable to other areas of cross-modality/domain information retrieval and transfer learning.		
	versions		**versions**			